NATIONAL CENTER FOR EDUCATION STATISTICS

Working Paper Series *August 2001*

The Working Paper Series was initiated to promote the sharing of the valuable work experience and knowledge reflected in these preliminary reports. These reports are viewed as works in progress, and have not undergone a rigorous review for consistency with NCES Statistical Standards prior to inclusion in the Working Paper Series.

U.S. Department of Education
Office of Educational Research and Improvement.

NATIONAL CENTER FOR EDUCATION STATISTICS

Working Paper Series

Assessing the Lexile Framework: Results of a Panel Meeting

Working Paper No. 2001-08 **August 2001**

Sheida White, Ph.D.
Assessment Division
National Center for Education Statistics

John Clement, Ph.D.
Education Statistics Services Institute

U.S. Department of Education
Office of Educational Research and Improvement.

U.S. Department of Education
Rod Paige
Secretary

Office of Educational Research and Improvement
Grover J. Whitehurst
Assistant Secretary

National Center for Education Statistics
Gary W. Phillips
Acting Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high quality data to the U.S. Department of Education, the Congress, the state, other education policymakers, practitioners, data users and the general public.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to:

National Center for Education Statistics
Office of Educational Research and Improvement
U.S. Department of Education
1990 K Street, NW
Washington, DC 20006

August 2001

The NCES World Wide Web Home Page address is: http://nces.ed.gov/
The NCES World Wide Web Electronic Catalog is: http://nces.ed.gov/pubsearch/

Suggested Citation
U.S. Department of Education, National Center for Education Statistics. *Assessing the Lexile Framework: Results of a Panel Meeting*, NCES 2001-08, by Sheida White and John Clement. Washington, DC: 2001.

Contact:

Sheida White
sheida.white@ed.gov
202-502-7473

Working Paper Foreword

In addition to official NCES Publications, NCES staff and individuals commissioned by NCES produce preliminary research reports that include analyses of survey results, and presentations of technical, methodological, and statistical evaluation issues.

The ***Working Paper Series*** was initiated to promote the sharing of the valuable work experiences and knowledge reflected in these preliminary reports. These reports are viewed as works in progress, and have not undergone a rigorous review for consistency with NCES Statistical Standards prior to inclusion in the Working Paper Series.

Copies of Working Papers can be downloaded as pdf files from the NCES Electronic Catalog (http://nces.ed.gov/pubsearch/), or contact Sheilah Jupiter by phone at (202) 502-7444, or by e-mail at sheilah.Jupiter@ed.gov, or by mail at U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, 1990 K Street NW, Room 9048, Washington, DC 20006.

Marilyn Seastrom
Chief Mathematical Statistician
Statistical Standards Program

Ralph Lee
Mathematical Statistician
Statistical Standards Program

Assessing the Lexile Framework: Results of a Panel Meeting

Prepared by
Sheida White, Ph.D.
Assessment Division

John Clement, Ph.D.
Education Statistics Services Institute

Prepared for:
U.S. Department of Education
Office of Educational Research and Improvement
National Center for Education Statistics

August 2001

Table of Contents

1. Background…….. 2
2. A brief description of the Lexile Framework…….. 2
3. Recommendations from the panel discussion……. 3
    3.1. Possible applications…… 4
    3.2. Areas of concern – semantic theory……. 5
    3.3. Areas of concern – syntactic theory……. 6
    3.4. Areas of concern – psychometric issues…….. 6
4. Summary and conclusions……... 9

Attachment 1. Lexile Framework Assessment Panel Members
    and Invited Observers…….. 11
Attachment 2. Invited Papers:
    Adams, M.J. …… 15
    Kamil, M. ……. 22
    Larson, R. ……. 27
    Smith, C. …….. 34
    Wrigley, H.S. …… 39
Attachment 3. Comments from Lexile Framework developers…. 46

1. Background

This working paper summarizes the results of an expert panel review of the Lexile Framework (LF). The review was conducted by five panel members through readings, the preparation of brief individual reports, and participation in a meeting held on April 26, 2001 in Washington, D.C. The list of panel members and invited observers, along with brief biographical statements, is contained in Attachment 1.

The Lexile Framework is a linguistic-theory-based method for measuring the reading difficulty of prose text. Information about the Framework can be found in a number of published documents, and at the website (http://www.lexile.com).

The panel's charge was to provide NCES with informed judgment on the theoretical underpinnings and construct validity of the Framework. Specifically, panel members were asked to address the Lexile Framework's basis in linguistic theory. The following questions framed the panel's work and deliberations:

1. Are word frequency and sentence length *solid* criteria to use in determining text difficulty?

2. Are these criteria *sufficient* to determine text difficulty? If not, can they be improved or used for only a subset of reading passages?

3. Are the *procedures* used to determine word frequency and sentence length, and their calibration through Rasch analysis, adequate? If not, how can they be improved? Are there alternative procedures for assessing readability?

4. What is the relationship between the Lexile Framework and other measures of text difficulty?

It must be noted that these questions do not directly deal with issues of educational effectiveness. The panel was not asked to address the worth of the Framework as an educational resource.

2. A brief description of the Lexile Framework

The LF measures the reading difficulty of prose texts, and the reading capacity of people. It is not intended for use with poetry or texts with genre-specific features such as menus, documents, etc.

Measuring reading difficulty: The process is based on a mathematical formula that assigns reading difficulty values to passages of text (known as "slices"). These values

are scaled; the scale unit is called a **lexile**. The mean lexile value of all the slices in a book is the lexile value of that book.

The LF formula is based on two components. The first is a measure of sentence length, which by hypothesis indicates the level of syntactic complexity. Wright and Stenner 1998 base this on what they call *the syntactic axiom:* the shorter the sentences, the easier the passage is to read. The second component is a measure of semantic complexity, based on word familiarity in terms of inverse frequency of appearance in a corpus of text works. Again, the inventors of the LF base this measure on *the semantic axiom*: the more familiar the words, the easier the passage is to read. Details of the computation of the formula, and information on how scaling factors were obtained, can be found in Stenner (n.d.); Stenner (1996); Stenner and Burdick (1997); and Wright and Stenner (1998).

Measuring reading capacity: The authors have developed a process that assigns a value, also in lexile units, to the reading capacity of a person. The central idea is that, when a person is reading with 75% comprehension, they are at optimal reading capacity. The process therefore assesses a person's level of reading comprehension, and then calculates the lexile value of texts they can read with 75% comprehension. This is the measure of their reading capacity.

On the basis of these assumptions and a good deal of testing, the creators of the framework have established anchor points (the value in lexiles of first grade basal readers is set at 200, and the value in lexiles of twelfth-grade texts is set at 1200). The LF scale is set at 1000 points between first and twelfth grades. The scale itself extends from 0 to 1800 lexiles.

3. Recommendations from the panel discussion

The panel affirmed the value of both sentence length and word frequency as overall measures of semantic and syntactic complexity, although participants diverged on whether these constructs were best viewed as proxies or as direct measures; they can in fact be viewed as both.

Although most efforts to establish readability measures use similar indicators for semantic and syntactic complexity, the LF, in the words of one panel member's report, appears "…exceptional in the psychometric care with which it has been developed; the extent of its formal validation with different populations of texts, tests, and children; in its automation; and in its developers' continual quest to improve it (Adams, attached)."

Researchers in psychology, education, and reading have carried out extensive research to identify the factors that underlie reading comprehension and difficulty. This research has not, however, led to indicators that improve substantially upon those in the LF in power or practical utility.

There remain a number of concerns, or areas in which further research is needed. However, some potential areas of application for the LF with regard to student assessments of interest to the Center can be contemplated.

3.1. Possible applications:

- **Selection of candidate reading passages for assessment items**. It seems possible to use the LF database of tens of thousands of Lexiled passages in thousands of books to select candidate passages that have a desired Lexile level and other desirable properties for test use. For example, a passage out of The Pickwick Papers (1160 Lexiles overall) might be selected for the 12th grade NAEP (1150 Lexiles overall); knowing the overall level, an expert item developer can focus on other qualities that make a good item (passage interest level and contextual clues, for instance)[1]. The LF would be especially useful for this purpose if texts and passages were indexed by grade and topic. At present, the Lexile Web site does not provide such indexing for texts.

- **Obtaining Lexile values for text passages in use in existing assessments**. Prose passages from NAEP and NAAL assessments can be Lexiled retroactively, through use of a formula computation routine at the Lexile web site (http://www.lexile.com). Examination of passages in this fashion might show that a particular passage in a fourth grade NAEP assessment might be more suited to the eighth-grade level, for instance. Alternatively, such examination might show that prose passages selected for an upcoming NAAL assessment had a desirable range (say, 460 to 1200 Lexiles) of reading difficulties.

- **Assessment item and assessment comparisons**. There are some limited contexts in which it may be useful to NCES to compare different reading assessment instruments – for example, to compare the 12th grade NAEP to the NAAL. The objective would be to compare reading difficulty ranges, and to confirm coverage of a proper range of reading difficulty levels. Other comparisons at the passage and assessment level could include examination of Lexile levels in passages administered in different years.

- **Contrasting performance levels with Lexile scores**. In NAEP and NAAL, it may be useful to compare respondents' performance on open-ended items built around prose text with the predicted performance based on the Lexile value of the prose text.

- **Assessing the reading difficulty level of NAEP and NAAL questions**. In addition to assessing the reading difficulty of NAEP and NAAL text passages, the LF might be used to assess the difficulty of the actual questions, if the framework can be adapted for this purpose.

[1] One of the panel members notes that imprecision in Lexile scores is an important issue, and that a score for a slice of text may have a confidence interval that approaches one grade level in size. See also the discussion in section 3.4.

3.2. Areas of concern – semantic theory:

- **Variability in the middle range of the word frequency corpus, and other frequency-related issues**. Concern was expressed by panel members that the middle range of word frequencies had large numbers of words with similar global frequencies of occurrence. This may mask differences in the middle ranges. Moreover, many passages would have approximately equal Lexile values, thus diminishing the discriminating power of the LF in the middle range of reading difficulties. A possible course of action would be to experiment with other ways (either mathematical or semantic) of defining word frequency that would provide more discrimination in the middle ranges.

  Underlying issues of variability and imprecision is the basic fact that word frequency distributions in any real corpus are highly skewed. The extent to which true frequencies of occurrence of relatively uncommon words can be measured depends very strongly on the size of the corpus, which inevitably has to be very large. Increasing the size of the corpus, therefore, is an essential requirement for improving the estimation of text difficulty.

- **Separating ideational complexity from rarity**. As Adams points out in her report to the panel (see attached), "word frequency and semantic accessibility hold an asymmetrical relationship to one another." The infrequency of appearance of a word in a corpus can be due to the fact that it occurs rarely in print – for reasons unrelated to its conceptual complexity; frequency of appearance alone does not simply predict the degree of semantic sophistication. Whether it is possible to achieve this kind of separation in refining a measure of reading difficulty is not known.

- **Other semantic issues.** Published word frequency lists don't adequately represent new and/or technical terms, informal or slang words in current use; function words; alternative forms of presentation of equivalent terms (i.e. possessives, plurals, capitalized forms); or proper nouns.

- **The LF does not assess other aspects of reading, including subject difficulty, information in the context, or discourse features**. Research affirms that the comprehensibility of text is significantly influenced by such inter- and extra-sentential factors as contextual clues, discourse markers such as "this" or "so", or particular syntactic structures such as foregrounding or other syntactic ordering, used to make certain information more salient. Because the LF makes use of within-sentence factors only, its potential for assessing the effects of any such contextual, pragmatic, or discourse features is inherently limited.

3.3. Areas of concern– syntactic theory:

- **Adequacy of sentence length as a measure of complexity**. A number of possible modifications to the estimation of sentence length were suggested, including the addition of modifications to the measure based on such concepts as transparency (sentences are transparent if their form mirrors the way they are understood), embedded clauses or cleft sentences (excellent examples provided in Smith, attached). In particular, Kamil (attached) cites research by D. Pearson in which shortened sentences become more, rather than less, difficult, due to more complex syntax. It might be valuable to examine how such an insight could be used to improve the item selection and design process.

- **Relative contribution of syntactic and semantic complexity**. The panel suggested that, although syntactic and semantic complexity were admittedly highly correlated, any possible independent contribution of the two components of reading difficulty might be worth examination through more detailed statistical analysis. Furthermore, within the semantic component, it would be useful to know the relative contribution to reading difficulty of the truly semantic features (content classification and part of speech) in comparison to such proxy features (to the LF) as number of letters, number of syllables, modal grade of appearance of a word, or relative frequency.

It is important to acknowledge, however, that since the LF scaling is based on sentence length and word frequency proxies, not much variation in Lexile-based measures is likely to be left over for other aspects of comprehension. In practice, the LF may simply not offer enough residual variation in the general population of readers to explore other factors. [However, it may be productive to explore sources of residual variation in other reader populations; cf. Section 3.4 below.]

3.4. Areas of concern – psychometric issues:

- **Explaining "within-ensemble" variation and other issues related to item and passage validity**. There is persistent variation (or imprecision) in a Lexile score assigned to a slice of text or to an individual. Such variation can approach, in the aggregate, a grade level in size. Accounting for the imprecision of Lexile scores is an important issue.

Person-related imprecision relates to the methods by which a Lexile score is assigned to an individual. Such imprecision appears to amount to approximately 40-100 Lexiles, depending on reader and text characteristics (cf. Stenner and Burdick 1997: 23-30, especially Table 8).

There are both text measure errors and method-related errors. Text-measure errors appear to amount to 25-75 Lexiles, and could be eliminated entirely (cf. Stenner and Burdick 1997:21-22). With regard to method-related errors, one can think of a text passage (or Lexile "slice") and an associated question (intentionally selected and framed

as a “rational cloze”) as a single unit. No matter what question one asks in relation to a passage, however, the LF formula returns the same value. There are different ways in which a question might be asked from that passage. The different questions can lead to differences in performance; these differences, amounting to an imprecision in the Lexile score for a passage, have not so far been accounted for. Testable questions can be asked about the precision or accuracy of the LF formula in assigning a value to a text passage. Similar questions can be asked of the use of the “rational cloze” approach as opposed to other cloze methods or other means of asking a question to show understanding of the meaning of a passage.

- **Generalizing to other genres**. It is important to find alternatives to the LF for assessing ability with special genres such as quantitative text and document-based forms.

"Document literacy" (forms, applications; schedules; signs, labels; notices; announcements) is such an important part of reading for adults and a significant part of the NAAL - yet no work has been done in terms of applying the framework to documents with a fair amount of discontinuous texts. This constitutes a severe limitation. A framework might be developed that gives a sense of what the "cognitive load" of such a document might be (that is, how much processing ability it takes to unpack the information). If we don't have that kind of information, it must be constantly be highlighted that the framework only applies to prose texts and that most of the reading that adults do who don't have strong educational backgrounds falls into the document category.

It becomes important to know how many adults were studied as the framework was established, and to what extent authentic texts were used (as opposed to textbooks) and whether scores differed in the lower levels between children and adults - for example, what kinds of things can an adult at a Lexile level of, say, 1000, read, as opposed to a child with the same score. There has been a long battle fought to avoid having grade level scores applied to adults, since they are meaningless in terms of the kinds of abilities that adults have when it comes to interpreting print in their daily lives. The differences between adult readers and young readers will need to be discussed if not sorted out; otherwise, if scores are applied to materials developed for schoolchildren and adults in the same way, the results are likely to be skewed for both groups.

- **Applicability of the LF to the reading assessment of second-language readers**.

The LF has only had a relatively small amount of work done in languages other than English. It would not be possible to make a definitive statement about whether second-language learners should be scored the same way as native speakers of English. The Panel expressed concern over the applicability of the LF to second-language readers, citing the following reasons:

- For those who are not fully proficient in the language they are trying to read, syntax often plays a much greater role than mere sentence length. Fairly short sentences that are easily understood by native speakers (passives for example)

will present difficulties ("He was run over by a truck"). Similarly, idiomatic expressions, such as "You want to bet 20 bucks on this? You're on." are difficult to comprehend for those who are fairly new to English.

- Words familiar to native speakers may not be familiar to non-native readers, and familiarity with English vocabulary will differ, depending on the language of origin. Those coming from Romance languages for example will recognize Latin based words, such as "edifice" and "malnutrition" but may have difficulty with every day words such as "mop" or "turnip" , since these rarely appear in ESL/EFL textbooks.

- Texts with strong references to cultural contexts are likely to present significant difficulties as well where. While native speakers are much more likely to understand references to sports such as American football and baseball, comprehension breaks down for non-native speakers who lack the necessary background knowledge to interpret a text that is U.S. specific.

- The ability to read in English and the ability to read in general should not be confused. Educated non-English speakers may have fairly high levels of reading proficiency in the native language, but my not know sufficient English to transfer that knowledge to English texts. Assigning a Lexile score based on the ability to process texts in English is likely to offer a false picture of a person's ability to comprehend what is written, unless the distinction between the ability to read a text in English and the ability to process print in general is made clear.

To maximize the usefulness of the LF when assessing second-language readers, one needs to understand the differences between text in English and other languages. A study of these differences, on a language by language basis (perhaps starting with Spanish) might be recommended. The study might start with the two components of the LF, sentence length and word frequency, perhaps also considering other factors. The study would be designed in consultation with people who have experience in these areas. Another possibility is to undertake analyses of assessments of populations of second-language readers, assuming that a sufficient number can be identified. These groups should be assigned separately, so that comparisons can be made to study whether or not the LF applies equally to native and non-native speakers of English. If these studies cannot be conducted, then disclaimers should be inserted in publications, saying that item difficulties are likely to be different for those not yet proficient in English.

- **Indexing the corpus of Lexiled texts by grade level and subject area.** The desirability of doing this has been mentioned earlier.

- **Exploring a modification of the Lexile Framework (LF) scale.** The LF is constructed on the basis of the assumption that the "optimal" reading level for a student occurs when they can comprehend correctly around 75% of the passages they read. This assumption differs from the standard psychometric assumption used in the design of assessment instruments: the optimal scaling point for an item (say, a passage of text in a reading assessment) occurs when approximately 50% of the test

takers can pass it. While noting that these figures are approximate and it is not straightforward to relate them directly, rescaling the LF to more closely fit psychometric standards is likely to make items move toward being administered at higher ages or grade levels. It is worth considering for NCES to explore what consequences such a change might have for the correspondence of Lexile assessment scores with other assessment measures or for the scale itself.

- **Cross-validating NAEP and NAAL psychometry**. One way to establish validity is through cross-validation: comparing reading difficulty levels for prose passages estimated from Lexile numbers with NAEP or NAAL item passing percentages. Suppose, for instance that a fourth-grade NAEP reading assessment item is assigned a reading difficulty level of 650 Lexiles. This amounts to a prediction of what percentage of fourth graders could pass the item (general fourth-grade student Lexile levels range from 620 to 800); this prediction can be compared to percentage of test-takers actually passing the item.

- **Associating books and texts with NAEP reading achievement levels**. NAEP reading assessment results group students into Basic, Proficient, and Advanced categories. Communicating what these categories mean to the general public is sometimes challenging. Lexile believers hope that the LF could be used to link, say, the "below Proficient" fourth-grade NAEP reader with a set of books that they would find optimal to read, which in turn would better convey to the public what such a reading performance level means. The measures of reading comprehension on which the LF is based, however, represent only one of a number of comprehension tasks. It is possible that other NAEP comprehension tasks based on the NAEP reading framework will perform differently.

- It would be helpful to **elucidate the features of linguistic complexity for passages at different difficulty levels** e.g., at 600 and 1200 Lexiles. But since the LF components of sentence length and word frequency are 'proxies' for complexity, the LF in itself cannot supply such information. Additional studies, more fine-grained than the LF assessments, would be needed to determine whether there are consistent differences between the linguistic features at different levels of difficulty. It's not clear that there are such differences. This would be a project for the future. If it were successful, it might complement other aspects of the NAEP reading framework, which addresses such matters as interpretation and critical analysis.

4. Summary and conclusions

The panel affirmed the value of both sentence length and word frequency as overall measures of semantic and syntactic complexity (although participants diverged on whether these constructs were best viewed as proxies or as direct measures).

Possible applications include:

- Selection of candidate reading passages for assessment items.
- Obtaining Lexile values for text passages in use in existing assessments.
- Assessment item and assessment comparisons.
- Contrasting performance levels with Lexile scores.
- Assessing the reading difficulty level of NAEP and NAAL questions.

The LF has limited uses for NCES assessments, however, because of linguistic and psychometric issues, and there are therefore certain limitations on how the LF might be applied.

Issues raised by the panel that are at once matters of concern and possible areas for future research can be grouped into three categories: semantic theory; syntactic theory; and psychometric issues.

Semantic theory issues:
- Words in the middle frequency ranges.
- The separation of ideational complexity from rarity in the definition of semantic complexity.
- Exclusion of novel or technical terms, function words, proper nouns, or alternative forms.
- Effect of pragmatics and discourse features.

Syntactic theory issues:
- Adequacy of sentence length as a proxy for syntactic complexity.
- Relative contribution of syntactic and semantic complexity.

Psychometric issues:
- Explaining within-ensemble variation.
- Assessing the generalizability of the LF to other genres.
- Applicability to the reading assessment of second-language readers.
- Indexing the corpus of Lexiled texts.
- Modifying the psychometric grounds of the LF.
- Cross-validating NAEP and NAAL psychometry.
- Associating books and texts with NAEP reading performance levels.
- Elucidating linguistic features of passages at various difficulty levels.

Attachment 1. Lexile Framework Assessment Panel Members and Invited Observers

Panel Members:

Marilyn Jager Adams (Ph.D., Brown University) is with Bolt Beranek and Newman (now BBN Technologies), and is also a Research Associate at the Harvard University Graduate School of Education. Her work includes many chapters and journal articles on issues of cognition and education as well as the book, *Beginning to read: Thinking and learning about print*, written on behalf of the U.S. Secretary of Education.

Ms. Adams has also been active in efforts to translate research to practice. She chaired the Planning Committee and was a member of the Study Committee for the National Academy of Science's report, *Preventing Reading Difficulties in Young Children.* She has also worked closely with several states and districts on developing standards, policy, and staff development strategies in the area of reading. Ms. Adams is principal author of *Fox in a Box*, a kit of informal reading and writing assessments for K-2 classroom teachers and tutors, and of three empirically validated instructional resources including Open Court's *Collections for Young Scholars,* a classroom reading program; *Odyssey: A curriculum for thinking* for middle-grade classrooms; and *Phonemic awareness in young children: A classroom curriculum* for preschool, kindergarten, and special needs children. madams@bbn.com

Michael L. Kamil (Ph.D., University of Wisconsin – Madison) is Professor of Education at Stanford University School of Education. He is a member of the Psychological Studies in Education Committee and is on the faculty of the Learning, Design, and Technology Program. His research explores the effects of computer technologies on literacy and the acquisition of literacy in both first and second languages. Another line of research focuses on the uses of expository text for reading instruction in first and second grade.

He has been editor of Reading Research Quarterly, Journal of Reading Behavior, and The Yearbook of The National Reading Conference. He was a member of the National Reading Panel, producing a synthesis of instructional research in reading. For the Panel, he chaired the subgroups working on comprehension, technology, and teacher education. He is currently a member of the Rand Corporation Reading Study Group, developing a long-term reading research agenda for the Office of Educational Research Improvement. mkamil@stanford.edu

Richard Larson (Ph.D., University of Wisconsin – Madison) is Professor and Chair of the Department of Linguistics, SUNY – Stony Brook. His research has examined a wide variety of topics in syntax and semantics, including relative and adverbial clauses, NP adverbs, disjunctions, prepositional phrases, double objects and clausal complements. He has published numerous research articles and is co-author of a recent textbook on formal semantic theory, which frames the subject matter from a Chomskyan viewpoint. His current research is in adjectival semantics and event theory. rlarson@semlab1.sbs.sunysb.edu

Carlota Smith (Ph.D., University of Pennsylvania) is Centennial Professor of Linguistics, University of Texas at Austin. Her current interests center on the interface of syntax, semantics, and pragmatics in text structure. She has published extensively in several areas including relative clauses, pragmatics and pronoun systems, tense, aspect, the notion of topic, lexical structure in Navajo, and language acquisition. She is the author of a book, "The Parameter of Aspect" and is writing another on local structure in texts. She was director of the Center for Cognitive Science and organizes cognitive science courses at UT. She is a member of the Usage Panel for the American Heritage Dictionary. carlotasmith@mail.utexas.edu

Heide Spruck Wrigley (Ph.D., University of Southern California) is the senior researcher for Language, Literacy and Learning at Aguirre International in San Mateo, CA. Dr. Wrigley is the key content expert for a national study on adult ESL literacy funded by the U.S. Department of Education, conducted jointly by the American Institutes for Research and Aguirre International. She has extensive experience in studying various areas of adult education, with a special emphasis on minority populations. Her international work has included conducting an evaluation for the Peace Corps for its English as a Second Language (ESL) program in Poland, technical assistance to teacher training institutions in the People's Republic of China (with a focus on EFL), and an evaluation of teacher training in English as a Foreign Language in Egypt. She has been key in two bi-national studies on the U.S. Mexico border.

Ms. Wrigley has also been involved in various statewide and national efforts related to the education of language minority adults. She has written extensively on issues related to adult ESL, including the need to develop assessments that capture what learners new to literacy are able to do. She is co-author of "Assessing Success in Family Literacy and Adult ESL" and primary author of "Bringing Literacy to Life: Issues and Options in Adult ESL Literacy." hwrigley@aiweb.com

Invited Observers and NCES and ESSI Staff

Don S. Burdick is Associate Professor of Statistics and Mathematics in the Department of Mathematics at Duke University. Professor Burdick has also been involved with MetaMetrics since its inception, as a consultant on the statistical methodology which underlies the Lexile Framework. His efforts during the development of the Framework have resulted in several publications, including a doctoral dissertation completed under his supervision.

Patricia L. Donahue (M.A., M.Litt., Middlebury College, M.Phil., Columbia University) is a program administrator with the National Assessment of Educational Progress (NAEP) at Educational Testing Service, where she is responsible for overseeing test development and scoring activities for the NAEP reading assessment. She has worked as an author on numerous NAEP reports, including *The Nation's Report Card: Fourth-Grade Reading 2000, The NAEP 1998 Reading Report Card for the Nation and the States,* and *Students Selecting Stories: The Effects of Choice in Reading Assessment.*

Elizabeth Greenberg is Senior Research Analyst at the American Institutes for Research (AIR). As deputy project director of the National Assessment of Adult Literacy (NAAL), she is interested in comparing different readability frameworks to assess which is most appropriate for linking between the 1992 National Adult Literacy Survey and the 2002 NAAL. In addition to her responsibilities on the NAAL, Ms. Greenberg works on background questionnaire design, analysis, and reporting for the National Assessment of Educational Progress (NAEP). Her research interests focus on using survey and assessment data to inform the discussion of public policy issues related to education. She has a Masters degree in political science and advanced to candidacy for a Ph.D. in political science at the University of California at Berkeley. Prior to being employed at AIR, she worked at the Economic Research Service of the U.S. Department of Agriculture, where she conducted research on rural schools and rural labor markets.

Andrew Kolstad was the government project officer responsible for the 1992 National Adult Literacy Survey and co-authored "Adult Literacy in America," the first in a series of reports based on that survey. Dr. Kolstad works as a mathematical statistician for the National Center for Education Statistics, which is a part of the U.S. Department of Education's Office of Educational Research and Improvement. He now serves as senior technical advisor to the National Assessment of Educational Progress and contributed to the survey design of the 2002 National Assessment of Adult Literacy. Dr. Kolstad formerly worked on longitudinal studies at NCES, including the National Longitudinal Study of the High School Class of 1972, the High School and Beyond study, and the National Education Longitudinal Study of 1988. Dr. Kolstad earned his professional degrees in sociology from Columbia College (B.A., 1967) and Stanford University (M.A. 1971, Ph.D., 1976).

Mark Kutner is Deputy Director of the Pelavin Research Center of the American Institutes for Research. He is directing the NCES-sponsored National Assessment of Adult Literacy, which will assess the literacy skills of adults in the United States in 2002.

Terry Salinger is Managing Associate at the Pelavin Research Center of the American Institutes for Research. She is responsible for corporate management and supervision of assessment projects and for research on assessment and literacy issues and policy. More specific responsibilities include developing testing plans for standards-based assessment; early learning assessment; coordination of teacher assessment advisory committees; focus groups on assessment issues; policy and instructional issues in literacy instruction; teachers professional development; supervision and training of staff. Recent projects have focused on large-scale assessment in kindergarten to grade 12 classrooms literacy instruction in kindergarten to grade 3, and adult literacy assessment.

Ramsay Selden is Director of the Education Statistics Services Institute (ESSI) of the American Institutes for Research. The ESSI was created by the National Center for Education Statistics to support its efforts, essentially functioning as an extension of NCES. ESSI develops new data collections in areas that are not addressed well in current educational statistics, helps analyze and disseminate NCES findings, and supports NCES programs in functions such as planning, evaluation, project management, and training.

ESSI is also intended to help enhance the field of education statistics generally. Before joining AIR to direct the ESSI, Selden directed the assessment programs of the Council of Chief State School Officers.

A. Jackson Stenner is co-founder and CEO of MetaMetrics, Inc. MetaMetrics is a privately held corporation which specializes in research and development in the field of education. He has been Principal Investigator on a series of five grants from the National Institute of Health (1984 – 1996) dealing with the measurement of literacy. He holds a Ph.D. degree from Duke University in Durham, North Carolina. He also holds Bachelors degrees in Psychology and Education from the University of Missouri at St. Louis. He is the author of numerous professional papers, chapters, and books on measurement and psychometrics.

Suzanne Triplett is director of Development and Operations for the National Assessment of Educational Progress (NAEP)in the National Center for Educational Statistics (NCES). Her responsibilities include the sampling and data collection components of NAEP, the NAEP web site, and state relations. Before joining NAEP, Triplett headed the North Carolina state testing and accountability program, the first state program to incorporate Lexiles into their student reporting system.

Sheida White is Project Officer for the National Assessment of Adult Literacy in the Assessment Division of the National Center for Education Statistics. For the past 10 years, she has monitored the NAEP Reading and Writing assessments. Before coming to NCES, she was a reading researcher in Hawaii where she worked directly with teachers and students. She has a Masters degree in English as a Second Language and a Ph.D. in sociolinguistics from Georgetown University. She has published articles in reading and literacy journals.

John Clement (Convenor) is Principal Research Analyst, Education Statistics Services Institute of the American Institutes for Research. His academic interests merge statistics, mathematics, and educational and evaluation research. He has a Ph.D. in Anthropology from the University of California at Irvine. Prior to joining AIR, John worked in research and policy analysis arenas for the National Academy of Sciences, the National Science Foundation, and EDUCOM. He has a long-term interest in the use of the Internet by educators, and was the founding Executive Director of the Consortium for School Networking (CoSN). He serves as liaison to the National Forum on Education Statistics, and has undertaken a number of projects in survey development and conceptualization in both K-12 and postsecondary education.

Attachment 2. Invited Papers.

On the Lexile Framework
M.J.Adams, BBN Technologies

1. Is the theoretical basis for the Lexile Framework logically solid? Does it correspond to a widely-accepted view among researchers of the syntactic and semantic events underlying reading?

Across the twentieth century, dozens of different formulae have been devised to estimate the readability of texts (see Klare, 1984). Associated with these efforts, many textual dimensions that might affect reading ease or difficulty have been considered, including for example illustrations, format, subject matter, abstractness, and interestingness. Nevertheless, of factors investigated, empirical investigations have repeatedly affirmed that in proper combination word familiarity and sentence complexity are especially good predictors of text comprehensibility and reading difficulty.

In terms of construct validity, these two factors, word familiarity and sentence complexity, are held to reflect the semantic and syntactic demands of a text, respectively. The reigning view within cognitive science is that language, whether written or spoken, conveys meaning through consensual reference. More specifically, written text conveys its author's thoughts only by means of directing readers to recall or create similar thoughts from their own prior knowledge. The semantic or conceptual knowledge that the author wishes to evoke in the readers' minds is conveyed principally through her or his choice of content words. Meanwhile, the author's syntax serves to constrain or specify the intended interrelations among those fields of meaning. The predictive strength of word familiarity and syntactic complexity follows directly: Even the best written text can be comprehensible to its readers only to the extent that they can apprehend the significance of the words and the syntax on the page.

That having been said, readability formulae do not gauge the semantic and syntactic complexity of a text directly but only estimate it by means of some correlate or indicator variable that can be quantified. Although, over the years, a number of more sophisticated methods have been devised, the most frequently used measure of syntactic complexity in readability formulae is sentence length (or, sometimes, clause length) as measured in number of words. The widespread use of this metric is owed largely to the fact that the number of words in a sentence can be counted easily, reliably and, now, automatically. In any case, sentence length has been shown time and again to correlate strongly with interpretive ease and difficulty of text. Moreover, an "on average" relationship between sentence length and syntactic complexity is logically compelling: The longer the sentence, the greater the number of concepts in reference; the greater the number of concepts in reference, the greater the potential number or complexity of the interrelations that must be understood between and among them. In common with classic readability formulae, the Lexile Framework has adopted sentence length as its index of syntactic complexity.

In the same spirit, readability formulae most often use indices of word familiarity to gauge the semantic demands of text. In many older readability formulae, word

familiarity was estimated from the lengths of words, as counted in letters or syllables. The construct validity of this measure is bolstered by a variety of considerations, including Zipf's (1935) well-established law that in every language common words tend to be short. Beyond that, word length proves to be a statistically decent proxy for semantic complexity. It is nevertheless a proxy, and with this in mind, some formulae have included a various "adjustment" factors, such as not counting repeated words or proper nouns, in the effort to increase sensitivity.

As enabled by computational advances, the Lexile framework estimates the familiarity of words from the (dispersion adjusted) frequency with which they were found to occur in a highly respected study of school children's reading materials (Carroll, Davies, and Richman, 1971). In terms of construct validity, the justification for this measure is straightforward: The less often a word appears in children's reading materials, the less likely its familiarity. Through empirical investigation, moreover, the Lexile developers found word frequency to absorb more of the semantic component of the text difficulty space than any of the measures with which they competed it, including word length (Stenner & Burdick, 1997).

In all, as reviewed in their technical literature, the developers of the Lexile Framework have chosen their measures and developed their formula with methodical care and discipline. In the interest of validation and calibration, they have also evaluated the performance of the Framework with thousands of children across a wide range of texts and tests. In terms of psychometrics, the Rasch scaling of the readability scores is a definite plus. In terms of ease of use and extendability, so too is the fact that the system is computer-based; for any text, both metrics can be quantified reliably and, with computers, quickly and inexpensively as well.

2. The framework's assessments of text comprehension difficulty rest on two measures. The syntactic measure is sentence length; the semantic one, word frequency in a reference corpus.

a. To what extent is each of these measures valid proxies for underlying theoretical constructs of central importance to reading comprehension?

Again, as discussed above, Lexile's use of word frequency and sentence length for estimating the semantic and syntactic demands of texts is logically justifiable and historically grounded. However, both of these measures are exactly and only proxies for the underlying variables of interest and, as such, their validity is necessarily but correlational. The extent to which they are valid proxies therefore depends on the granularity of the application. That is, just as the power of these measures is statistical, and so are their hazards.

In proper combination and given an ample and representative sample of texts, word frequency and sentence length have been shown to capture a majority of comprehensibility variance. In a study of 11 basal reading series, for example, the Lexile readability indices were shown to align with the actual, developmental order of the units (K-6?) with a raw correlation of 0.86 (0.97 given corrections for range restriction). On the other hand, the smaller or shorter the text sample of interest, the greater the risk.

According to Wright and Stenner (1998), the root mean square measurement error or standard deviation for one test item is about 172 Lexiles. In comparison, the distance in Lexile units from first- to second-grade books is approximately 200 Lexile units, shrinking to approximately 50 units at secondary school levels.

Following are some of the factors that contribute to the estimation error on each dimension.

Word Frequency and Semantic Accessibility

adore	ahhh	amphibole	apogee	authorship
Both	Bysshe	cacophony	diaper	Disraeli
dissident	drat	Evansville	extirpation	fakes
gag	glaciated	gorse	grouchy	hairbrush
Hippocratic	hovel	hygroscope	ingots	interstitial
irascible	Kristy	leviathan	loquacious	Louvre
lunchbox	magnanimous	marmot	Mason-Dixon	mastication
microwave	Mohawks	mommy	Myself	nacreous
namastey	nipping	Oz	pitch	pomegranate
psittacosis	puppyhood	purloin	Robby	Robot
sallow	salubrious	saprophytic	Send	Sinatra
smock	Sophocles	Texan	tooting	truck's
wallets	weekdays	wristwatch	x-axis	zippers

Above is a selection of words that Carroll et al (1971) encountered just twice in their sample of 5,000,000 words of school children's text. Despite the fact that they are of equal sampled frequency, a quick read through these words reveals a relatively dramatic range of referential sophistication (e.g., grouchy, smock, diaper, zippers vs. irascible, namastey, saprophytic, Bysshe).

Part of the problem is, of course, sampling error. An unfortunate fact of life is that sampling error unavoidably increases as the true frequency of items in a population decreases. Still more unfortunate, then, is the fact that the word frequency distribution is extremely skewed. For example, although Carroll et al's corpus included nearly 87,000 different word types, 75% of the running text was accounted for by just 1,000 of those types and 90% of the text was accounted for by just 5000 different types. Precisely because of this skewness, it is the ability to accurately discern the true frequencies of relatively uncommon words that matters most in the quest to measure the semantic complexity of texts. At the same time, however, this extreme skewness means that a very, very large sample is needed to accurately estimate the relative frequencies for even a moderate literary vocabulary.

To be sure, misestimates of word frequency due to sampling error could be reduced by processing ever-larger collections of text. However, not even an infinite sample of text could correct for the more basic problem, that the frequency with which words appear in text is only a proxy for the semantic variables of core interest. At the root of this problem is the fact that word frequency and semantic accessibility hold an

asymmetrical relationship to one another. That is, to the extent that the meaning of a word is complex or esoteric, its true frequency of occurrence in text will be low; on the other hand, the infrequency with which a word appears in print is not a good predictor of its semantic sophistication.

Again, a glance at the words in the table above reveals many of the issues. First, for the benefit of the computer, Carroll et al defined a word as a specific graphic string bounded by spaces. Because of this, small changes in presentation, such as capitalization, pluralization, and other basic inflections, may reduce a word's frequency far more than they reduce its semantic accessibility (e.g., truck's, tooting, Myself, weekdays). Second, many of the capitalized words in the corpus are proper nouns. Given that the very purpose of proper nouns is to (uniquely) distinguish its referent, all are relatively infrequent. Nevertheless, some of these names suggest a greater degree of literary sophistication or maturity than others (e.g., compare Oz, Sinatra, Sophocles). Further, while some of the proper nouns are likely just names (e.g., Kristy, Evansville), the contextual import of others quite likely presumes relatively specific and/or extensive background knowledge (e.g., Louvre, Hippocratic, Mason-Dixon). Third, because the database for the Carroll et al sample was written text, the counted frequencies underestimate the semantic familiarity of any word that is more common in the spoken than the written language of schoolchildren (e.g, drat, ahhh, lunchbox, mommy, hairbrush). Fourth, even among words that would be rare by any count, some are semantically simpler than others (e.g., pomegranate, marmot vs. amphibole, leviathan, apogee).

Finally, and perhaps most important, it is not the reader's familiarity with the printed word that is of interest in analyses of semantic accessibility; rather, it is her or his familiarity with its contextually complete and appropriate meaning and usage. A recent study by Biemiller and Slonim (2001, in press) affirms that the pace of vocabulary growth ranges widely. However, his data also indicate that the order in which the meanings of words are acquired is relatively fixed across children. Of special interest in the present context, Biemiller and Slonim found virtually no correlation between the order of acquisition of word meanings and the printed frequency of their tokens. The reason for this null correlation, he surmises, related to the variations in the meanings and usages of words. Children may be quite familiar with one meaning or usage of a word, but effectively clueless with another. For example, a child may understand "catchers squat" but not "the squat, yellow-brick house," or the child may understand "legs and arms" but not "soldiers at arms." Inasmuch as counts of word frequency, such as Carroll et al.'s, are sensitive only to graphic identity, this is a critical finding with respect to the ultimate potential of such tools for gauging the semantic accessibility of a text.

Again, given well-written text, the length of a sentence should correlate with its syntactic demands on the reader. Again, however, sentence length is a proxy for syntactic complexity. Its validity is correlational, holding on average across ample samples of well-written text but easily violated on a sentence by sentence basis.

By concatenating clauses, for example, one can create a sentence that is indefinitely long, but syntactically easy to parse, e.g., "The grocery store has peaches, and it has oranges, and it has bread, and it has cookies, and it has…" Further, comprehensibility may differ considerably even for sentences of essentially the same length and same content, e.g., "The dog worried the cat that chased the rat that ate the malt" vs. "The rat that the cat that

the dog worried chased ate the malt." And reducing the lengths of sentences by eliminating entailed words and connectors often make them harder rather than easier to understand, e.g., "The rat the cat the dog worried chased ate the malt." Indeed, such violations of the complexity/length rule have been richly documented in the linguistics literature.

Beyond such violations of the complexity/length rule, are issues of the subtlety or precision of the syntactic requirements of a text. In terms of classroom implications, for example, the following two sentences differ significantly: "Children who make good academic progress enjoy school" vs. "Children who enjoy school make good academic progress." More generally, in a way that is inappropriate in conversational discourse and rarely required for narrative discourse, scientific, mathematical, and logical texts require active respect of the relations formally specified by their syntax even where such conflicts with the reader's expectations.

Finally, because it is only gradually that children's ability to understand a given syntactic structure becomes context-independent, the active syntactic demands of a text must effectively increase as the familiarity of its wording and topic decreases. Note, too, that this must be especially troublesome for children with limited English proficiency.

b. Are [word frequency and sentence length], in combination, sufficient to account for the major factors that theories of reading claim should be present?

My vote for most important neglected factor goes to intersentential meaning and cohesion or, as I like to think about it, to the argument structure of text. Given that any sentence of a passage that does not contribute substantively to its message or impact ought to be deleted, it must also be the case that the significance of every sentence of a well-written passage extents substantively beyond its boundaries. If such between-sentence factors are also part of the comprehension challenge, then they are doubly neglected by readability formulae since so many of the "indexicals" (this, that, instead, she, it, neither, before,….) are very high frequency words and thus counted as extra easy.

I hasten to add that even if this concern is apt, it pertains to the Lexile Framework no more and no less than to other readability formulae. Further, like readability formulae, the preponderance of hard-nosed research on readability is focused on words and within-sentence factors. In other words, although my bias here is strong, it remains relatively unfounded (and though I have tried, I have never succeeded in persuading any agency to support research on this hypothesis).

3. Are there other theoretical frameworks, constructs or measures in the published literature that can be contrasted with the Lexile Framework?

To my knowledge all efforts to quantify the readability of text that are both practically efficient and usefully powerful have in one way or another used measures of word familiarity and within-sentence word counts to estimate semantic and syntactic demands. Of these, the Lexile Framework is exceptional in the psychometric care with which it has been developed; the extent of its formal validation with different populations

of texts, test, and children; its automation; and in its developers continual quest to improve it.

Again, efforts to identify the factors that make text easier and harder to read and comprehend have been myriad. Indeed, broadly defined, this quest has been perhaps the major preoccupation of cognitive psychology. Although this work helps us to understand the limitations of word frequency and sentence length for estimating the complexity of text, it has not yet yielded alternatives that are superior in usability or power. With computational advances, natural language applications may be the best bets toward this end, and I was happy to learn that Metametrics is exploring this technology.

In their discussion of the measurement error of the Lexile Framework, Wright and Stenner (1997) argue firmly against its use for purposes of gauging the readability of individual test passages, and the Panel concurred. By the same logic, the Panel warned against use of the Lexile Framework for equilibrating standardized reading comprehension tests. As a cynical argument against any such strategy, it would be far too easy for test-makers to choose passages that were far easier than suggested by their Lexile scores – and, in this era of high-stakes testing, far too tempting.

For proper interpretation, Wright and Stenner argue, the framework should be applied to representative samples of text. Consonant with this, the Panel discussed two ways in which the framework might be used to improve the calibration and informativeness of the NCES assessments.

The Panel's first suggestion was to use the framework as a device for generating candidate passage sets, which could then be culled and evaluated by other methods to select the final test materials. For this purpose, Lexile's large, electronically searchable text inventory might also be of significant benefit.

The Panel's second suggestion was that NCES consider using the Lexile Framework as a tool for matrix sampling the passages in its tests. Minimally, this might serve to reduce any bias or error associated with the tested texts and response probes. Involving a larger student population, it also offers potential for increasing the kinds of information returned from the assessment.

References

Biemiller, A., & Slonim N. (2001, in press). Estimating root word vocabulary growth in normative populations: Evidence for a common sequence of vocabulary acquisition. Journal of Educational Psychology.

Klare, G. R. (1984). Readability. In P. D. Pearson, R. Barr, M. L. Kamil, and P. Mosenthal (eds.), Handbook of reading research, 681-744. New York: Longman.

Zipf, G.K. (1935). The psycho-biology of language. Boston: Houghton Mifflin.

Carroll, J.B., Davies, P., & Richman, B. (1971). Word frequency book. Boston: Houghton Mifflin.

Stenner, A. J. & Burdick, D. S. (1997). The objective measurement of reading comprehension: In response to technical questions raised by the California Department of Education Technical Study Group. January 3, 1997.

Wright, B. D., & Stenner, A. J. (1998). Readability and reading ability. Presentation to Australian Council on Education Research (ACER), June 1998.

Comments on Lexile Framework
Michael L. Kamil, Stanford University

The basic notion of having a metric to match reading ability and difficulty of materials for reading is certainly one that has much to recommend it. It is an idea with a long history, dating at least back as far as the 1920s with the readability work of Lively and Pressey (1923) and Vogel and Washburne (1928). Readability is one dimension of what is needed to match readers and materials. An excellent summary of work on readability can be found in Klare (1984).

One difficulty in calculating readability is that it was time intensive. The development of computer analysis of text eliminated much of the time element. However, the variables that comprise a readability calculation must be relatively easy for a computer to determine. For example, sentence length is easy to calculate, but word difficulty (as, for example, in the Dale-Chall formula) may not be. To solve these problems, most computerized version of readability have resorted to proxies or look-up lists.
The Lexile Framework measures depend on the mean sentence length and the mean word frequencies. The measure of word frequency is taken from the Carroll, Davies, and Richman (1971) list that is based on a corpus drawn from school materials. These measures are calibrated through a Rasch analysis and subsequently converted into a Lexile scale ranging from (approximately) 200 to 1200. This is done for a range of texts from primer to encyclopedic materials. Standardized tests were also subjected to a similar analysis. (Materials can exceed these scale limits, however.)
Similarly, Lexile scores can be assigned to performances on tests, based on performance on those tests. The assumed criterion is that students should be able to comprehend 75% of material having the same Lexile score as their performance.

Three questions are addressed in these comments:

1. Is the theoretical basis for the Lexile Framework solid?
2. To what extent are the two measures (word frequency and sentence length) valid proxies for underlying theoretical constructs of central importance to reading comprehension?
3. Are there other theoretical frameworks, constructs, or measures in the published literature that can be contrasted with the Lexile Framework?
4.

The following remarks are grouped in three general categories that affect the interpretation of metrics of reading difficulty and Lexile measures in particular.

Locus Of Text Difficulty

There are three loci for text difficulty. One is in the reader's head; the other is on the page. Lexile analysis accounts for latter while ignoring the former. The third locus is the transaction and between reader and author. That is, the difficulty of text partly resides in the reader's head and partly in the text itself. Lexile also ignores this, focusing, as noted above on the text.

Lexile Framework claims to be appropriate for any text. However, the development of the Framework based on test items may limit the external validity of the measure. In short, the samples of text used in developing Lexile scores are limited, and are probably somewhat different from what students would encounter in their "normal" reading. There are two parts to this issue. The first is that the samples of text that appear on tests may not be sufficiently interesting for readers to show difficulty effects. That is, they may simply not be sufficiently motivated to spend the effort necessary to perform at their highest capacity.

A second dimension on which these texts seem to differ is in genre. The overwhelming genre to which students are exposed in school is some form of narrative. Estimates suggest that from 75-95% of all material in basal readers is of this text structure. Duke (2000) estimates that students encounter only enough expository information to account for 3.6 minutes of reading per day. The difficulty is that standardized tests of reading comprehension are typically weighted evenly across these text forms. Thus the sample of texts on which Lexile analysis is based is not representative of the materials available to a student reader.

The example given in 2.7 (Stenner & Burdick, 1997) seems to take a position contrary to both contemporary practice and evidence. They suggest that one learns to read by reading. Another claim is that they need to read easy materials. The optimal difficulty of materials is an intensively-debated issue. It is often the case that students achieve higher scores when instructed in more difficult materials (e.g., Barr, 1989; Kamil and Rauscher, 1990) However, the issue is far from definitively settled. Thus, it is not clear that the student who wished to read the Chris Evert biography would not have expended more energy than in reading something else of appropriate Lexile difficulty. Further, the Report of the National Reading Panel (2000), for example, found no evidence that reading, of itself, led to higher achievement levels. While there was very little evidence on this matter, what evidence the Panel found was negative. Instruction plays an important role in learning to read. Practice is also important, but is not sufficient for learning to read.

Finally, this formulation does not account for specific interest, motivation, or technical prior knowledge. A student who is familiar with Chris Evert before reading a difficult book about her may have the specific vocabulary knowledge to make the text, effectively, much less difficult. For that same student, a book of the same difficulty level about Herbert Hoover would probably not be as readable.

A related construct is the purpose for reading. Not every text is read for the same purpose. The importance of that purpose for the reader may determine reading effectiveness. In turn, this may limit the accuracy of Lexile scores, since they are based on a subset of reading behaviors with an identifiable bias.

The final issue of text difficulty applies to a new text form-- electronic and multimedia text. This type of text has become more prevalent in classrooms, at work, and at home. While it has always been the case that readers have had to combine textual and non-

textual information in comprehension, electronic and multimedia text make that skill even more important. Given the increasing reliance on this type of text, it is important to consider whether or not it would also be subject to the same difficulty metrics. The interactive nature of electronic text can provide for support for difficult elements in the text. This type of reader support might make the concept of difficulty irrelevant, or at least VERY different, for such texts. However, some current research suggests that, even without this capability, very different metrics are needed for assessing the difficulty of electronic and multimedia text.

Utility for Decision-Making

Whether or not Lexile measures (or any other measure of text difficulty) are useful is a crucial, although separate, pragmatic question. In fact, the range of readable text for students is large, depending on the support available to the reader. Put another way, the utility of any measure of text difficulty is the success in matching text to student abilities in optimal ways for instruction or for independent reading.

The difficulty with the Lexile measure is that it is probably correct, on the average. Instructional decisions, however, must be made individually, not for groups. What is needed is to validate the use of Lexile scores in practice. The question is whether these scores be used to match texts to students in a way that produces greater learning. Corollary to this is whether or not the costs and benefits justify this procedure.
It is not clear from the available documents how Lexile measures might be used in the classroom. In fact, the examples suggest that there is a far greater range of uses than those in school. There is a suggestion that parents could use these scores to select materials for students. Whether or not this would be a reasonable procedure should be tested empirically. That is, do such parental decisions have any effect on reading or reading ability?

Statistical Dimensions of the Lexile Framework

Obviously, a great deal of work has gone into the development of the Lexile Framework. It assumes that the construct of text difficulty can be represented in a single measure that is a proxy for all of the variables of text difficulty. In fact, as noted above, motivation, interest, and prior knowledge are important sources of variation in text difficulty. They might represent one or more dimensions that are ignored in a single dimension model. The procedures used to establish the Lexile scores clearly disregard these variables. There was no attempt to determine whether the se other variables might represented significant sources of variation. Thus, the model may fit to a subset of available data on reading, but not to others. If this were the case, its external validity would be limited to texts similar to those in the subset.

While the Lexile model does fit some data, it is not clear from the papers whether there are sources of measurement error unaccounted for. The linking error analysis is interesting, but represents only a single source of error (and perhaps one which is not all that important). The standard errors of measurement for the individual reading tests appear to be left unanalyzed. This source of error may have potentially greater effects on the ultimate uses of Lexile scores. That is, given that many standardized tests have

relatively large standard errors of measurement, reliance on those scores will produce less reliable derivative measures.

In addition, the confidence interval for any Lexile score is important. This is given scant treatment in the papers reviewed. The standard deviation seem to be about 118 points (for the data presented in Stenner and Burdick (1997, p.27). This seems to be about the equivalent of a grade level. Whether or not measures with this size error are usable should be tested empirically, rather than statistically.

The corpus of text used for performance measures is that found on tests of reading. It is important for these tests that the passages are chosen to be psychometrically valid in separating students. However, these passages are typically emotionally and motivationally neutral, so as not to give particular students an advantage over others. Consequently, such passages may be unusual samples of text that reflect only one dimension of performance in reading comprehension. Thus the statistical model may hold for the corpus involved, but not for text more generally.

One more concern: The Lexile type items used for testing are reminiscent of cloze items, although there are important difference. Since cloze items have been shown to tap only low level comprehension processes, the same may be true of these items. If that is the case, Lexile may "work" for many of the same reasons that cloze appears to work. Foremost among these reasons is probably the case that much of comprehension is low level. The difficulty is that the measure then fails to capture the full range of reading comprehension.

Conclusion

The answers to the three questions can be summarized as follows:

1. The theoretical basis for the Lexile Framework is as solid as any of many other readability formulations. Since it attempts to produce a scale that can account for reader performance, it represents at least some improvement over traditional measures. However, the concerns above show that Lexile scores are not free from difficulties.

2. Again, as suggested above, there are many constructs in reading comprehension that have been omitted. Some candidates for important dimensions seem to be (at least) reader motivation, reader interest, and reader knowledge. The notion of purpose in reading is excluded in the Lexile Framework. This is a serious oversight because of the dramatic effects that purpose can have on reading. Of perhaps somewhat lesser importance is the notion of sentence length. Work by Pearson, for example, has shown that shorter sentences do not always yield better comprehension.

3. There is a small body of work on text usability and learnability (Armbruster and Anderson, Irwin and Davis, to name a couple). These have the advantage of being much more useful than single dimension indicators. The disadvantage is that the analyses required are far more difficult than conventional readabilities of Lexile scores.

An alternative procedure for assessing readability that seemed to have some promise is that of using anchor passages (e.g. Singer or Carver). In this procedure, sample anchors are used to compare passages to find which is the closed in difficulty to a target passage. There is limited research on this, but perhaps it is time to begin to explore it anew.
A major unresolved issue is the way in which reading comprehension is to be conceptualized. If reading is conceived of as a transmission process, where information is transferred from the page to the reader, Lexile analysis is certainly usable--and potentially valuable. However, if it is conceptualized as a multidimensional process, then Lexile analysis will be appropriate for only a subset of reading contexts.

Empirical work can help to answer these questions by determining how well decisions made on the basis of Lexile scores "work" in instructional contexts. Since there are few instructional alternatives there is not necessarily a great need for fine-grained analysis of texts. There may be more cost effective ways of arriving at a decision than Lexile analysis.

The concluding quote from Klare (1984) is interesting in its cautions and suggestions for what needs to be done in readability research. Lexile analysis could benefit from the application of some of these suggestions.

> We clearly <u>can</u> go too far with readability findings, both in attempts at predicting and producing readable writing, but we need not. Arguable issues remain in the areas of criteria and grade-level assignments and in application to the preparation of reading materials. Formulas selected and used properly can be helpful as screening devices, but this does not mean they can at the same time serve as guides for readable writing. Humans and language are too complex to expect such simple cause-effect relationships. Research has now clearly begun, however, to move in the direction of understanding <u>when</u> and <u>how</u> changes in readability can work for the individual reader. (Klare, 1984, p. 731)

Report on the Lexile Framework
Richard K. Larson (Linguistics, SUNY - Stony Brook)

The following is my summary assessment of the Lexile Framework, structured around the following three questions that ESSI panel members were asked to address:
• Is the Lexile Framework logically solid? Does it correspond to a widely-accepted view among researchers of the syntactic and semantic events underlying reading?
• To what extent are sentence length and word frequency (measured against a standard reference corpus) valid proxies for syntactic and semantic complexity, the two basic concepts that underlie the theory according to the authors?
• Are there other theoretical frameworks, constructs, or measures in the published literature that can be contrasted with the Lexile Framework?

1.0 The Logical Basis of the Lexile Framework

Conceptually, the Lexile Framework appears to embody a very simple view of reading, issuing from a simple, general theory of communication. Stenner & Burdick (1997) state: "All symbol systems share two features: a semantic component and a syntactic component...In language, the semantic units are words. Words are organized according to rules of syntax into thought units and sentences...In all cases, the semantic units vary in familiarity and the syntactic structures vary in complexity. The comprehensibility or difficulty of a message is dominated by the familiarity of the semantic units and by the complexity of the syntactic structures used in constructing the message." Two elements are encompassed here: a picture of language processing, and a picture of where complexity arises.

1.1 Language Processing

This Lexile picture is broadly compatible with widely held "bottom-up" views about the processing of spoken and written language (Baker, Atwood and Duffy 1988). Linguistic input is taken to be parsed into its constituent sentences, phrases and words; word meanings are looked up, and meanings of larger constituents are computed from them. However, other important aspects of language processing must be accommodated in addition to words + syntax. First, linguistic understanding must allow for **context-dependent element**s, including pronouns, demonstratives, deictic verbs and adverbs, and tenses. A listener hearing (1), for example, will not be able to understand what is conveyed by the utterance unless he/she is in a position to identify which male is intended by *h*e, which object is intended by *that too*l, what time the sentence is spoken (in order to identify the reference of *tomorro*w) and where the speaker is (implicitly referred to by the use of *brin*g)
(1) He will bring that tool tomorrow.
None of this information can be achieved by lexical lookup. Dictionaries (mental or otherwise) do not provide fixed referents for *he, that, tomorrow,* etc. Rather, the reference of pronouns and other context-dependent elements must be calculated using the listener's knowledge of the extra-linguistic context in which the sentence is uttered. Context-dependent items are pervasive in natural language, and their processing is known to introduce serious complexity into the task of natural language understanding by computer (refs.). There is also evidence that such items can induce processing complexity in

reading. For example, in an eye-tracking study, Frazier (1984) found that although young readers encountered no greater difficulty in processing left-branching sentences (2a) versus their right-branching equivalents (2b), insertion of a pronoun into the embedded clause caused left-branch examples (3a) to be processed more slowly, implying greater complexity.

(2) a. [That the traffic in this town is unregulated] bothers me.
b. It bothers me [that the traffic in this town is unregulated].
(3) a. [That people look at him strangely] bothers Mary.
b. It bothers Mary [that people look at him strangely].

This result strongly suggests that resolution of context-dependent items can interfere with, or amplify otherwise insignificant complexities in, syntactic processing by a reader.

A second important aspect of language processing going beyond words + syntax is **pragmatic**s. Words + sentence + context yield the literal meaning of a sentence, however the informational content of a sentence very typically exceed its literal meaning. This is because speaking is fundamentally an interactive enterprise between persons, taking place against a background of shared assumptions about cooperative behavior (Grice 1990). These assumptions typically allow a listener to conclude more than what is literally said. To illustrate, consider the situation in which Y throws a party and the dialog in (4) subsequently occurs.

(4) X: So, how did your party go?
Y: Some of the people I invited didn't show up.

Y's statement literally asserts that there were invited individuals that didn't come to the party, but it also conveys the (logically independent) proposition that there were invited individuals who did come. Grice shows this inference to result from a complex calculation on Y's utterance, based on the assumption that he/she is a cooperative speaker - one supplying truthful, complete and relevant information .

Pragmatic computations like those involved in (4) are present in both spoken and written text, and, once again, are known to introduce complexity into the task of natural language understanding by computer. Furthermore pragmatic principles like those studied by Grice have been proposed to be subject to maturational development (Crain (refs)). It is highly likely, therefore, that pragmatic complexity affects reading processing.

In summary, the bottom up words + syntax model of processing embodied in the Lexile Framework appears to overlap with what is believed about NL processing, but the correspondence is not complete. There appear to be significant aspects of language processing, with implications for reading comprehension, that do not appear to be expressed in the Lexile Framework. Furthermore, it is not clear how to incorporate these variables in a way that would yield machine-calculable values

1.2 Processing Complexity

The words + syntax model embodied in the Lexile Framework takes processing complexity to arise predominantly from the familiarity of the semantic units (words) + syntactic complexity of the string.

The view that word familiarity correlates with ease of lexical processing has some linguistic support. Miller (1991) has argued that frequency is a crucial component in word learning; children must hear a word numerous times in different contexts to master it. Furthermore, studies of reading show that efficient processing of words involves a form of holistic perception that is achieved only by prior, repeated attention to sequences and combinations of letters that occur in the words (Adams 1990). Presumably frequency of exposure is crucial to achieving this holistic stage. Finally, explicit word decoding typical of poor readers is to known to place a heavy burden on processing resources, which may block or delay higher-level operations (Perfetti and Lesgold 1977). Given these results it is plausible to conclude that familiarity plays a role in word learning and efficient word processing.

The second assumption - that processing difficulty arises through syntactic complexity – is also plausible and widely assumed, although interpretations of what complexity amounts to vary considerably. Some views hold complexity to be a property of the linguistic object per se; others take it to be determined relationally, by the interaction between the language processor with the linguistic object. Thus Smith (1988) proposes that linguistic complexity is a property of individual sentences and provides a thorough and explicit taxonomy of complexity types, including systematic complexity, surface syntactic complexity, interpretive complexity and phonological complexity. Most parsing studies hold complexity to be arise when specific properties of the human processor interact with particular aspects of lexical items and constituent structure.

2.0 The Variables Employed by the Lexile Framework

The Lexile Framework operationalizes its view of reading comprehension & complexity in three variables, two derived from the text, and one derived from the reader. The text variables are word familiarity, measured against a standard reference corpus, and average sentence length. The reader variable is score on a set of cloze questions. These variables are correlated by means of a regression or "calibration" equation.

The general questions that arise for the Lexile Framework are the following:
• Are the two text variables sound? Do they measure some important components of the reading process?
• Are the two text variables complete? Do the two text variables measure all important components of the reading process?
• Is the reader variable sound and complete? Does it measure some or all important components of reading comprehension?

I am not in a position to evaluate the third question. I do not know whether a cloze question task of the sort described in the Lexile materials represents an adequate test of comprehension. It may well be that there are a number of valid, progressively inclusive definitions of sentence comprehension, for example: (a) ability to construct a coherent literal meaning for the individual sentence; (b) ability to do (a) plus integrate the literal meaning with that of the surrounding text; (c) ability to do (b) plus integrate the results into a wider reader knowledge base. And so on. It is conceivable that the Lexile materials

do in fact test a certain basal level of comprehension. The sentence completion tasks typically appear to require the computation of entailments that in turn require computation of the strict and literal meaning of the sentences.

2.1 Soundness of the Lexile Framework Variables

We have already noted that word familiarity is arguably an important component in reading comprehension and reading complexity. Hence this variable does appear sound. The variable of sentence length is less clear. The authors of the Lexile Framework materials acknowledge explicitly that sentence length is not causally related to sentence complexity.

This assumption has been strongly confirmed in a variety of studies. It has been shown that comprehension can be increased by revision that does not affect average sentence length (Davison and Kantor 1982; Charrow 1988). And it has been shown that comprehension can be decreased by revisions that actually decrease length, as in the study of Irwin and Pulver (1984) which showed that omission of sentence connectives reduced text comprehension, despite decreasing sentence length.

Stenner (nd.) offers a rationalization for sentence length as a useful proxy for syntactic complexity, based on the approach to reading comprehension by Crain and Shankweiler (1988). In brief, the latter argue that differences between good and poor readers do not reflect fuller vs. lesser mastery of linguistic knowledge, but rather processing limitations on the ability to use that knowledge. More specifically, Crain and Shankweiler argue that slower orthographic decoding by poor vs. good readers taxes verbal working memory, and preempts higher processing. Thus "working memory limitations create a 'bottleneck' that restricts the utilization of higher level language processing systems, preventing proper comprehension of what is read." (p.184). Stenner suggests that this general view of reading comprehension difficulties might explain why sentence length could play an important role in processing complexity. If verbal working memory has limited capacity and decays rapidly, then a seemingly shallow variable like sentence length might well be an important variable affecting processing difficulty.

Anderson and Davison (1988) anticipate a problem with this reasoning, however. They note that "even if features of words or sentence structure delay comprehension, or simply make it more difficult, the influence of these factors will not necessarily be reflected in failure to answer comprehension questions correctly... Answering comprehension or cloze questions... is based more on a memory of representation of a sentence than on a sentence piece by piece while it is being processed." (p.41) Thus, even with the rationalization Stenner suggests, there is a considerable gap between the variable (sentence length) and the result with which it is correlated by the Lexile Framework (performance on a cloze test).

2.2 Completeness of the Lexile Framework Variables

A very persistent critique of readability formulae is that the variables they employ are clearly incomplete. For example, Bruce and Rubin (1988) criticize such formulae for ignoring (among other things): "...discourse cohesion characteristics, the number of

inferences required, the number of items to remember, the complexity of ideas, rhetorical structure and dialect." (p.8) They also fault readability formulae for ignoring "such reader-specific factors as motivation, interest, values or purpose." (p.8) Anderson and Davison (1988) cite these additional variables as well, adding prior knowledge by the reader. (See also Baker, Duffy and Atwood 1988; Charrow 1988; Green and Olsen 1988). All of these factors, ignored in readability formulae, have been shown in experimental studies to affect reading comprehension in a significant way.

I will not attempt to summarize these results, but I note that none of these critiques is answered in the Lexile Framework materials that I received. So far as I can see, the charge of incompleteness leveled against readability frameworks also applies to the Lexile Framework, whose heart is a readability formula based on the same variables as its predecessors.

2.3 Evidence for the Lexile Framework Variables

Apart from very brief discussion attempting to rationalize its two text variables, proponents of the Lexile Framework appear to rely exclusively on statistical evidence in support of their approach.

Stenner and Burdick (1987) gives what appears to be the core statistical arguments for the Lexile Framework. In one case "1,780 reading comprehension test items appearing on 9 nationally normed tests were analyzed...The study correlated empirical item difficulties provided by the publisher with the Lexile calibrations specified [by the Lexile framework]."

In the second case, Lexile scores were computed for units in 11 different basal reader series. A comparison was made between the unit sequence given by the publisher and the unit sequence determined by Lexile scores. In each case, the resulting corrected correlations are very high, .91 and .97 respectively. Stenner and Burdick conclude that the first set of results "... show[s] that most attempts to measure reading comprehension ... measure the common comprehension factor specified by the Lexile theory. (p.13)" A similar conclusion is implied for textbook sequencing in the second set of results.
The wording in Stenner and Burdick's conclusion is at least as interesting in what it doesn't claim as what it does. Notice that it does not claim that the Lexile Framework measures reading comprehension. Rather it says that the Lexile Framework measures the same factors that comprehension tests do, and, by extension, the same factors that are relevant to the sequencing of basal readers. This point invites a number of questions. We might naturally ask: How did the designers of the nationally normed tests assess comprehension level? And how did the textbook publishers determine the sequence of their readers? What objective factors did they use to rank the respective items? So far as I can tell, Stenner and collaborators do not comment on this point. Nonetheless the question is clearly important. As Bruce and Rubin (1988) point out, readability formulae are often used by test designers and textbook publishers for just the two purposes mentioned above: to rank the difficulty of reading test questions, and to sequence texts. Furthermore, these readability formulae invariably include some measures of word familiarity and sentence length as core variables. Plainly if the data sets employed by the

Lexile researchers were ranked by reading formulae, then the fact that the Lexile Framework extracts the same ranking using the same variables would be unsurprising, and would provide no evidence for the general correctness of the approach. At most it would it would show that test designers, text publishers and readability researchers are locked in a disturbingly tight embrace. To be able to assess the statistical we therefore plainly need to know how the national normed test questions and basal reader series were ranked in the first place.

3.0 Comparable Approaches & Constructs

The final question ESSI panel members were asked to address was whether there are other theoretical frameworks, constructs or measures in the literature that can be contrasted with the Lexile Framework. I am not aware of any such frameworks that involve easily derived numerical values plugged into a formula. There are developed theories of reading comprehension, but none yet make quantifiable predictions about the readability of a given text.

References

Adams, M. (1990) *Beginning to Rea*d. Cambridge: MIT Press.

Anderson, R. and A. Davison (1988) "Conceptual and Empirical Bases of Readability Formulas," in A. Davison and G. Green (eds.) *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidere*d. Hillsdale, NJ: Lawrence Erlbaum Associates.

Baker, Atwood and Duffy (1988) "Cognitive Approaches to Assessing the Readability of Text,"in A. Davison and G. Green (eds.) *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidere*d. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bruce, B. and A. Rubin (1988) "Readability Formulas: Matching Tool and Task," in A. Davison and G. Green (eds.) *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidere*d. Hillsdale, NJ: Lawrence Erlbaum Associates.

Charrow, V. (1988) "Readability vs. Comprehensibility: A Case Study in Improving a Real Document," in A. Davison and G. Green (eds.) *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidere*d. Hillsdale, NJ: Lawrence Erlbaum Associates.

Crain, S. and D. Shankweiler (1988) "Syntactic Complexity and Reading Acquisition," in A. Davison and G. Green (eds.) *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidere*d. Hillsdale, NJ: Lawrence Erlbaum Associates.

Davison, A. and R. Kantor (1982) "On the Failure of Readability Formulas to Define Readable Texts: A Case Study from Adaptations," *Reading Research Quarterly 1*7: 187-209.

Frazier, L. (1984) "Syntactic Complexity," In D. Dowty et al (eds.) *Natural Language Processing: Psychological, Computational and Theoretical Perspective*s. Cambridge, England: Cambridge University Press.

Green, G. and M. Olsen (1988) "Preferences for and Comprehension of Original and Readability Adapted Materials," in A. Davison and G. Green (eds.) *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidere*d. Hillsdale, NJ: Lawrence Erlbaum Associates.

Grice, P (1990) *Studies in the Way of Word*s. Cambridge: Harvard University Press.
Miller, G. (1991) *The Science of Word*s. New York: Scientific American.

Perfetti, C. and A. Lesgold (1977) "Discourse Comprehension and Sources of Individual Differences," in M. Just and P. Carpenter (eds.) *Cognitive Processes in ÿÿÿÿ*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Smith, C. (1988) "Factors of Linguistic Complexity and Performance," in A. Davison and G. Green (eds.) *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidere*d. Hillsdale, NJ: Lawrence Erlbaum Associates.

Stenner, A.J. (n.d.) "Theory Referenced Measurement."

Stenner, A.J. (1996) "Measuring Reading Comprehension with the Lexile Framework"

Stenner, A.J., and D. Burdick (1997) "The Objective Measurement of Reading Comprehension"

Assessing the Lexile Framework -- Comments
Carlota S. Smith, University of Texas

The panel of ESSI met to discuss the Lexile framework. The framework allows independent measure of the comprehensibility of a text, and the ability of a reader to comprehend. With these tools one can test reading ability and match texts with readers for study and development after testing. There were two main topics at the meeting; consideration of the Lexile framework itself, and whether the framework would be appropriate for the work of ESSI in the programs known as NAEP and NAAL. (Since many questions that I mentioned in my earlier comments were answered at the April 26 meeting, I omit them here.)

1. The framework. As several panel members pointed out, the Lexile framework offers a severely limited approach to reading assessment. There are many aspects of the reading process, and facts about readers, that it doesn't deal with - nor does it claim to. I would emphasize that this does not mean that the approach itself is not useful; rather, its limitations must be understood and allowed for.

The situation is a familiar one for linguists. In the field of linguistics, those of us who work with formal sentence analysis are far removed from the communicative aspects of language and the means by which subtle messages can be conveyed (and received). The formal rules that one attempts to discover underlie language, but in a sense do not deal with the complexities of communication and language use. This does not mean that there is no connection, or that one is not interested in communication. Actually the interface between syntax, semantics, and pragmatics - a special interest of mine - is s particularly strong and challenging area. All of this is to say that, in thinking about the Lexile framework, it is important to see both what it does and does not offer. The framework offers a way of assessing reading ability and text complexity that abstracts away from factors of context, situation, personality, etc.

2. The Lexile analysis is based on two features: the frequency of words, and the length of the sentences, in a given text I have some comments about possible improvements and pitfalls for each one; they might be useful for future work as the Lexile framework becomes more sophisticated. They also might be useful for ESSI if it at some later point wants to ask the Lexile people to work with them.

• Word frequency: There is solid evidence in psycholinguistics for the importance of word frequency in language processing - and reading is a form of language processing. All studies involving language must control for word frequency. The correlation of word length with frequency is well-established, going back at least to work of George Zipf in the late 1940s. Indeed, word frequency is an indispensable measure.

However, there are some weaknesses in the standard frequency counts often used in psychology and experimental psycholinguistics. I have observed that the middle range word frequencies cover a very wide span - rather like the range covered by a grade of B

in a system that allows only two acceptable grades, A and B, with C and anything below an unacceptable grade. The same mid-range categories tend to include words that are quite common and words that are quite unusual. This observation comes from my experience in developing test items for experiments in language processing and language acquisition, and from comments of graduate students with whom I worked. One might hope that the frequency counts on which it is based would be more fine-grained.

Another problem with published word frequency lists was that few of the texts on which the lists were based represent current words and popular usage (in the Brown corpus for instance). Quite often we found that words in current use were not in the corpus, and therefore not listed in the frequency counts. Yet for Lexile materials it would be desirable, I think, to include current materials with current words. One can imagine that certain children might do better with reading materials with many current words.

The component of word frequency would be an area to consider in further development of the Lexile format. There are a number of frequency counts and corpora now available in addition to general lists such as The American Heritage word frequency book. (edited by John Carroll) and other published studies. For instance, at the University of Pennsylvania Cognitive Science Center there is a large corpus of English that includes frequency counts, I believe. IBM and other research institutes also have corpora, many of them quite varied, and including some relatively modern texts. It would be worth while to make a detailed study of the available lists, especially at the middle ranges of frequency. Doing this would, I think, increase the sensitivity of the frequency measure and be useful for texts and readers at relatively advanced levels.

The articles about the Lexile framework give a list of word features, under the rubric 'semantic': part of speech, number of letters, number of syllables, the modal grade at which a word appears in school materials, content classification, and frequency. The significance of these features wasn't clear: I don't know whether they are correlated to the frequency measurements in some way.

I want to point out that, of the 6 features, 2 are semantic: content classification and, possibly, part of speech. The others are not semantic. This does not mean that they are unimportant, but one should be clear about their contribution. If content classification is important, it would be useful to know more about it: what the categories are, how they are decided, etc.

It is well-known that 'function words' such as articles (*a, the, this, that, to*) are more frequent than other classes of words. Given this, I wonder whether there is any reason to include function words in the frequency counts. Perhaps doing so is simpler than designing a computer program to find and exclude them.

• Syntactic complexity and sentence length: We are told that sentence length correlates with measures of complexity, and I have no doubt that this is correct. I am somewhat familiar with the field of writing pedagogy, in which this is an old and much-discussed problem Apparently, there is a good deal of statistical support in the literature

for the claim that a simple count of sentence length correlates with sentence complexity over a wide range of material.

Two points came up at the meeting concerning this topic. The first is that sentence length is important not only because it correlates with other measures, but directly, as a factor that affects reading in its own right The reasons for this are still the subject of active research and debate.

Secondly, people are interested in the components of sentence complexity that are indirectly assessed by the measure of sentence length. The main aspect of complexity that it captures is amount of material in a sentence, often counted in terms of the number of distinct clauses. Clauses have full verbs, according to most counts. Sentences with many clauses are longer than sentences with few clauses. For instance, consider the following example (from an Op-ed article in the New York Times)*: In the United States, we have lost over 500,000 jobs that were available in the oil industry while we have grossly increased our dependency on foreign oil; we now import 55 percent of what we use*. This sentence is quite long: it has 4 full clauses (the verbs are *lost, were available*, *increased*, *import*) and a prepositional phrase at the beginning. It is not easy to read and understand such a sentence. Evidently, this kind of complexity correlates with sentence length. Other phrases also contribute to length: prepositional phrases, adverbials - all the material in a sentence.

But some aspects of sentence complexity are not directly assessed by the measure of length. There are grammatical ways to compress clauses that makes them shorter and denser in information than sentences with full clauses. Such sentences tend to be relatively difficult For instance, compare these two sentences; *John opened the door to the restaurant and walked in; he was singing loudly* and *Opening the door to the restaurant, John walked in, singing loudly*. The second is shorter: it has essentially the same information as the first but only one full clause. Another example: compare *That John won the race surprised us* with *It surprised us that John won the race*. The second sentence is longer and, most people agree, easier to read and understand. Finally, I give an example of compression from the same New York Times article is as above: *With the shift toward the information and service economy, and away from manufacturing, the United States is less dependent on oil*. The first part of this sentence - introduced by *with* - is a nominal, a condensed version of a full clause (something like *The United States has shifted toward the information and service economy*); the second part is understood in terms of material that appears in the context, but is omitted in the sentence (*the US is less dependent on oil - than ..is*). Such compressed sentences are shorter but arguably more complex than sentences in which each clause is complete.

A related and equally important point is that of transparency. The topic was addressed briefly at our meeting in Washington. Sentences are transparent if their form mirrors the way they are understood. For instance, the sentence *Mary ate the spinach* is transparent because the 'logical subject' is also the sentence subject. The corresponding passive sentence, *The spinach was eaten by Mary*, is less transparent because the sentence subject differs from the logical subject. The linguistics terminology is 'surface

structure' and 'underlying structure'. So in a passive sentence, the surface structure subject is different from the underlying subject. The passive is also longer, of course, so that in this case length and transparency are confounded.. There are other possible variations besides active and passive, e.g. *The spinach, Mary ate*; or a pseudo-cleft or cleft sentence –*What Mary ate was the spinach, It was the spinach that Harry ate*; *It was Mary that ate the spinach* etc. These sentences are all less transparent than the simple active declarative.

I am not sure of the relation between transparency and length, but I don't think that it's consistent: in some cases transparent sentences are shorter than opaque sentences, in other cases they are not. There is reason to think that certain kinds of opacity are easier than others for people to deal with. One example is repeated material, which is often not present in a sentence; we may say that it is 'deleted' or 'in zero form'. For instance, compare the following: *Mary went to the store and Mary went to the movies*; *Mary went to the store and to the movies*. Or, *John is taller than Bill is*; *John is taller than Bill*. The sentences with deleted material are shorter and apparently not more difficult to understand than the longer, full sentences.

Transparency plays a role in language acquisition, according to Dan Slobin, Steven Pinker, and many others. Children first learn the most transparent sentence structures and gradually acquire others. I do not know whether the role of transparency in reading has been investigated, but I suspect that it has; this is an interesting topic that would be worth pursuing (I personally would be interested in pursuing it, given time and resources).

I would suggest that teachers be told as much as possible about the complexity measures that justify the sentence length measure. Syntactic complexity is an important aspect of reading (and writing) that is not emphasized in most current approaches to teaching these subjects. In fact syntactic details are often given short shrift, due to the belief that motivation and student involvement are the most important factors for reading and writing. While I do not question the importance of the latter, I think that there is a real need to emphasize the syntactic aspects of the materials that children encounter in reading

Many teachers may not know what the important correlations are between sentence length and sentence complexity, much less how to help children become familiar with the complex syntactic structures that add significantly to the difficulty of a text. Making such information available would be another direction for the Lexile framework; it is, at least, a connection worth making explicitly for teachers when possible.

2. Would the Lexile framework be appropriate for the testing done in the NAEP and NAAL and programs? This topic was discussed quite extensively at the meeting. It is clear that many aspects of the tests - including the formatting of documents and other special materials - are simply not covered by Lexile.

In the 1996 article, there is a discussion of the test items in the section on Testing the Lexile Equation. We are told that that the passages but not the questions and answers were analyzed in the Lexile framework. This is unfortunate. For testing relatively unprepared children, and speakers of other languages, it would be very important to control the questions and answers on tests. Unless and until this changes, the usefulness of the Lexile approach for NAEP and NAAL programs will be limited.

Nevertheless it would be useful to have baseline materials that have been assessed by Lexile, and might be the first step toward a more rigorous approach to the testing materials.

There is a possibility that children with relatively little academic preparation, and native speakers of languages other than English, might be underrated by the Lexile approach. I think that it would be important to use words in current usage for such populations; and to make some attempt to consider compression and transparency (other aspects of sentence complexity besides length). Perhaps special care might be taken for these aspects of test materials.

Reading in a Language Not Your Own: The Applicability of the Lexile Framework for Second Language Readers[1]

Heide Spruck Wrigley

A. Summary of Concerns

I reviewed the Lexile Framework with an eye towards its applicability for measuring the reading skills of non-native speakers of English, particularly those with limited English proficiency. I can see the framework working well in terms of sorting non-native speakers into those who are proficient readers of various texts and those who struggle with English texts. I can also see how the framework can be used to sort texts along a continuum, ranging from texts that can be read with relative ease by individuals who are new to English and texts that require a high degree of proficiency. Finally, I can see the framework's applicability to assessing the various levels of biliteracy that bilingual children and adults may possess. There is a further possibility for using the framework with immigrant families: The Lexile process could be used to indicate dual scores, one for reading proficiency in the language of the home and the other for proficiency in reading texts in English. Obtaining a measure of biliteracy for various groups of English learners could be invaluable in language policy, curriculum development, instruction and testing.

I have a number of concerns about both the theoretical foundation and the ultimate use of the framework with English language learners. First, I worry that using only two features of a text (syntactic and semantic) and using only one element for each (sentence length and vocabulary respectively) might not adequately capture the difficulties that text written in English represent to those who have not yet mastered the language. More importantly, I worry that the framework might not allow testers, researchers, or practitioners to see to what extent second language learners are able to derive meaning from a text when certain elements are present. I am also concerned that the abilities of second language learners to comprehend and make meaning will be underestimated and opportunities to select or develop texts appropriate for this group will be lost. Finally, I am troubled because the framework can easily lead to false interpretations of the assessment results, as limitations in language proficiency become confounded with reading difficulties.

B. The Theoretical Basis

The field generally agrees that reading is a highly complex process and that comprehension arises from interactions between the reader and the text. As such, comprehension is dependent on both features inherent in a text (syntactical, lexical, pragmatic) and on elements that the reader brings to the text, e.g., the ability to decode fluently and accurately, various levels of background knowledge, a stance toward the text (interested, skeptical, resistant to the information presented), as well as affective factors such as motivation. While the dimensions of readability that the Lexile Framework addresses (sentence lengths and word frequency) are strong determinants in establishing what types of texts can be understood by what kind of reader, these dimensions are by no

means the only factors that influence comprehension. Other factors, such as the need to know, the effective use of text-processing strategies or the ability to activate real world knowledge in trying to make sense of print are significant as well. While these may wash out in large scale assessments, they are nevertheless important enough to consider in discussions of what the framework can and cannot tell us either about a specific text or an individual reader.

The reading process is significantly more complex for those attempting to read in a language not their own, since success in reading not only depends on one's ability to process print (an ability one may or may not have in the native language), but also on one's ability to manipulate and understand the language in question. In other words, a person might be a fluent reader in Arabic, but her lack of English proficiency makes it difficult to make sense of a textbook in English that she could easily read in Arabic.

It is axiomatic that we cannot read texts in a language that we don't understand, but we are far from knowing how much language proficiency and socio-cultural knowledge must be present before a text that is comprehensible to someone who grew up speaking English can be understood by a non-native speaker. There may be a threshold level of proficiency that allows second language readers to make sense of written texts in the same way that native speakers make meaning, but these levels tend to differ by texts (notably genres) and tasks (answering comprehension questions, summarizing, oral or written recall). If those who speak English as a Second Language with various levels of proficiency are included in the framework along with those who are native speakers or have near native competence, distinctions between the groups disappear and results are likely to be skewed. Unless non-native speakers and the texts they are able to read are assessed as a separate set and the results analyzed by proficiency levels, the Lexile Framework will have limited usability for those seeking to establish threshold levels of reading competence for second language readers.

Here are some additional considerations that need to be taken into account:

1. **Sentence Construction:** For those who are not fully proficient in the language of the text, syntax often plays a much greater role in facilitating or impeding comprehension than does sentence length. Relatively short sentences easily understood by native speakers often present difficulties to readers new to a language. For example, sentences such as "he was run over by a pick-up" are difficult to process by English learners since passive constructions are acquired relatively late in the language acquisition process. The use of the verb plus particle construction in "run over" and the inclusion of "pick-up" as a noun indicating a truck inhibits comprehension further.

2. **Word Frequency:** Vocabulary familiar to native speakers may not be familiar to non-native readers, and familiarity with English vocabulary will differ depending on the language of origin. Those coming from Romance languages for example will recognize Latin based words such as "edifice" and "malnutrition" but may have difficulty with every day words such as "mop" or "turnip," since these rarely

appear in ESL/EFL textbooks. Similar difficulties are presented by texts using idiomatic or slang expressions, common in literature for young adults, such as "She saw him at a party and did her best to pick him up." In essence the difficulties with text comprehension exhibited by non-native readers may have much more to do with language difficulties than with difficulties processing texts.

3. **Cultural literacy:** Texts with strong references to cultural contexts are likely to present significant difficulties for language learners as well. While native speakers are much more likely to understand references to television shows or to sports such as American football and baseball, comprehension breaks down for non-native speakers who lack the necessary background knowledge to interpret a text that has a cultural context specific to the country in question. For example, "betting on the Superbowl" is not an expression that is transparent to those relatively new to the United States.

4. **Content knowledge:** The importance of content knowledge should not be underestimated as a factor influencing comprehension for those readers who are relatively new to English. In second language reading, content knowledge includes (1) an understanding of how English works (in terms of formal rules that govern syntax and morphology), (2) subject matter knowledge necessary to understand vocabulary and concept related to a particular domain (such as health or civics), (3) pragmatic knowledge or familiarity with certain text types (prose and document texts, for example), and (4) the cultural knowledge necessary to understand references to traditions, customs, and ideas shared by native speakers of English but not necessarily by the foreign-born.

Here is an example to illustrate the importance of content knowledge in comprehension. In doing research in the schools in Long Beach serving Cambodian students, we found that almost all the Khmer speaking teenagers failed their health classes, although their English was adequate for other subjects. These youth simply had not grown up with learning and talking about the body and its systems the way children raised in United States were. (From the elementary grades on, American children read about topics such as "our friend the heart.") The Cambodian teens had never studied what goes on inside a body and found health concepts entirely puzzling even when relatively common words, such as "lungs" or "kidney function" were used. Difficulties persisted even when these terms were accompanied by illustrations in the textbooks. What's more, any talk about the body was seen as embarrassing, creating further barriers to understanding. If these students were assessed in their ability to make sense out of simple health related texts, their scores would have been much lower than for other texts that did not present the same kinds of barriers.

Finally, any theory of comprehension applicable to second language speakers should take into account the different kinds of knowledge required to obtain meaning from a text written in a language they are still trying to master. Such a theory must also acknowledge that in reading in a second language, two systems interact: (1) the language proficiency system that tells us how much of the language an individual has acquired and (2) the text

processing system that indicates how well the person can deal with written information in the target language. (In many cases, the latter is influenced by reading proficiency in the first language as well.) If the different competencies related to each of these systems remain confounded, we will not be able to get a true picture of reading abilities of English language learners and will not be able to place their performance on a common scale with those who have grown up with English.

B. Comments on the Syntactic and Semantic Measures Proposed

For non-native speakers of English, an understanding of the syntax of English is a key component in being able to derive meaning from print. Similarly, a large store of vocabulary allows for access to English even in cases where syntactic knowledge is still a bit shaky. However, looking only at sentence length as a measure of syntactic complexity has a number of limitations, given the other factors that inhibit or promote comprehensibility. Beyond those discussed in the previous section, factors that influence reading comprehension in a second language include the following:

1. **Cohesion and Coherence:** The syntactic measure proposed (sentence length) operates merely on a sentence level. Yet there is a body of research that suggests that relative text coherence and cohesion play a significant role in comprehension. Texts that logically don't hang together well, present real difficulties for non-native speakers who find them especially confusing. For many non-native speakers, understanding is aided by text markers (such as those indicating sequence or enumeration) which, while they may add words to a sentence, nevertheless act as roadmaps to understanding. Those text markers can appear in various forms within a sentence ("there are three significant reasons why refugees leave their countries") or a separate sentence ("The third reason will become clear shortly.") In either case, they provide non-native speakers with signals as to what to pay attention to, facilitating comprehension in the process. By not analyzing the role for these text aids, along with other features related to cohesion and coherence, we are likely to miss what non-native speakers can understand if logical relationships are made clear and some textual support is given.

2. **Sentence Length:** The framework assumes that, all things being equal shorter sentences are easier to process than longer sentences. Many L2[2] researches suggest otherwise. They argue that it is not sentence length per se, but the number of subordinations within them that matters. There is now a body of research that indicates that making the relationship between clauses transparent (and creating a longer sentence in the process) actually increases comprehension rather than inhibiting it. For example, the sequence "The chicken crossed the road. It wanted to get to the other side." is more difficult to process than the sentence "The chicken crossed the road because it wanted to get to the other side." There are reasons to believe that non-native readers still struggling with language are likely to have

[2] L2 refers to a learner's second language (although in some cases it may be the third or fourth language studied or learned), while L1indicates the first language.

higher levels of comprehension if the relationship between ideas are made explicit even if sentence length is increased in the process.

3. **Syntactic complexity:** Syntax, often referred to simply as "grammar," is a key determinant of readability in second language reading. Relative complexity of the syntactic construction appears to matter a great deal more than sentence length per se. This is particularly true in cases where syntactic features used in English do not exist in the native language, creating an additional cognitive load that needs to be considered in assessing reading proficiency. For example, processing the sequence "He did not get the memo. Neither did I." presents a significant challenge to those with limited proficiency in English**,** since the rule that triggers the construction is not commonly found in other languages. We can easily see that although the sentence "Neither did I." is quite short and easily comprehensible to native speakers, it presents a triple challenge: (1) understanding the negative "neither," (2) understanding the inversion ("did I"), and (3) understanding the use of the auxiliary "did." While sentence inversion caused by negative adverbs such as hardly, never, and barely ("barely had he spoken when…") might be considered a literary convention, there are other much more common syntactic features of English that present great difficulties for non-native speakers. These include hypotheticals, such as "had I had the money, I would have gone" and passives, both formal and informal ("They were laughed at." "She got cheated in that deal."). In the end, then, it is the relative complexity of the syntax, what Berman (1984) calls "heaviness," along with the transparency or opacity of the sentence structures that causes difficulties for non-native speakers, not sentence length itself.

4. **Vocabulary:** The sophistication and "commonness" of the vocabulary used in a text has consistently been shown to have a significant effect on understanding for second language readers (as it does for L1 readers). The effect of difficult vocabulary is compounded if topics are unfamiliar, creating a double challenge for immigrant and refugee students who may be unfamiliar with both the words themselves and the concepts underlying the words used. For example, a text written in German might make references to Weltanschauung, Fahrvergnuegen, or Schadenfreude, ideas immediately accessible to native speakers of German. Those trying to learn German often have great difficulties, not only because of the size of these words, but because the underlying concepts have no direct equivalent in English. Similarly, Asian students might be thrown by the word "privacy" since the concept is culturally foreign to many of them. Homonyms, words that sound the same but are spelled differently and words with multiple, unrelated meanings are especially difficult to process for non-native speakers, particularly for those with lower levels of proficiency in English. These readers are often not able to use context clues to reduce the potential ambiguity of common words as an aid to comprehension. For example, when seeing the sentence "You'll have to get rid of that bug." readers new to English might not be able to determine if "bug" refers to a VW, a small beetle, a computer glitch, or a spying device, making it very difficult to derive the intended meaning from a passage. While native speakers

might be able to use context clues to disambiguate meaning, non-native speakers with limited proficiency are not able to do so to the same degree since they focus mostly on sentence level meaning. As a result, they often are stopped cold by terms of this sort since they may not be familiar with one of the possible meanings of these homonyms. What's more, the word "bug" itself (in any of its meanings), although quite familiar with native speakers, might not be familiar to second language learners since it seldom appears in English as a Second Language (ESL) or English as a Foreign Language (EFL) textbooks.

5. **Text types:** To what extent a non-native speaker will be able to process a text written in English does not solely depend on syntactic and lexical knowledge. Text types and tasks play a role as well, along with the motivation to engage in these tasks. For now, the Lexile Frameworks deals only with prose texts. However, if the framework is to be applied to items such as those used in the NAAL, which includes the use of "document literacy" and "quantitative literacy," the level of difficulty of non-continuous texts will have to be analyzed as well. To what extent non-native readers can cope with information laid out in columns, graphs or charts will need to be assessed, since being able to interpret these print forms will be influenced by one's familiarity with these formats. In addition, the density of the text will need to be considered (medicine labels, for example, tend to be extremely dense), as will the complexity of the information presented (simple linear formats are easier to process than a flow-chart or a matrix, for example).

6. **Questions and tasks:** Not addressed through the framework are the difficulties that instructions and questions present for non-native speakers. For those at the lower end of the language proficiency scale, instructions regarding a passage or task often present greater difficulties than the task itself, which could be accomplished rather easily if one only understood what to do. If the framework is to be used for the assessment of second language readers, instructions and questions must be subjected to the same kind of rigorous analysis as the passages themselves.

D. Competing Frameworks

As others have pointed out (see the attached paper by M. J. Adams), the Lexile Framework presents advantages to over earlier schemes to assess readability. Sentence length and word frequency are no doubt essential factors in sorting texts into those which can be easily read and those likely to present difficulties to many readers. Similarly, readers who have a wide store of vocabulary and are able to process long sentences are likely to deal with a much wider range of texts than those who know fewer words and get lost in complex syntax. However, before the framework is used to include non-native speakers of English, additional features that allow or inhibit readability for second language learners be taken into account. One potential alternative framework might be the scheme being developed by the American Institutes for Research (AIR), designed to determine the complexity of items and tasks in the current NAAL. This scheme might

provide guidance in assessing the cognitive load in other assessment items such as those used by NAEP.

E. Final Recommendations

To ensure the usefulness of the Lexile Framework (LF) when assessing second-language readers who are not yet proficient in English, comprehension needs to be assessed along two dimensions: (1) the ability to process text in a language one speaks and understands (the native language in most cases), and (2) the ability to read in the language one is trying to learn (English in our case). Such study could begin by assessing Spanish speakers who are learning English and be expanded to include other languages. The study could start with two components of the LF, sentence length and word frequency, but might also consider other factors, such as cultural load and syntactic complexity. It should be designed in consultation with scholars in biliteracy who have experience analyzing dual language systems.

Another possibility is to undertake analyses of different populations of second language readers, stratified by proficiency levels. These groups should be assessed separately, so that comparisons can be made, and it can be demonstrated whether or not the LF applies equally to native and non-native speakers of English. If these studies cannot be carried out, then disclaimers should be inserted in publications, saying that the system has not been tested with second language readers and text difficulty is likely to vary for those not yet proficient in English.

If features that influence second language reading are not taken into account, we are likely to risk the validity of the assessments, the generalizability of the findings, and the interpretation of performance of those new to English.

Attachment 3. Comments from Lexile Framework Developers.

A Response to "Assessing The Lexile Framework: Results of a Panel Meeting"

A. Jackson Stenner, Donald S. Burdick, Eleanor E. Sanford, Hal Burdick

Thank you for your interest in The Lexile Framework™ for Reading. We appreciate the professionalism and thoroughness with which Drs. Sheida White and John Clement directed the panel review. The panelists were energetic, thoughtful and extraordinarily "fast studies" on the complexities of the Lexile Framework. The full-day panel review was an invigorating, intellectual exchange, which has helped us formulate directions for future communication with psychometricians, linguists, and reading specialists. We appreciate the opportunity to respond to the panel's summary report, "Assessing the Lexile Framework: Results of a Panel Meeting" (2001). To make it easier for the reader to relate our comments with those of the panel, we quote from the report to focus our comments.

1. Background

Here are our responses to the four framing questions of the report.

1. *Are word frequency and sentence length solid criteria to use in determining text difficulty?*

During the last century, dozens of "readability formulae" have been developed for describing the comprehensibility of text. Almost every one of these formulae has a semantic and syntactic component. Hundreds of studies have demonstrated the value of word frequency and sentence length as proxies for these two linguistic variables. Is sentence length a good predictor of text comprehensibility because increasingly complex syntax generally requires more words or because sentence length is an indirect index of the load a sentence places on short-term memory? The explanatory power of the two proxies for semantic difficulty and syntactic complexity is impressive and well documented. However, this does not mean that the observed regularity of the Lexile formula completes the analysis of reading. Whatever the eventual answer, the "solidity" of the empirical relationship between text comprehensibility and log mean sentence length is not in doubt.

2. Are these criteria sufficient to determine text difficulty? If not, can they be improved or used for only a subset of reading passages?

A construct specification equation (Stenner, Smith, and Burdick, 1983) is useful to the extent that it predicts item difficulties. Such an equation is complete to the extent that all reliable variation in item difficulties is accounted for by the construct theory and associated specification equation.

A reading item has two "hurdles." The first hurdle is whether the reader can read the passage making up the item. The second hurdle is whether the reader can successfully negotiate the response requirement. Native-Lexile items use an "embedded completion" four-option multiple-choice format. The reader is asked to read the passage and choose the word that best restores the embedded-completion sentence. The task involves making a generalization based on the content in the passage.

The Lexile measure, based on the text features of the passage, has a root mean square error (RMSE) of approximately 150 Lexiles per item. Different item writers append different embedded-completion sentences and response options to the same passage. Since the Lexile Theory produces the same theoretical calibration for each possible item, the RMSE is primarily due to this "ensemble variation"—the unintended, but unavoidable, "easiness" or "hardness" that an item writer adds in the process of building an item.

If most of the RMSE is ensemble variation, there would be nothing left for the Lexile Theory to explain. The theory, as operationalized in the specification equation, could be characterized as complete. Presently, we are conducting two lines of research to check for completeness. The first involves estimating the magnitude of ensemble variation by having multiple item writers response illustrate the same passages. If the within-passage standard deviation approaches 150 Lexiles, then the only variance in item difficulties not explained by the Lexile Theory is an item-writing effect.

The second line of research designed to examine the completeness of the Lexile Theory asserts that if the Lexile Theory provides a complete account of text-to-text variation, then an item type that has no item-writer effect and thus no ensemble variation should produce an R^2 near unity. One possible item type is a read-aloud fluency score computed by counting the number of correct words read per minute (conditioning on text Lexile measure). If the Lexile Theory is complete and read-aloud fluency is determined to be a precise indicator of comprehension, then all of the reliable variation in passage read-aloud difficulty should be explained by the Lexile specification equation.

3. *Are the procedures used to determine word frequency and sentence length, and their calibration through Rasch analysis, adequate? If not, how can they be improved? Are there alternative procedures for assessing readability?*

When the Lexile Analyzer (available at www.lexile.com) is used to measure a book, the whole text is "sliced" into standard "paragraphs" of 125 words (or to the end of the sentence). The result for *Ivanhoe* (190,789 words) is approximately 1,500 slices of text to be calibrated. Each text slice is calibrated according to the Lexile specification equation. The Analyzer then treats these 1,500 "paragraph" calibrations as if they were test item calibrations and uses the Rasch Model to solve for the Lexile measure corresponding to a "raw score" of 1,125 correct (i.e., 75% comprehension = 1125/1500). The resulting Lexile measure for *Ivanhoe* is 1410L.

Question #3 also asks about the "adequacy" of the Rasch Model. Measurement is the process of converting observations (counts) into measures (quantities) via a construct theory. The Rasch Model states a conjoint additive requirement for the way observations and construct theory must combine in a probability model to produce measures. There is no other combination of observation and theory that produces sufficient estimation and parameter separation from reading test scores. This is why we speak of "data that fit the Rasch Model" and not "finding a model to fit data.

There is an "experiment" that shows how well we understand a construct. If presented with an instrument purportedly measuring a construct, can we use the construct theory to produce a copy of the instrument such that the raw score-to-measure table for the copy is identical to that of the original instrument? If the copy produces measures that are statistically equivalent to those produced by the original instrument, then we have demonstrated that we understand what we are measuring.

We first attempted this type of "experiment" for reading comprehension. Now, some four years and 30 studies later, we can build reading comprehension tests based on theory that produce raw score-to-measure correspondences that conform to design requirements.

4. *What is the relationship between the Lexile Framework and other measures of text difficulty?*

Every readability equation in use today uses proxies for semantic familiarity and syntactic complexity. Ten of the most commonly used readability equations estimate one common readability construct (Wright and Stenner, 1998). Do these findings mean that the only features of text that impact comprehensibility are the semantic and syntactic components found in the Lexile equation? No. There are dozens of highly correlated text characteristics that predict measures of text comprehensibility. Other researchers may choose different variables or different operationalizations of the same variable but the result will be an equation that produces predicted text difficulties that are highly correlated with the predictions made by other equations.

A Brief Description of the Lexile Framework

"The Lexile Framework measures the reading difficulty of prose texts and the reading capacity of people." (Page 3)

The Lexile Framework expresses the difficulty of text and the ability of readers in a common unit called a "Lexile." The forecasted comprehension rate experienced by a particular reader reading a particular text is determined by the difference between the text Lexile measure and the reader's Lexile measure. We remind all of our users that there are other factors that affect the relationship between a reader and a book. These factors include content, age of the reader, interest, suitability of the text, and text difficulty. The Lexile measure of a book, a measure of text difficulty, should be the first step in the book selection process with other factors then being considered. By targeting a book for a

reader, the reading experience will be more successful and enjoyable. The Lexile measure should never be the only factor considered when selecting a book.

"Measuring reading difficulty: The process is based on a mathematical formula that assigns reading difficulty values to passages of text (known as "slices"). These values are scaled; the scale unit is called a ***lexile****. The mean lexile value of all the slices in a book is the lexile value for that book."* (Pages 3 and 4)

The mean slice measure is not used to determine the Lexile measure for a book. Instead, counts correct are converted into Lexiles using the one-parameter logistic model (called the Rasch Model). When a reader is measured, it is the examinee's actual count correct (raw score) that is converted into a Lexile measure given the test item calibrations. When a book is measured, we use the book as though it were many paragraph-sized test items with known calibrations (provided by the theory). We ask the Rasch Model for the measure that corresponds to a relative raw score (percent correct) of 75%. This measure is the book measure. This procedure ensures that the distribution (not just the mean) of the slice calibrations is accounted for when solving for the measure. *"The central idea is that, when a person is reading with 75% comprehension, they are at optimal reading capacity."* (Page 4)

The primary utility of the Lexile Framework is its ability to forecast what happens when readers encounter text. A reader with a measure of 600L who is given a text measured at 600L is expected to have a 75-percent comprehension rate. This is the "default" setting within the Lexile Framework. This value was selected to ensure that texts selected are not so hard that the reader experiences frustration and loses the meaning-thread of the text, but, at the same time, are not so easy that the reader does not experience any challenge.

This "default" setting of 75% comprehension is just that, a "default." By understanding the interaction between reader measures and text measures, any level of comprehension can be used as a benchmark. A reader can modulate his or her own target by lowering the difficulty (i.e., increase to 90% comprehension) or increasing the difficulty (i.e., lower to 50% comprehension). This flexibility allows the teacher, parent, librarian, or student the ultimate control to modulate the forecasted comprehension rate to the reading purpose.

"The LF scale is set at 1000 points between first and twelfth grades. The scale itself extends from 0 to 1800 lexiles." (Page 4)

Historically, the anchor points for the Lexile Scale were obtained text from seven basal primers for the low end and text from *The Electronic Encyclopedia* (Grolier, Inc., 1986) for the high end. These points corresponded to the middle of first grade text and the midpoint of workplace text. The Lexile unit was designed to be 1/1000th of the difference between the difficulty of the primers and the difficulty of the encyclopedia.

The Lexile Scale ranges from below zero to above 2000L. There is a not an explicit bottom or top to the scale, but rather two anchor points on the scale that describe

different levels of reading comprehension. The Lexile Map, a graphic representation of the Lexile Scale from 200L to 1700L, provides a context for understanding reading comprehension.

3. Recommendations from the panel discussion

3.1 Possible Applications

"Selection of candidate reading passages for assessment items. … The Lexile Framework would be especially useful for this purpose if texts and passages were indexed by grade and topic. At present, the Lexile Web site does not provide such indexing for texts." (Page 5)

On-line at Lexile.com, it is possible to search 30,000+ books by keyword and Lexile measure (or range). EBSCO provides a similar on-line search capability for over four million Lexile-linked non-fiction articles. Off-line, MetaMetrics has indexed a 400-million-word corpus created from scanning and measuring 30,000 fiction and non-fiction books. It is now possible to search for "photosynthesis" and retrieve dozens of candidate passages, all with Lexile measures in the range of 1100L to 1200L. These passages can be screened for word length, presence of clichés, etc. Finally, the passages can be formatted to look like passages administered on the NAEP or NAAL assessments and key words in the passages can be highlighted to assist item writers in formulating questions. This type of item-writing support tool can be designed to work with any textbase including the MetaMetrics corpus and/or the EBSCO corpus. We have built such a tool for writing native-Lexile items and single sentence cloze reading items.

"Obtaining Lexile values for text passages in use in existing assessments." and *"Assessment item and assessment comparisons. There are some limited contexts in which it may be useful to NCES to compare different reading assessment instruments—for example, to compare the 12th grade NAEP to the NAAL. … Other comparisons at the passage and assessment level could include examination of the Lexile levels in passages administered in different years."* (Page 5)

MetaMetrics collaborates with a dozen test publishers in the development of reading and writing tests. These publishers routinely use the Lexile Analyzer at Lexile.com to pre-calibrate text passages prior to field-testing the items. There is no charge for using the Lexile Analyzer in this kind of product-development activity.

3.2 Areas of concern—semantic theory

"Separating ideational complexity from rarity." (Page 6)

Adams points out the difference between ideational complexity and rarity. This problem has been the focus of much research at MetaMetrics. From the data collected, we can infer that the heart of the problem is the confinement of rarity measures to the written word.

Words like "kitty" are frequently spoken, but infrequently written. Large frequency corpora of natural speech are hard, if not impossible, to find, and are naturally

biased towards words that are written more than spoken. Words such as "colossal," "enigma," and "ecstatic," are synonyms for "big," "puzzle," and "happy." The fundamental idea behind each of these words is easy to grasp. It's the infrequency of use—not the hardness of the concept—that makes these words harder than their same-meaning counterparts.

During MetaMetrics' research on the prediction of word difficulties on vocabulary tests, numerous variables have been examined (i.e., word length, part of speech, word frequency). The variables that show the most promise are the ones that take into account the Lexile difficulty of the text at which the word crosses the threshold from "rare inclusion" to "general inclusion." The problem with using threshold, or "first appearance," variables as a predictor of semantic demand in reading comprehension however is that they were built into the Lexile equation in the first place.

The good news is that the Lexile equation measures slices of text instead of individual words. Using the mean log word frequency prevents single words from holding too much sway.

"Other semantic issues. Published word frequency lists don't adequately represent new and/or technical terms, informal or slang words in current use; function words; alternate forms of presentation of equivalent terms; or proper nouns." (Page 6)

During the development of the Lexile Framework, the first word frequency measure used was the raw count of how often a given word appeared in a corpus of 5,088,721 words sampled from a broad range of school materials (Carroll, Davies, and Richman, 1971).

Currently, the Lexile Framework has a 305-million-word corpus that is also used to examine the semantic component of text. This corpus was assembled from approximately 15,000 texts measured by MetaMetrics for publishers from 1998 through January 2000. This corpus includes technical terms such as "byte" and "biodiversity." MetaMetrics is currently "tagging" words to indicate alternate forms (e.g., "color" and "colour"). When enough additional texts have been analyzed to make an adjustment to the corpus necessary and desirable, a study will be conducted to adjust the calibration equation such that the Lexile measure of a text based on the current corpus will be equivalent to the Lexile measure based on the new corpus.

3.3/3.4 Areas of concern—syntactic theory and psychometric issues

"Adequacy of sentence length.… It might be valuable to examine how such an insight could be used to improve the item selection, and design process." and *"Relative contribution of syntactic and semantic complexity. The panel suggested that, although syntactic and semantic complexity were admittedly highly correlated, any possible independent contribution of the two components of reading difficulty might be worth examination through more detailed statistical analysis."* and "*Explaining 'within-ensemble' variation and other issues related to item and passage validity."* (Page 7)

For five years, summer interns from Duke University and The University of North Carolina at Chapel Hill have operationalized dozens of new text variables and developed alternative transformations of existing variables in an attempt to explain the 150 Lexiles of RMS item error. Repeated failure to identify additional variables that could account for the persistent residual variation has led to an alternative hypothesis. It follows that although most of the reliable between-text passage variation is accounted for by the Lexile equation, when these text passages are turned into test items, uncontrolled "easiness" and "hardness" are added to the underlying text passage difficulty. It is the item engineering process that produces the RMSE, not the incompleteness of the Lexile Theory and the associated Lexile equation. This hypothesis is currently under investigation.

3.4 Areas of concern—psychometric issues

"Generalizing to other genres." (Page 8)

The Lexile Framework was developed to measure how well readers read literary and expository texts. The Lexile calibration equation was not developed to measure poems, recipes, lists, and plays. With these types of text, the features of the reading environment (for example, format and sentence structure) and interpretative skills take on extra importance when determining reader comprehension.

The Lexile levels do not translate directly to grade levels. In every classroom there is a range of readers and a range of materials. In a fifth-grade classroom there are readers that are far ahead of the rest and there are readers that are far behind the rest. To say that some books are "just right" for fifth graders assumes that all fifth graders are reading at the same level. The goal of the Lexile Framework is to match readers with texts at whatever level the reader is reading.

Just because a student is an excellent reader does not mean that he or she would comprehend a text typically found at a higher grade level. Without background knowledge some words might not have meaning. A high Lexile measure for a grade indicates that the student can read grade-level appropriate materials at a higher comprehension level (say 90%).

The real power of the Lexile Framework is in examining the growth of readers—wherever the reader may be in the development of his or her reading skills. Readers can be matched with texts that they are forecasted to read with 75% comprehension. As a reader grows, he or she can be matched with more demanding texts. And, as the texts become more demanding, then the reader grows.

"Applicability of the Lexile Framework to the reading assessment of second-language readers." (Page 8)

The panel report sets forth a tantalizing set of possible causes for differential item functioning (DIF) in second-language readers. It is suggested that such a research agenda could begin with the Spanish language.

The Lexile Framework now has a Spanish counterpart—"El Sistema Lexile." We have found that the Spanish Lexile Analyzer explains variation in Spanish item difficulties as well as it's English counterpart explains variation in English item difficulties. But, that does not mean that when either version is used on a second-language population that the measures are as valid as when the respective frameworks are used on the intended first language readers. The first research task would be to document DIF when the English framework is used on first language Spanish readers and when the Spanish framework is used with first language English readers. Given that DIF can be documented, then the interesting task begins of identifying the unwanted new dimensions that cause DIF. The Panel Report suggests several candidate causes: (1) use of idiomatic expressions and clichés, (2) common and uncommon roots making some "hard" words self-evident ("malnutrition") and some "easy" words quite difficult ("mop"), (3) shared and unshared cultural background, and (4) distinctions between the ability to read a text in English and the ability to process print in general. MetaMetrics would welcome an opportunity to collaborate on this important research topic or to support work by others.

"Exploring a modification of the Lexile Framework (LF) scale." (Pages 9 and 10)

It is not correct to assert that the Lexile Framework is based on an assumption that a 75% comprehension rate is "optimal." The accurate assertion is that 75% comprehension is the implied interpretation of what it means when a reader and text have the same Lexile measure. The assertion of "optimality" requires an objective, or stated purpose, for reading. Different objectives will lead to different optimal comprehension rates. If the objective is for the developing reader to improve reading skills by self-engaging with the text, then a 75% comprehension rate is a target that is probably close to optimal. On the other hand, if the objective is to design items for a reading test to maximize measurement efficiency, the optimal rate is probably closer to 50%. In the latter context there is no need to alter the Lexile Scale. Just use items that are 200 Lexiles higher than the center of the target population.
"Associating books and texts with NAEP reading achievement levels. NAEP reading assessment results group students into Basic, Proficient, and Advanced categories. Communicating what these categories mean to the general public is sometimes challenging." *(Page 10)*

NAEP currently annotates RPS scale values and proficiency categories with test items taken from released item pools. These annotations help the public appreciate what a "Proficient" reader can do that a "Basic" reader cannot.

One way to extend this annotation process is to identify well-known fiction and non-fiction texts, periodicals, and textbooks that have text measures corresponding to key proficiency ranges or scale score locations. Similarly, it would be possible to calculate the comprehension rate that say, an eighth-grade NAEP "Basic" reader would have with a sample of eighth-grade textbooks. "Basic" eighth-grade readers would experience 75% comprehension. At this level, the texts are not so hard that the reader experiences frustration and loses the meaning-thread of the text, but, at the same time, is not so easy

that the reader does not experience any challenge. "Proficient" eighth-grade readers would enjoy 90% comprehension of the material and "Advanced" readers would enjoy 95% comprehension. It may be that this correspondence between comprehension rate and on-grade level texts can be observed across all three grades assessed (4, 8, and 12).

The Panel Report states: "The measures of reading comprehension on which the Lexile Framework is based, however, represent only one of a number of comprehension tasks. It is possible that other NAEP comprehension tasks based on the NAEP reading framework will perform differently" (Page 10). It is "possible," but highly unlikely, that the NAEP comprehension tasks measure something different from what is specified in the Lexile equation.

We have now analyzed the observed item difficulty variation on some 100+ tests varying widely in item format, nominal grade level, type of passages, etc; fits to the Lexile theory are uniformly high. Similarly, person measures produced by a target test (e.g., SAT-9) and native-Lexile linking tests correlate in the range $r = .75$ to $r = .85$ (within grade). These results are consistent with alternate-form reliabilities for the respective tests. Disattenuated correlations among reading comprehension tests of comparable length approach unity. Factor analytic work beginning with Thurstone's (1946) reanalysis of Fred Davis' data (1944), the anchor test study (Loret et al, 1974; Bashaw and Rentz, 1977), Zwick's (1987) full information factor analysis of reading test data, and Reder (1996) all support a unidimensional reading comprehension construct identical to the construct "specified" by the Lexile Theory (Stenner, Smith, Horabin, and Smith, 1987).

One implication of these findings is that item format may not be as important as we have believed in the past. This does not mean that there is no item format main effect. Some ways of measuring reading comprehension are easier than others. However, within a format it is the Lexile equation that orders items/passages as to difficulty. Given the overwhelming consistency with which item difficulties on tests calling themselves reading comprehension tests are predictable from the Lexile Theory, developers of a test that did not fit would need to specify what construct other than "reading comprehension" their instrument was measuring, because that construct label has been appropriated for the dozens of tests measuring the common reading comprehension construct.

4. Summary and Conclusions

"The Lexile Framework has limited uses for NCES assessments, however, because of linguistic and psychometric issues, and therefore certain limitations on how the Lexile Framework might be applied." (Page 11)

The word "limited" in the above statement contradicts the bulk of the Panel's report. If there are outstanding limitations, they should be listed and justified. A rewording of the statement consistent with the rest of the Panel's report would be: "The Lexile Framework has several promising uses for NCES assessments. The Panel Report

has identified a number of research directions that could be undertaken to inform how broadly the Lexile Framework can be used in future applications."

References

Bashaw, W.L. & Rentz, R.R. (1977). The National Reference Scale for Reading: An application of the Rasch model. *Journal of Educational Measurement*, *14*, 161-179.

Carroll, J.B., Davies, P., & Richman, B. (1971). *Word frequency book*. Boston: Houghton Mifflin.

Davis, F. (1944). Fundamental factors of comprehension in reading. *Psychometrika*, *9*, 185-197.

Loret, P.G., Seder, A., Bianchini, J.C., & Vale, C.A. (1974). *Anchor Test Study Final Report: Project Report* and Vols 1-30. Berkeley, CA: Educational Testing Service. (ERIC Document Nos. ED 092 601 – ED 092 631).

Reder, S. (1996). Dimensionality and construct validity of the NALS assessment. In M.C. Smith (Ed.) *Literacy for the 21st Century research and practice: Forty-fifth yearbook of the National Reading Conference*.

Stenner, A.J., Smith, D.R., Horabin, I. & Smith, M. (1987). *Fit of the Lexile Theory to item difficulties on fourteen standardized reading comprehension tests*. Durham, NC: MetaMetrics, Inc.

Stenner, A.J., Smith, M., & Burdick, D.S. (1983). Toward a theory of construct definition. *Journal of Educational Measurement*, *20*(4), 305-315.

Thurstone, L.L. (1946). Note on a reanalysis of Davis' Reading Tests. *Psychometrika*, *11*(2). 185.

Wright, B.D. & Stenner, A.J. (1998). *Readability and reading ability*. Paper presented at the meeting of the Australian Council on Educational Research, Melbourne.

Zwick, R. (1987). Assessing the dimensionality of the NAEP reading data. *Journal of Educational Measurement*, *24*, 293-308.

LISTING OF NCES WORKING PAPERS TO DATE

Working papers can be downloaded as pdf files from the NCES Electronic Catalog (http://nces.ed.gov/pubsearch/). You can also contact Ralph Lee at (202) 502–7381 (ralph.lee@ed.gov) if you are interested in any of the following papers.

Listing of NCES Working Papers by Program Area

No.	Title	NCES contact
Baccalaureate and Beyond (B&B)		
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
Beginning Postsecondary Students (BPS) Longitudinal Study		
98–11	Beginning Postsecondary Students Longitudinal Study First Follow-up (BPS:96–98) Field Test Report	Aurora D'Amico
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
1999–15	Projected Postsecondary Outcomes of 1992 High School Graduates	Aurora D'Amico
2001-04	Beginning Postsecondary Students Longitudinal Study: 1996-2001 (BPS:1996/2001) Field Test Methodology Report	Paula Knepper
Common Core of Data (CCD)		
95–12	Rural Education Data User's Guide	Samuel Peng
96–19	Assessment and Analysis of School-Level Expenditures	William J. Fowler, Jr.
97–15	Customer Service Survey: Common Core of Data Coordinators	Lee Hoffman
97–43	Measuring Inflation in Public School Costs	William J. Fowler, Jr.
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
1999–03	Evaluation of the 1996–97 Nonfiscal Common Core of Data Surveys Data Collection, Processing, and Editing Cycle	Beth Young
2000–12	Coverage Evaluation of the 1994–95 Common Core of Data: Public Elementary/Secondary School Universe Survey	Beth Young
2000–13	Non-professional Staff in the Schools and Staffing Survey (SASS) and Common Core of Data (CCD)	Kerry Gruber
Data Development		
2000–16a	Lifelong Learning NCES Task Force: Final Report Volume I	Lisa Hudson
2000–16b	Lifelong Learning NCES Task Force: Final Report Volume II	Lisa Hudson
Decennial Census School District Project		
95–12	Rural Education Data User's Guide	Samuel Peng
96–04	Census Mapping Project/School District Data Book	Tai Phan
98–07	Decennial Census School District Project Planning Report	Tai Phan
Early Childhood Longitudinal Study (ECLS)		
96–08	How Accurate are Teacher Judgments of Students' Academic Performance?	Jerry West
96–18	Assessment of Social Competence, Adaptive Behaviors, and Approaches to Learning with Young Children	Jerry West
97–24	Formulating a Design for the ECLS: A Review of Longitudinal Studies	Jerry West
97–36	Measuring the Quality of Program Environments in Head Start and Other Early Childhood Programs: A Review and Recommendations for Future Research	Jerry West
1999–01	A Birth Cohort Study: Conceptual and Design Considerations and Rationale	Jerry West
2000–04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
2001–02	Measuring Father Involvement in Young Children's Lives: Recommendations for a Fatherhood Module for the ECLS-B	Jerry West
2001–03	Measures of Socio-Emotional Development in Middle Childhood	Elvira Hausken
2001–06	Papers from the Early Childhood Longitudinal Studies Program: Presented at the 2001 AERA and SRCD Meetings	Jerry West

No.	Title	NCES contact
Education Finance Statistics Center (EDFIN)		
94–05	Cost-of-Education Differentials Across the States	William J. Fowler, Jr.
96–19	Assessment and Analysis of School-Level Expenditures	William J. Fowler, Jr.
97–43	Measuring Inflation in Public School Costs	William J. Fowler, Jr.
98–04	Geographic Variations in Public Schools' Costs	William J. Fowler, Jr.
1999–16	Measuring Resources in Education: From Accounting to the Resource Cost Model Approach	William J. Fowler, Jr.
High School and Beyond (HS&B)		
95–12	Rural Education Data User's Guide	Samuel Peng
1999–05	Procedures Guide for Transcript Studies	Dawn Nelson
1999–06	1998 Revision of the Secondary School Taxonomy	Dawn Nelson
HS Transcript Studies		
1999–05	Procedures Guide for Transcript Studies	Dawn Nelson
1999–06	1998 Revision of the Secondary School Taxonomy	Dawn Nelson
International Adult Literacy Survey (IALS)		
97–33	Adult Literacy: An International Perspective	Marilyn Binkley
Integrated Postsecondary Education Data System (IPEDS)		
97–27	Pilot Test of IPEDS Finance Survey	Peter Stowe
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
2000–14	IPEDS Finance Data Comparisons Under the 1997 Financial Accounting Standards for Private, Not-for-Profit Institutes: A Concept Paper	Peter Stowe
National Assessment of Adult Literacy (NAAL)		
98–17	Developing the National Assessment of Adult Literacy: Recommendations from Stakeholders	Sheida White
1999–09a	1992 National Adult Literacy Survey: An Overview	Alex Sedlacek
1999–09b	1992 National Adult Literacy Survey: Sample Design	Alex Sedlacek
1999–09c	1992 National Adult Literacy Survey: Weighting and Population Estimates	Alex Sedlacek
1999–09d	1992 National Adult Literacy Survey: Development of the Survey Instruments	Alex Sedlacek
1999–09e	1992 National Adult Literacy Survey: Scaling and Proficiency Estimates	Alex Sedlacek
1999–09f	1992 National Adult Literacy Survey: Interpreting the Adult Literacy Scales and Literacy Levels	Alex Sedlacek
1999–09g	1992 National Adult Literacy Survey: Literacy Levels and the Response Probability Convention	Alex Sedlacek
2000–05	Secondary Statistical Modeling With the National Assessment of Adult Literacy: Implications for the Design of the Background Questionnaire	Sheida White
2000–06	Using Telephone and Mail Surveys as a Supplement or Alternative to Door-to-Door Surveys in the Assessment of Adult Literacy	Sheida White
2000–07	"How Much Literacy is Enough?" Issues in Defining and Reporting Performance Standards for the National Assessment of Adult Literacy	Sheida White
2000–08	Evaluation of the 1992 NALS Background Survey Questionnaire: An Analysis of Uses with Recommendations for Revisions	Sheida White
2000–09	Demographic Changes and Literacy Development in a Decade	Sheida White
2001-08	Assessing the Lexile Framework: Results of a Panel Meeting	Sheida White
National Assessment of Educational Progress (NAEP)		
95–12	Rural Education Data User's Guide	Samuel Peng
97–29	Can State Assessment Data be Used to Reduce State NAEP Sample Sizes?	Steven Gorman
97–30	ACT's NAEP Redesign Project: Assessment Design is the Key to Useful and Stable Assessment Results	Steven Gorman
97–31	NAEP Reconfigured: An Integrated Redesign of the National Assessment of Educational Progress	Steven Gorman
97–32	Innovative Solutions to Intractable Large Scale Assessment (Problem 2: Background Questionnaires)	Steven Gorman
97–37	Optimal Rating Procedures and Methodology for NAEP Open-ended Items	Steven Gorman

No.	Title	NCES contact
97–44	Development of a SASS 1993–94 School-Level Student Achievement Subfile: Using State Assessments and State NAEP, Feasibility Study	Michael Ross
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
1999–05	Procedures Guide for Transcript Studies	Dawn Nelson
1999–06	1998 Revision of the Secondary School Taxonomy	Dawn Nelson
2001-07	A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA)	Arnold Goldstein
2001-08	Assessing the Lexile Framework: Results of a Panel Meeting	Sheida White
National Education Longitudinal Study of 1988 (NELS:88)		
95–04	National Education Longitudinal Study of 1988: Second Follow-up Questionnaire Content Areas and Research Issues	Jeffrey Owings
95–05	National Education Longitudinal Study of 1988: Conducting Trend Analyses of NLS-72, HS&B, and NELS:88 Seniors	Jeffrey Owings
95–06	National Education Longitudinal Study of 1988: Conducting Cross-Cohort Comparisons Using HS&B, NAEP, and NELS:88 Academic Transcript Data	Jeffrey Owings
95–07	National Education Longitudinal Study of 1988: Conducting Trend Analyses HS&B and NELS:88 Sophomore Cohort Dropouts	Jeffrey Owings
95–12	Rural Education Data User's Guide	Samuel Peng
95–14	Empirical Evaluation of Social, Psychological, & Educational Construct Variables Used in NCES Surveys	Samuel Peng
96–03	National Education Longitudinal Study of 1988 (NELS:88) Research Framework and Issues	Jeffrey Owings
98–06	National Education Longitudinal Study of 1988 (NELS:88) Base Year through Second Follow-Up: Final Methodology Report	Ralph Lee
98–09	High School Curriculum Structure: Effects on Coursetaking and Achievement in Mathematics for High School Graduates—An Examination of Data from the National Education Longitudinal Study of 1988	Jeffrey Owings
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
1999–05	Procedures Guide for Transcript Studies	Dawn Nelson
1999–06	1998 Revision of the Secondary School Taxonomy	Dawn Nelson
1999–15	Projected Postsecondary Outcomes of 1992 High School Graduates	Aurora D'Amico
National Household Education Survey (NHES)		
95–12	Rural Education Data User's Guide	Samuel Peng
96–13	Estimation of Response Bias in the NHES:95 Adult Education Survey	Steven Kaufman
96–14	The 1995 National Household Education Survey: Reinterview Results for the Adult Education Component	Steven Kaufman
96–20	1991 National Household Education Survey (NHES:91) Questionnaires: Screener, Early Childhood Education, and Adult Education	Kathryn Chandler
96–21	1993 National Household Education Survey (NHES:93) Questionnaires: Screener, School Readiness, and School Safety and Discipline	Kathryn Chandler
96–22	1995 National Household Education Survey (NHES:95) Questionnaires: Screener, Early Childhood Program Participation, and Adult Education	Kathryn Chandler
96–29	Undercoverage Bias in Estimates of Characteristics of Adults and 0- to 2-Year-Olds in the 1995 National Household Education Survey (NHES:95)	Kathryn Chandler
96–30	Comparison of Estimates from the 1995 National Household Education Survey (NHES:95)	Kathryn Chandler
97–02	Telephone Coverage Bias and Recorded Interviews in the 1993 National Household Education Survey (NHES:93)	Kathryn Chandler
97–03	1991 and 1995 National Household Education Survey Questionnaires: NHES:91 Screener, NHES:91 Adult Education, NHES:95 Basic Screener, and NHES:95 Adult Education	Kathryn Chandler
97–04	Design, Data Collection, Monitoring, Interview Administration Time, and Data Editing in the 1993 National Household Education Survey (NHES:93)	Kathryn Chandler
97–05	Unit and Item Response, Weighting, and Imputation Procedures in the 1993 National Household Education Survey (NHES:93)	Kathryn Chandler
97–06	Unit and Item Response, Weighting, and Imputation Procedures in the 1995 National Household Education Survey (NHES:95)	Kathryn Chandler

No.	Title	NCES contact
97–08	Design, Data Collection, Interview Timing, and Data Editing in the 1995 National Household Education Survey	Kathryn Chandler
97–19	National Household Education Survey of 1995: Adult Education Course Coding Manual	Peter Stowe
97–20	National Household Education Survey of 1995: Adult Education Course Code Merge Files User's Guide	Peter Stowe
97–25	1996 National Household Education Survey (NHES:96) Questionnaires: Screener/Household and Library, Parent and Family Involvement in Education and Civic Involvement, Youth Civic Involvement, and Adult Civic Involvement	Kathryn Chandler
97–28	Comparison of Estimates in the 1996 National Household Education Survey	Kathryn Chandler
97–34	Comparison of Estimates from the 1993 National Household Education Survey	Kathryn Chandler
97–35	Design, Data Collection, Interview Administration Time, and Data Editing in the 1996 National Household Education Survey	Kathryn Chandler
97–38	Reinterview Results for the Parent and Youth Components of the 1996 National Household Education Survey	Kathryn Chandler
97–39	Undercoverage Bias in Estimates of Characteristics of Households and Adults in the 1996 National Household Education Survey	Kathryn Chandler
97–40	Unit and Item Response Rates, Weighting, and Imputation Procedures in the 1996 National Household Education Survey	Kathryn Chandler
98–03	Adult Education in the 1990s: A Report on the 1991 National Household Education Survey	Peter Stowe
98–10	Adult Education Participation Decisions and Barriers: Review of Conceptual Frameworks and Empirical Studies	Peter Stowe
National Longitudinal Study of the High School Class of 1972 (NLS-72)		
95–12	Rural Education Data User's Guide	Samuel Peng
National Postsecondary Student Aid Study (NPSAS)		
96–17	National Postsecondary Student Aid Study: 1996 Field Test Methodology Report	Andrew G. Malizio
2000–17	National Postsecondary Student Aid Study:2000 Field Test Methodology Report	Andrew G. Malizio
National Study of Postsecondary Faculty (NSOPF)		
97–26	Strategies for Improving Accuracy of Postsecondary Faculty Lists	Linda Zimbler
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
2000–01	1999 National Study of Postsecondary Faculty (NSOPF:99) Field Test Report	Linda Zimbler
Postsecondary Education Descriptive Analysis Reports (PEDAR)		
2000–11	Financial Aid Profile of Graduate Students in Science and Engineering	Aurora D'Amico
Private School Universe Survey (PSS)		
95–16	Intersurvey Consistency in NCES Private School Surveys	Steven Kaufman
95–17	Estimates of Expenditures for Private K–12 Schools	Stephen Broughman
96–16	Strategies for Collecting Finance Data from Private Schools	Stephen Broughman
96–26	Improving the Coverage of Private Elementary-Secondary Schools	Steven Kaufman
96–27	Intersurvey Consistency in NCES Private School Surveys for 1993–94	Steven Kaufman
97–07	The Determinants of Per-Pupil Expenditures in Private Elementary and Secondary Schools: An Exploratory Analysis	Stephen Broughman
97–22	Collection of Private School Finance Data: Development of a Questionnaire	Stephen Broughman
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
2000–04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
2000–15	Feasibility Report: School-Level Finance Pretest, Private School Questionnaire	Stephen Broughman
Recent College Graduates (RCG)		
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
Schools and Staffing Survey (SASS)		
94–01	Schools and Staffing Survey (SASS) Papers Presented at Meetings of the American Statistical Association	Dan Kasprzyk
94–02	Generalized Variance Estimate for Schools and Staffing Survey (SASS)	Dan Kasprzyk

No.	Title	NCES contact
94–03	1991 Schools and Staffing Survey (SASS) Reinterview Response Variance Report	Dan Kasprzyk
94–04	The Accuracy of Teachers' Self-reports on their Postsecondary Education: Teacher Transcript Study, Schools and Staffing Survey	Dan Kasprzyk
94–06	Six Papers on Teachers from the 1990–91 Schools and Staffing Survey and Other Related Surveys	Dan Kasprzyk
95–01	Schools and Staffing Survey: 1994 Papers Presented at the 1994 Meeting of the American Statistical Association	Dan Kasprzyk
95–02	QED Estimates of the 1990–91 Schools and Staffing Survey: Deriving and Comparing QED School Estimates with CCD Estimates	Dan Kasprzyk
95–03	Schools and Staffing Survey: 1990–91 SASS Cross-Questionnaire Analysis	Dan Kasprzyk
95–08	CCD Adjustment to the 1990–91 SASS: A Comparison of Estimates	Dan Kasprzyk
95–09	The Results of the 1993 Teacher List Validation Study (TLVS)	Dan Kasprzyk
95–10	The Results of the 1991–92 Teacher Follow-up Survey (TFS) Reinterview and Extensive Reconciliation	Dan Kasprzyk
95–11	Measuring Instruction, Curriculum Content, and Instructional Resources: The Status of Recent Work	Sharon Bobbitt & John Ralph
95–12	Rural Education Data User's Guide	Samuel Peng
95–14	Empirical Evaluation of Social, Psychological, & Educational Construct Variables Used in NCES Surveys	Samuel Peng
95–15	Classroom Instructional Processes: A Review of Existing Measurement Approaches and Their Applicability for the Teacher Follow-up Survey	Sharon Bobbitt
95–16	Intersurvey Consistency in NCES Private School Surveys	Steven Kaufman
95–18	An Agenda for Research on Teachers and Schools: Revisiting NCES' Schools and Staffing Survey	Dan Kasprzyk
96–01	Methodological Issues in the Study of Teachers' Careers: Critical Features of a Truly Longitudinal Study	Dan Kasprzyk
96–02	Schools and Staffing Survey (SASS): 1995 Selected papers presented at the 1995 Meeting of the American Statistical Association	Dan Kasprzyk
96–05	Cognitive Research on the Teacher Listing Form for the Schools and Staffing Survey	Dan Kasprzyk
96–06	The Schools and Staffing Survey (SASS) for 1998–99: Design Recommendations to Inform Broad Education Policy	Dan Kasprzyk
96–07	Should SASS Measure Instructional Processes and Teacher Effectiveness?	Dan Kasprzyk
96–09	Making Data Relevant for Policy Discussions: Redesigning the School Administrator Questionnaire for the 1998–99 SASS	Dan Kasprzyk
96–10	1998–99 Schools and Staffing Survey: Issues Related to Survey Depth	Dan Kasprzyk
96–11	Towards an Organizational Database on America's Schools: A Proposal for the Future of SASS, with comments on School Reform, Governance, and Finance	Dan Kasprzyk
96–12	Predictors of Retention, Transfer, and Attrition of Special and General Education Teachers: Data from the 1989 Teacher Followup Survey	Dan Kasprzyk
96–15	Nested Structures: District-Level Data in the Schools and Staffing Survey	Dan Kasprzyk
96–23	Linking Student Data to SASS: Why, When, How	Dan Kasprzyk
96–24	National Assessments of Teacher Quality	Dan Kasprzyk
96–25	Measures of Inservice Professional Development: Suggested Items for the 1998–1999 Schools and Staffing Survey	Dan Kasprzyk
96–28	Student Learning, Teaching Quality, and Professional Development: Theoretical Linkages, Current Measurement, and Recommendations for Future Data Collection	Mary Rollefson
97–01	Selected Papers on Education Surveys: Papers Presented at the 1996 Meeting of the American Statistical Association	Dan Kasprzyk
97–07	The Determinants of Per-Pupil Expenditures in Private Elementary and Secondary Schools: An Exploratory Analysis	Stephen Broughman
97–09	Status of Data on Crime and Violence in Schools: Final Report	Lee Hoffman
97–10	Report of Cognitive Research on the Public and Private School Teacher Questionnaires for the Schools and Staffing Survey 1993–94 School Year	Dan Kasprzyk
97–11	International Comparisons of Inservice Professional Development	Dan Kasprzyk
97–12	Measuring School Reform: Recommendations for Future SASS Data Collection	Mary Rollefson
97–14	Optimal Choice of Periodicities for the Schools and Staffing Survey: Modeling and Analysis	Steven Kaufman
97–18	Improving the Mail Return Rates of SASS Surveys: A Review of the Literature	Steven Kaufman
97–22	Collection of Private School Finance Data: Development of a Questionnaire	Stephen Broughman

No.	Title	NCES contact
97–23	Further Cognitive Research on the Schools and Staffing Survey (SASS) Teacher Listing Form	Dan Kasprzyk
97–41	Selected Papers on the Schools and Staffing Survey: Papers Presented at the 1997 Meeting of the American Statistical Association	Steve Kaufman
97–42	Improving the Measurement of Staffing Resources at the School Level: The Development of Recommendations for NCES for the Schools and Staffing Survey (SASS)	Mary Rollefson
97–44	Development of a SASS 1993–94 School-Level Student Achievement Subfile: Using State Assessments and State NAEP, Feasibility Study	Michael Ross
98–01	Collection of Public School Expenditure Data: Development of a Questionnaire	Stephen Broughman
98–02	Response Variance in the 1993–94 Schools and Staffing Survey: A Reinterview Report	Steven Kaufman
98–04	Geographic Variations in Public Schools' Costs	William J. Fowler, Jr.
98–05	SASS Documentation: 1993–94 SASS Student Sampling Problems; Solutions for Determining the Numerators for the SASS Private School (3B) Second-Stage Factors	Steven Kaufman
98–08	The Redesign of the Schools and Staffing Survey for 1999–2000: A Position Paper	Dan Kasprzyk
98–12	A Bootstrap Variance Estimator for Systematic PPS Sampling	Steven Kaufman
98–13	Response Variance in the 1994–95 Teacher Follow-up Survey	Steven Kaufman
98–14	Variance Estimation of Imputed Survey Data	Steven Kaufman
98–15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
98–16	A Feasibility Study of Longitudinal Design for Schools and Staffing Survey	Stephen Broughman
1999–02	Tracking Secondary Use of the Schools and Staffing Survey Data: Preliminary Results	Dan Kasprzyk
1999–04	Measuring Teacher Qualifications	Dan Kasprzyk
1999–07	Collection of Resource and Expenditure Data on the Schools and Staffing Survey	Stephen Broughman
1999–08	Measuring Classroom Instructional Processes: Using Survey and Case Study Fieldtest Results to Improve Item Construction	Dan Kasprzyk
1999–10	What Users Say About Schools and Staffing Survey Publications	Dan Kasprzyk
1999–12	1993–94 Schools and Staffing Survey: Data File User's Manual, Volume III: Public-Use Codebook	Kerry Gruber
1999–13	1993–94 Schools and Staffing Survey: Data File User's Manual, Volume IV: Bureau of Indian Affairs (BIA) Restricted-Use Codebook	Kerry Gruber
1999–14	1994–95 Teacher Followup Survey: Data File User's Manual, Restricted-Use Codebook	Kerry Gruber
1999–17	Secondary Use of the Schools and Staffing Survey Data	Susan Wiley
2000–04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
2000–10	A Research Agenda for the 1999–2000 Schools and Staffing Survey	Dan Kasprzyk
2000–13	Non-professional Staff in the Schools and Staffing Survey (SASS) and Common Core of Data (CCD)	Kerry Gruber
2000–18	Feasibility Report: School-Level Finance Pretest, Public School District Questionnaire	Stephen Broughman
Third International Mathematics and Science Study (TIMSS)		
2001–01	Cross-National Variation in Educational Preparation for Adulthood: From Early Adolescence to Young Adulthood	Elvira Hausken
2001-07	A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA)	Arnold Goldstein

Listing of NCES Working Papers by Subject

No.	Title	NCES contact
Adult education		
96–14	The 1995 National Household Education Survey: Reinterview Results for the Adult Education Component	Steven Kaufman
96–20	1991 National Household Education Survey (NHES:91) Questionnaires: Screener, Early Childhood Education, and Adult Education	Kathryn Chandler
96–22	1995 National Household Education Survey (NHES:95) Questionnaires: Screener, Early Childhood Program Participation, and Adult Education	Kathryn Chandler
98–03	Adult Education in the 1990s: A Report on the 1991 National Household Education Survey	Peter Stowe
98–10	Adult Education Participation Decisions and Barriers: Review of Conceptual Frameworks and Empirical Studies	Peter Stowe
1999–11	Data Sources on Lifelong Learning Available from the National Center for Education Statistics	Lisa Hudson
2000–16a	Lifelong Learning NCES Task Force: Final Report Volume I	Lisa Hudson
2000–16b	Lifelong Learning NCES Task Force: Final Report Volume II	Lisa Hudson
Adult literacy—see Literacy of adults		
American Indian – education		
1999–13	1993–94 Schools and Staffing Survey: Data File User's Manual, Volume IV: Bureau of Indian Affairs (BIA) Restricted-Use Codebook	Kerry Gruber
Assessment/achievement		
95–12	Rural Education Data User's Guide	Samuel Peng
95–13	Assessing Students with Disabilities and Limited English Proficiency	James Houser
97–29	Can State Assessment Data be Used to Reduce State NAEP Sample Sizes?	Larry Ogle
97–30	ACT's NAEP Redesign Project: Assessment Design is the Key to Useful and Stable Assessment Results	Larry Ogle
97–31	NAEP Reconfigured: An Integrated Redesign of the National Assessment of Educational Progress	Larry Ogle
97–32	Innovative Solutions to Intractable Large Scale Assessment (Problem 2: Background Questions)	Larry Ogle
97–37	Optimal Rating Procedures and Methodology for NAEP Open-ended Items	Larry Ogle
97–44	Development of a SASS 1993–94 School-Level Student Achievement Subfile: Using State Assessments and State NAEP, Feasibility Study	Michael Ross
98–09	High School Curriculum Structure: Effects on Coursetaking and Achievement in Mathematics for High School Graduates—An Examination of Data from the National Education Longitudinal Study of 1988	Jeffrey Owings
2001-07	A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA)	Arnold Goldstein
Beginning students in postsecondary education		
98–11	Beginning Postsecondary Students Longitudinal Study First Follow-up (BPS:96–98) Field Test Report	Aurora D'Amico
2001-04	Beginning Postsecondary Students Longitudinal Study: 1996-2001 (BPS:1996/2001) Field Test Methodology Report	Paula Knepper
Civic participation		
97–25	1996 National Household Education Survey (NHES:96) Questionnaires: Screener/Household and Library, Parent and Family Involvement in Education and Civic Involvement, Youth Civic Involvement, and Adult Civic Involvement	Kathryn Chandler
Climate of schools		
95–14	Empirical Evaluation of Social, Psychological, & Educational Construct Variables Used in NCES Surveys	Samuel Peng

No.	Title	NCES contact
Cost of education indices		
94–05	Cost-of-Education Differentials Across the States	William J. Fowler, Jr.
Course-taking		
95–12	Rural Education Data User's Guide	Samuel Peng
98–09	High School Curriculum Structure: Effects on Coursetaking and Achievement in Mathematics for High School Graduates—An Examination of Data from the National Education Longitudinal Study of 1988	Jeffrey Owings
1999–05	Procedures Guide for Transcript Studies	Dawn Nelson
1999–06	1998 Revision of the Secondary School Taxonomy	Dawn Nelson
Crime		
97–09	Status of Data on Crime and Violence in Schools: Final Report	Lee Hoffman
Curriculum		
95–11	Measuring Instruction, Curriculum Content, and Instructional Resources: The Status of Recent Work	Sharon Bobbitt & John Ralph
98–09	High School Curriculum Structure: Effects on Coursetaking and Achievement in Mathematics for High School Graduates—An Examination of Data from the National Education Longitudinal Study of 1988	Jeffrey Owings
Customer service		
1999–10	What Users Say About Schools and Staffing Survey Publications	Dan Kasprzyk
2000–02	Coordinating NCES Surveys: Options, Issues, Challenges, and Next Steps	Valena Plisko
2000–04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
Data quality		
97–13	Improving Data Quality in NCES: Database-to-Report Process	Susan Ahmed
Data warehouse		
2000–04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
Design effects		
2000–03	Strengths and Limitations of Using SUDAAN, Stata, and WesVarPC for Computing Variances from NCES Data Sets	Ralph Lee
Dropout rates, high school		
95–07	National Education Longitudinal Study of 1988: Conducting Trend Analyses HS&B and NELS:88 Sophomore Cohort Dropouts	Jeffrey Owings
Early childhood education		
96–20	1991 National Household Education Survey (NHES:91) Questionnaires: Screener, Early Childhood Education, and Adult Education	Kathryn Chandler
96–22	1995 National Household Education Survey (NHES:95) Questionnaires: Screener, Early Childhood Program Participation, and Adult Education	Kathryn Chandler
97–24	Formulating a Design for the ECLS: A Review of Longitudinal Studies	Jerry West
97–36	Measuring the Quality of Program Environments in Head Start and Other Early Childhood Programs: A Review and Recommendations for Future Research	Jerry West
1999–01	A Birth Cohort Study: Conceptual and Design Considerations and Rationale	Jerry West
2001–02	Measuring Father Involvement in Young Children's Lives: Recommendations for a Fatherhood Module for the ECLS-B	Jerry West
2001–03	Measures of Socio-Emotional Development in Middle School	Elvira Hausken
2001–06	Papers from the Early Childhood Longitudinal Studies Program: Presented at the 2001 AERA and SRCD Meetings	Jerry West

No.	Title	NCES contact
Educational attainment		
98–11	Beginning Postsecondary Students Longitudinal Study First Follow-up (BPS:96–98) Field Test Report	Aurora D'Amico
Educational research		
2000–02	Coordinating NCES Surveys: Options, Issues, Challenges, and Next Steps	Valena Plisko
Employment		
96–03	National Education Longitudinal Study of 1988 (NELS:88) Research Framework and Issues	Jeffrey Owings
98–11	Beginning Postsecondary Students Longitudinal Study First Follow-up (BPS:96–98) Field Test Report	Aurora D'Amico
2000–16a	Lifelong Learning NCES Task Force: Final Report Volume I	Lisa Hudson
2000–16b	Lifelong Learning NCES Task Force: Final Report Volume II	Lisa Hudson
2001–01	Cross-National Variation in Educational Preparation for Adulthood: From Early Adolescence to Young Adulthood	Elvira Hausken
Engineering		
2000–11	Financial Aid Profile of Graduate Students in Science and Engineering	Aurora D'Amico
Faculty – higher education		
97–26	Strategies for Improving Accuracy of Postsecondary Faculty Lists	Linda Zimbler
2000–01	1999 National Study of Postsecondary Faculty (NSOPF:99) Field Test Report	Linda Zimbler
Fathers – role in education		
2001–02	Measuring Father Involvement in Young Children's Lives: Recommendations for a Fatherhood Module for the ECLS-B	Jerry West
Finance – elementary and secondary schools		
94–05	Cost-of-Education Differentials Across the States	William J. Fowler, Jr.
96–19	Assessment and Analysis of School-Level Expenditures	William J. Fowler, Jr.
98–01	Collection of Public School Expenditure Data: Development of a Questionnaire	Stephen Broughman
1999–07	Collection of Resource and Expenditure Data on the Schools and Staffing Survey	Stephen Broughman
1999–16	Measuring Resources in Education: From Accounting to the Resource Cost Model Approach	William J. Fowler, Jr.
2000–18	Feasibility Report: School-Level Finance Pretest, Public School District Questionnaire	Stephen Broughman
Finance – postsecondary		
97–27	Pilot Test of IPEDS Finance Survey	Peter Stowe
2000–14	IPEDS Finance Data Comparisons Under the 1997 Financial Accounting Standards for Private, Not-for-Profit Institutes: A Concept Paper	Peter Stowe
Finance – private schools		
95–17	Estimates of Expenditures for Private K–12 Schools	Stephen Broughman
96–16	Strategies for Collecting Finance Data from Private Schools	Stephen Broughman
97–07	The Determinants of Per-Pupil Expenditures in Private Elementary and Secondary Schools: An Exploratory Analysis	Stephen Broughman
97–22	Collection of Private School Finance Data: Development of a Questionnaire	Stephen Broughman
1999–07	Collection of Resource and Expenditure Data on the Schools and Staffing Survey	Stephen Broughman
2000–15	Feasibility Report: School-Level Finance Pretest, Private School Questionnaire	Stephen Broughman
Geography		
98–04	Geographic Variations in Public Schools' Costs	William J. Fowler, Jr.
Graduate students		
2000–11	Financial Aid Profile of Graduate Students in Science and Engineering	Aurora D'Amico

No.	Title	NCES contact
Imputation		
2000–04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
Inflation		
97–43	Measuring Inflation in Public School Costs	William J. Fowler, Jr.
Institution data		
2000–01	1999 National Study of Postsecondary Faculty (NSOPF:99) Field Test Report	Linda Zimbler
Instructional resources and practices		
95–11	Measuring Instruction, Curriculum Content, and Instructional Resources: The Status of Recent Work	Sharon Bobbitt & John Ralph
1999–08	Measuring Classroom Instructional Processes: Using Survey and Case Study Field Test Results to Improve Item Construction	Dan Kasprzyk
International comparisons		
97–11	International Comparisons of Inservice Professional Development	Dan Kasprzyk
97–16	International Education Expenditure Comparability Study: Final Report, Volume I	Shelley Burns
97–17	International Education Expenditure Comparability Study: Final Report, Volume II, Quantitative Analysis of Expenditure Comparability	Shelley Burns
2001–01	Cross-National Variation in Educational Preparation for Adulthood: From Early Adolescence to Young Adulthood	Elvira Hausken
2001-07	A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA)	Arnold Goldstein
Libraries		
94–07	Data Comparability and Public Policy: New Interest in Public Library Data Papers Presented at Meetings of the American Statistical Association	Carrol Kindel
97–25	1996 National Household Education Survey (NHES:96) Questionnaires: Screener/Household and Library, Parent and Family Involvement in Education and Civic Involvement, Youth Civic Involvement, and Adult Civic Involvement	Kathryn Chandler
Limited English Proficiency		
95–13	Assessing Students with Disabilities and Limited English Proficiency	James Houser
Literacy of adults		
98–17	Developing the National Assessment of Adult Literacy: Recommendations from Stakeholders	Sheida White
1999–09a	1992 National Adult Literacy Survey: An Overview	Alex Sedlacek
1999–09b	1992 National Adult Literacy Survey: Sample Design	Alex Sedlacek
1999–09c	1992 National Adult Literacy Survey: Weighting and Population Estimates	Alex Sedlacek
1999–09d	1992 National Adult Literacy Survey: Development of the Survey Instruments	Alex Sedlacek
1999–09e	1992 National Adult Literacy Survey: Scaling and Proficiency Estimates	Alex Sedlacek
1999–09f	1992 National Adult Literacy Survey: Interpreting the Adult Literacy Scales and Literacy Levels	Alex Sedlacek
1999–09g	1992 National Adult Literacy Survey: Literacy Levels and the Response Probability Convention	Alex Sedlacek
1999–11	Data Sources on Lifelong Learning Available from the National Center for Education Statistics	Lisa Hudson
2000–05	Secondary Statistical Modeling With the National Assessment of Adult Literacy: Implications for the Design of the Background Questionnaire	Sheida White
2000–06	Using Telephone and Mail Surveys as a Supplement or Alternative to Door-to-Door Surveys in the Assessment of Adult Literacy	Sheida White
2000–07	"How Much Literacy is Enough?" Issues in Defining and Reporting Performance Standards for the National Assessment of Adult Literacy	Sheida White
2000–08	Evaluation of the 1992 NALS Background Survey Questionnaire: An Analysis of Uses with Recommendations for Revisions	Sheida White

No.	Title	NCES contact
2000–09	Demographic Changes and Literacy Development in a Decade	Sheida White
2001-08	Assessing the Lexile Framework: Results of a Panel Meeting	Sheida White
Literacy of adults – international		
97–33	Adult Literacy: An International Perspective	Marilyn Binkley
Mathematics		
98–09	High School Curriculum Structure: Effects on Coursetaking and Achievement in Mathematics for High School Graduates—An Examination of Data from the National Education Longitudinal Study of 1988	Jeffrey Owings
1999–08	Measuring Classroom Instructional Processes: Using Survey and Case Study Field Test Results to Improve Item Construction	Dan Kasprzyk
2001-07	A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA)	Arnold Goldstein
Parental involvement in education		
96–03	National Education Longitudinal Study of 1988 (NELS:88) Research Framework and Issues	Jeffrey Owings
97–25	1996 National Household Education Survey (NHES:96) Questionnaires: Screener/Household and Library, Parent and Family Involvement in Education and Civic Involvement, Youth Civic Involvement, and Adult Civic Involvement	Kathryn Chandler
1999–01	A Birth Cohort Study: Conceptual and Design Considerations and Rationale	Jerry West
2001–06	Papers from the Early Childhood Longitudinal Studies Program: Presented at the 2001 AERA and SRCD Meetings	Jerry West
Participation rates		
98–10	Adult Education Participation Decisions and Barriers: Review of Conceptual Frameworks and Empirical Studies	Peter Stowe
Postsecondary education		
1999–11	Data Sources on Lifelong Learning Available from the National Center for Education Statistics	Lisa Hudson
2000–16a	Lifelong Learning NCES Task Force: Final Report Volume I	Lisa Hudson
2000–16b	Lifelong Learning NCES Task Force: Final Report Volume II	Lisa Hudson
Postsecondary education – persistence and attainment		
98–11	Beginning Postsecondary Students Longitudinal Study First Follow-up (BPS:96–98) Field Test Report	Aurora D'Amico
1999–15	Projected Postsecondary Outcomes of 1992 High School Graduates	Aurora D'Amico
Postsecondary education – staff		
97–26	Strategies for Improving Accuracy of Postsecondary Faculty Lists	Linda Zimbler
2000–01	1999 National Study of Postsecondary Faculty (NSOPF:99) Field Test Report	Linda Zimbler
Principals		
2000–10	A Research Agenda for the 1999–2000 Schools and Staffing Survey	Dan Kasprzyk
Private schools		
96–16	Strategies for Collecting Finance Data from Private Schools	Stephen Broughman
97–07	The Determinants of Per-Pupil Expenditures in Private Elementary and Secondary Schools: An Exploratory Analysis	Stephen Broughman
97–22	Collection of Private School Finance Data: Development of a Questionnaire	Stephen Broughman
2000–13	Non-professional Staff in the Schools and Staffing Survey (SASS) and Common Core of Data (CCD)	Kerry Gruber
2000–15	Feasibility Report: School-Level Finance Pretest, Private School Questionnaire	Stephen Broughman
Projections of education statistics		
1999–15	Projected Postsecondary Outcomes of 1992 High School Graduates	Aurora D'Amico

No.	Title	NCES contact
Public school finance		
1999–16	Measuring Resources in Education: From Accounting to the Resource Cost Model Approach	William J. Fowler, Jr.
2000–18	Feasibility Report: School-Level Finance Pretest, Public School District Questionnaire	Stephen Broughman
Public schools		
97–43	Measuring Inflation in Public School Costs	William J. Fowler, Jr.
98–01	Collection of Public School Expenditure Data: Development of a Questionnaire	Stephen Broughman
98–04	Geographic Variations in Public Schools' Costs	William J. Fowler, Jr.
1999–02	Tracking Secondary Use of the Schools and Staffing Survey Data: Preliminary Results	Dan Kasprzyk
2000–12	Coverage Evaluation of the 1994–95 Public Elementary/Secondary School Universe Survey	Beth Young
2000–13	Non-professional Staff in the Schools and Staffing Survey (SASS) and Common Core of Data (CCD)	Kerry Gruber
Public schools – secondary		
98–09	High School Curriculum Structure: Effects on Coursetaking and Achievement in Mathematics for High School Graduates—An Examination of Data from the National Education Longitudinal Study of 1988	Jeffrey Owings
Reform, educational		
96–03	National Education Longitudinal Study of 1988 (NELS:88) Research Framework and Issues	Jeffrey Owings
Response rates		
98–02	Response Variance in the 1993–94 Schools and Staffing Survey: A Reinterview Report	Steven Kaufman
School districts		
2000–10	A Research Agenda for the 1999–2000 Schools and Staffing Survey	Dan Kasprzyk
School districts, public		
98–07	Decennial Census School District Project Planning Report	Tai Phan
1999–03	Evaluation of the 1996–97 Nonfiscal Common Core of Data Surveys Data Collection, Processing, and Editing Cycle	Beth Young
School districts, public – demographics of		
96–04	Census Mapping Project/School District Data Book	Tai Phan
Schools		
97–42	Improving the Measurement of Staffing Resources at the School Level: The Development of Recommendations for NCES for the Schools and Staffing Survey (SASS)	Mary Rollefson
98–08	The Redesign of the Schools and Staffing Survey for 1999–2000: A Position Paper	Dan Kasprzyk
1999–03	Evaluation of the 1996–97 Nonfiscal Common Core of Data Surveys Data Collection, Processing, and Editing Cycle	Beth Young
2000–10	A Research Agenda for the 1999–2000 Schools and Staffing Survey	Dan Kasprzyk
Schools – safety and discipline		
97–09	Status of Data on Crime and Violence in Schools: Final Report	Lee Hoffman
Science		
2000–11	Financial Aid Profile of Graduate Students in Science and Engineering	Aurora D'Amico
2001-07	A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA)	Arnold Goldstein
Software evaluation		
2000–03	Strengths and Limitations of Using SUDAAN, Stata, and WesVarPC for Computing Variances from NCES Data Sets	Ralph Lee

No.	Title	NCES contact
Staff		
97–42	Improving the Measurement of Staffing Resources at the School Level: The Development of Recommendations for NCES for the Schools and Staffing Survey (SASS)	Mary Rollefson
98–08	The Redesign of the Schools and Staffing Survey for 1999–2000: A Position Paper	Dan Kasprzyk
Staff – higher education institutions		
97–26	Strategies for Improving Accuracy of Postsecondary Faculty Lists	Linda Zimbler
Staff – nonprofessional		
2000–13	Non-professional Staff in the Schools and Staffing Survey (SASS) and Common Core of Data (CCD)	Kerry Gruber
State		
1999–03	Evaluation of the 1996–97 Nonfiscal Common Core of Data Surveys Data Collection, Processing, and Editing Cycle	Beth Young
Statistical methodology		
97–21	Statistics for Policymakers or Everything You Wanted to Know About Statistics But Thought You Could Never Understand	Susan Ahmed
Students with disabilities		
95–13	Assessing Students with Disabilities and Limited English Proficiency	James Houser
Survey methodology		
96–17	National Postsecondary Student Aid Study: 1996 Field Test Methodology Report	Andrew G. Malizio
97–15	Customer Service Survey: Common Core of Data Coordinators	Lee Hoffman
97–35	Design, Data Collection, Interview Administration Time, and Data Editing in the 1996 National Household Education Survey	Kathryn Chandler
98–06	National Education Longitudinal Study of 1988 (NELS:88) Base Year through Second Follow-Up: Final Methodology Report	Ralph Lee
98–11	Beginning Postsecondary Students Longitudinal Study First Follow-up (BPS:96–98) Field Test Report	Aurora D'Amico
98–16	A Feasibility Study of Longitudinal Design for Schools and Staffing Survey	Stephen Broughman
1999–07	Collection of Resource and Expenditure Data on the Schools and Staffing Survey	Stephen Broughman
1999–17	Secondary Use of the Schools and Staffing Survey Data	Susan Wiley
2000–01	1999 National Study of Postsecondary Faculty (NSOPF:99) Field Test Report	Linda Zimbler
2000–02	Coordinating NCES Surveys: Options, Issues, Challenges, and Next Steps	Valena Plisko
2000–04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
2000–12	Coverage Evaluation of the 1994–95 Public Elementary/Secondary School Universe Survey	Beth Young
2000–17	National Postsecondary Student Aid Study:2000 Field Test Methodology Report	Andrew G. Malizio
2001-04	Beginning Postsecondary Students Longitudinal Study: 1996-2001 (BPS:1996/2001) Field Test Methodology Report	Paula Knepper
2001-07	A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA)	Arnold Goldstein
Teachers		
98–13	Response Variance in the 1994–95 Teacher Follow-up Survey	Steven Kaufman
1999–14	1994–95 Teacher Followup Survey: Data File User's Manual, Restricted-Use Codebook	Kerry Gruber
2000–10	A Research Agenda for the 1999–2000 Schools and Staffing Survey	Dan Kasprzyk
Teachers – instructional practices of		
98–08	The Redesign of the Schools and Staffing Survey for 1999–2000: A Position Paper	Dan Kasprzyk
Teachers – opinions regarding safety		
98–08	The Redesign of the Schools and Staffing Survey for 1999–2000: A Position Paper	Dan Kasprzyk

No.	Title	NCES contact
Teachers – performance evaluations		
1999–04	Measuring Teacher Qualifications	Dan Kasprzyk
Teachers – qualifications of		
1999–04	Measuring Teacher Qualifications	Dan Kasprzyk
Teachers – salaries of		
94–05	Cost-of-Education Differentials Across the States	William J. Fowler, Jr.
Training		
2000–16a	Lifelong Learning NCES Task Force: Final Report Volume I	Lisa Hudson
2000–16b	Lifelong Learning NCES Task Force: Final Report Volume II	Lisa Hudson
Variance estimation		
2000–03	Strengths and Limitations of Using SUDAAN, Stata, and WesVarPC for Computing Variances from NCES Data Sets	Ralph Lee
2000–04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
Violence		
97–09	Status of Data on Crime and Violence in Schools: Final Report	Lee Hoffman
Vocational education		
95–12	Rural Education Data User's Guide	Samuel Peng
1999–05	Procedures Guide for Transcript Studies	Dawn Nelson
1999–06	1998 Revision of the Secondary School Taxonomy	Dawn Nelson