# Applying Usability Measures to Assess Textbooks

Philip Kortum, Michelle Hebl and Frederick L. Oswald
Rice University, Houston, TX

The assessment of usability is a common task for human factors professionals. The System Usability Scale (SUS) is a survey tool that has been supported empirically as a psychometrically robust, valid and highly adaptive way to measure subjective usability. One area in which the SUS has not been applied yet, but could have direct and important applicability, is in the assessment of textbooks. Although usability (and a SUS score) is not typically thought of as a textbook metric, the International Organization for Standardization (ISO) metrics of effectiveness, efficiency and satisfaction map well onto key attributes that describe a good textbook. Here we describe the results of an initial effort to explore using the SUS to measure textbook usability. A total of 319 participants completed the SUS and indicated the usability of the textbooks required in their classes in a given semester. In total, participants rated 169 unique textbook titles. Results showed that only 32% of Science, Technology, Engineering and Math (STEM) textbooks were rated as acceptable, with almost twice as many non-STEM textbooks (63%) rated as acceptable. Specific attributes of textbook design that support usability are discussed.

## INTRODUCTION

One of the very common tasks that human factors professionals do is to measure the usability of a given product, service, or system. They do this in order to insure that a target audience can use a product to perform a specified task successfully. There are a number of ways that assessing usability can be accomplished. If time and money are significant constraining factors, techniques such as heuristic evaluation (Nielsen, 1994) and cognitive walkthroughs (Wharton, Rieman, Lewis & Polson, 1994) can be used to gain a reasonable estimation of the product's usability. However, these techniques generally are viewed as substitutes for the benefit of employing real, representative users in the evaluation process. Users are brought into a laboratory, given the product and then asked to perform a number of tasks with that product. Measures of a product's effectiveness, efficiency and satisfaction are collected, along with observations of critical-use errors, and this helps to quantify how usable that product is under those circumstances. This type of laboratory testing is the gold standard for the assessment of usability, and it tends to produce reliable results if the experimental design is solid, if the users represent the true user of the product, and if the tasks selected mirror the tasks that real users would do in the field.

Another way to collect usability information is to ask users, in a retrospective fashion, to assess the usability of a product or service. There are a large number of survey instruments that measure usability in this way, including the Computer Usability System Questionnaire (CSUQ, Lewis, 1995), the Post-study System Usability Questionnaire (PSSUQ, Lewis, 1995), the Software Usability Measurement Inventory (SUMI, (Kirakowski & Corbett, 1993), the System Usability Scale (SUS, Brooke, 1996), the Usefulness, Satisfaction and Ease of Use Questionnaire (USE, Lund, 2001), and the Website Analysis and Measurement Inventory (WAMMI, Kirakowski, Claridge, & Whitehand, 1998), just to name a few of the most popular instruments. All of these instruments collect information about the usability of a system after it has been used, and provide a quick and easy way for the practitioner to gather quantitative data on the usability of a product.

As one might ascertain from the names of some of these instruments, they can be applied only to certain kinds of interfaces. The System Usability Scale is one of the most versatile assessment instruments in use today. It is non-proprietary, so it is free to use, and it is simple and quick to administer, making it useful in situations in which multiple administrations are planned. The SUS is also technology-agnostic, meaning it can be used to measure the usability of any number of products and services. Indeed, it has been used to measure everything from websites and voting machines to microwaves and telephones (e.g. Bangor, Kortum & Miller, 2008; Byrne, Greene & Everett, 2007; Kortum & Bangor, 2013; Sauro, 2011).

Because the SUS has been used extensively in industry and academia alike, it is known to have consistently high reliability and validity measures (Bangor, Kortum, & Miller, 2008; Sauro & Dumas, 2009). Further, this extensive use means that there are

Copyright 2014 Human Factors and Ergonomics Society. DOI 10.1177/1541931214581281

numerous comparative studies available to help researchers determine where their scores fit in the greater universe of scores (Lewis & Sauro, 2009; Bangor, Kortum, & Miller, 2008). Taken together, all of these attributes make the SUS an excellent instrument for measuring the usability of almost anything that people use.

There is one area that the SUS has not yet been applied, but one in which it would seem to have direct and important applicability: the assessment of textbooks. Instructors spend a significant amount of time trying to select textbooks that will convey the material in an understandable way, make students want to engage in the learning process, and help students master the material in the minimum time possible. These selection criteria map almost perfectly to the metrics of usability described by the International Organization for Standardization governing the measurement of usability, ISO 9241-11 (ISO, 1998).

In this standard, *usability* is described as having three dimensions: The first dimension is *effectiveness*, which describes a user's ability to accomplish a task with a minimum number of errors. The second dimension is *efficiency*, which describes the ability to complete those tasks in the most efficient manner possible. The third dimension is *satisfaction*, which describes the user's overall attitude about a product's fitness for its purpose and satisfaction with the manner in which it operates.

This direct mapping would suggest that the SUS might be a valuable tool to help assess the 'goodness' of textbooks. There seem to be precious few quantitative measures of textbooks. Much of the research has focused on using text readability measures, with mixed results (Flory, Phillips & Tassin, 1992; Spinks & Wells, 1993). Others have focused on more qualitative measures of good textbooks, such as structure, organization and pedagogy representation (Ciborowski, 1988; Lepionka, 2008), but provide no specific ways to measure those variables. Other sources that are used currently to assess the goodness of textbooks come from either marketing material (which is clearly biased) or in the form of user reviews on reseller sites, such as Amazon.com or BarnesandNoble.com. Although these user reviews may prove useful, they are not validated measures, and the dimensions along which users are rating the textbooks are almost completely unknown.

If the SUS could be used to assess the usability of textbooks, that usability might serve as a sufficient proxy for 'fitness for use' that would allow students to convey consistent information about the goodness of a textbook. Likewise, it might provide valuable information for instructors to help them in their textbook selection process using a common scale.

## METHODS

*Participants.* We recruited 319 subjects to evaluate the usability of the textbooks that they were currently using. All of the participants were undergraduate students at a private university who participated in the study for credit in partial fulfillment of course requirements. There were 175 female participants and 144 male participants in the study. No other selection criteria were used in recruitment

*Materials and Procedure.* After signing up for the experiment, participants were directed to a website where they completed an IRB approved consent form before beginning the study. They were then asked to record the name of a textbook they were using that semester, including relevant information regarding the textbook authors and version. The System Usability Scale was then presented (Table 1), along with 10 other questions adapted from past gender-related research and best-practices teaching techniques (Dasgupta, 2011; Dweck, 2008; Froyd, 2008) (Table 2), and one question that asked about their high school preparation in the content area covered by the textbook. All of the questions were rated on a 5 point Likert scale, ranging from 'Strongly Disagree' to 'Strongly Agree', except for the question regarding their degree of high school preparation, which used a 5-point Likert scale ranging from 'very poor' to 'excellent.' Scoring of the 10 SUS questions followed Brooke (1996), and resulted in a score ranging from 0-100.

Upon completion of all these ratings for a given textbook, participants were asked if they had another textbook they would like to rate. If they said yes, then another rating form was presented. If not, they were thanked, taken to a debriefing form, and the study ended.

**Table1:** System Usability Scale (SUS) Questions

| | |
|---|---|
| 1 | I think that I would like to use this product frequently. |
| 2 | I found the product unnecessarily complex. |
| 3 | I thought the product was easy to use. |
| 4 | I think that I would need the support of a technical person to be able to use this product. |
| 5 | I found the various functions in the product were well integrated. |
| 6 | I thought there was too much inconsistency in this product. |
| 7 | I imagine that most people would learn to use this product very quickly. |
| 8 | I found the product very awkward to use. |
| 9 | I felt very confident using the product. |
| 10 | I needed to learn a lot of things before I could get going with this product. |

**Table 2:** Textbook Attribute Questions

| | |
|---|---|
| 1 | This book included examples of humanitarian solutions or ways in which the knowledge gained from this field could be applied to help others. |
| 2 | This book focused on basic concepts of science. |
| 3 | This book involved active learning exercises, encouraged demonstrations, or got you to think about actual applications. |
| 4 | This book was focused on learning-by-doing rather than focused on reading about other people's findings. |
| 5 | This book included photographs of people. |
| 6 | This book included photographs of women. |
| 7 | This book included photographs of minorities. |
| 8 | The way this book was written and the content made you feel like YOU personally belong in this field. |
| 9 | The way this book was written made you think that people who want to succeed in this content area could do so if they simply put in enough effort. |
| 10 | The way this book was written made you think that people who want to succeed in this content area must be born with natural talents that make them good at this. |

## RESULTS

There were 746 textbook ratings comprising 169 unique textbook titles. The largest number of ratings for a given textbook was 74, with an average of 4.4 ratings per textbook. Average textbook SUS scores ranged from 0 to 100, with an overall average of 65.2. The distribution of these scores is shown in Figure 1.

Textbooks were then categorized as belonging to a STEM field (Science, Technology, Engineering, Math) or non-STEM field. There is not a single, universally accepted definition of what constitutes a STEM field (Wasem, 2012) and we took a conservative approach, excluding most of the social sciences, with the exception of neuropsychology. The average rating for non-STEM books (M=71.4) was statistically significantly higher than the rating for STEM textbooks (M=59.3), $p < .0001$. The distribution of SUS scores based on these categories is shown in Figure 2.

There were moderate to strong correlations between SUS scores and some of the textbook attributes. These correlations are shown in Table 3 for both STEM and non-STEM textbooks.

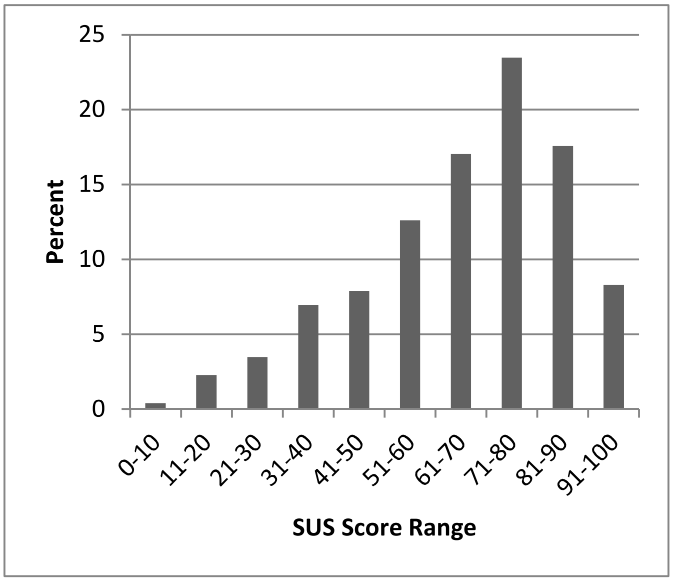

**Figure 1:** Distribution of Individual SUS Scores in the Study

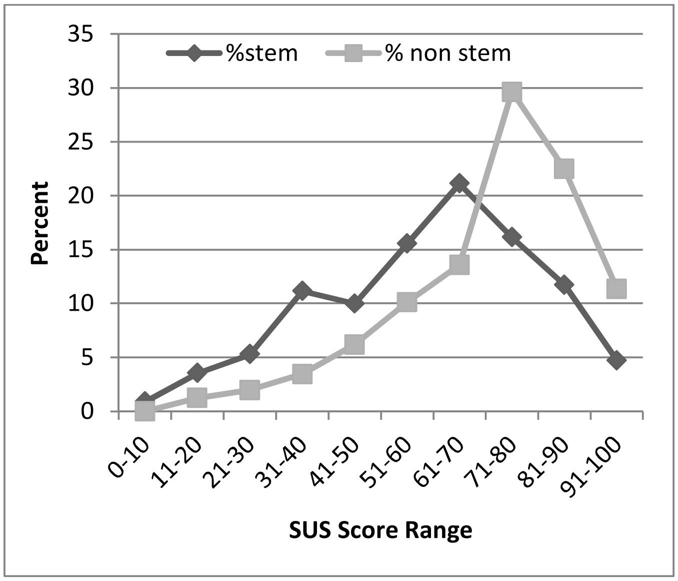

**Figure 2:** Distribution of Individual SUS Scores for Textbooks in STEM and Non-STEM Fields

**Table 3**: Correlations between Usability (SUS) score with Attributes of STEM and non-STEM Textbooks

| **Textbook Attribute** | **STEM** | | **Non-STEM** | |
|---|---|---|---|---|
| | ***r*** | ***p*** | ***r*** | ***p*** |
| This book included examples of humanitarian solutions or ways in which the knowledge gained from this field could be applied to help others. | .30 | **<.01** | .16 | .16 |
| This book focused on basic concepts of science. | .12 | .27 | -.14 | .21 |
| This book involved active learning exercises, encouraged demonstrations, or got you to think about actual applications. | .39 | **<.01** | .16 | .15 |
| This book was focused on learning-by-doing rather than focused on reading about other people's findings. | .15 | .17 | .17 | .12 |
| This book included photographs of people. | .23 | **.04** | .30 | **<.01** |
| This book included photographs of women. | .27 | **.01** | .30 | **<.01** |
| This book included photographs of minorities. | .11 | .31 | .28 | **.01** |
| The way this book was written and the content made you feel like YOU personally belong in this field. | .39 | **<.01** | .44 | **<.01** |
| The way this book was written made you think that people who want to succeed in this content area could do so if they simply put in enough effort. | .53 | **<.01** | .17 | .13 |
| The way this book was written made you think that people who want to succeed in this content area must be born with natural talents that make them good at this. | -.50 | **<.01** | -.26 | **.02** |
| How would you rate your high school preparation in the content area of this book? | -.22 | **.04** | .18 | .10 |

*Note.* Significant correlations (*p* < .05) are shown in boldface.

## DISCUSSION

Although 'usability' is probably not the first thing that comes to mind when describing a textbook, we believe that the ISO constructs of effectiveness, efficiency and satisfaction map well onto how users think about the textbook as a tool for learning. It is immediately clear that students are assessing the usability of their textbooks over a wide range. With scores ranging from 0 to 100, the entire range of the SUS is being utilized, suggesting that the SUS is capturing some element of the goodness of the textbook. This range of scores mirrors that of other less well-defined measures of goodness, such as the star-ratings found on popular on-line retailers like Amazon.com.

Page count is one metric that could trivially explain this wide range of SUS scores. Students could be basing their assessment of the usability of the book solely on how long it is. Short books would be 'usable' and long books would be 'unusable'. However, we found that page count is uncorrelated with the SUS scores obtained in this study (r = .02, *p* =.86).

The correlations between SUS scores and the attribute questions suggest some interesting elements of textbook style that may promote usability. As one might suspect, books that involve a lot of active learning in the form of exercises predict a higher usability score for STEM textbooks. Interestingly, the inclusion of photographs appears to have a positive impact across the board. Elements of writing style that allow the student to see themselves doing this kind of work in the field showed a strong positive correlation with textbook usability in STEM, as did a style that emphasized an ability to master the material through effort, not natural talent. Finally, preparation for the course in high school had little impact on usability associated with non-STEM textbooks, but a moderate effect on STEM texts.

Interpretation of the usabilty scores is clearly an important element of the SUS's utility as a rating mechanism. Bangor, Kortum, and Miller (2009) proposed an adjective rating scale that described the usability of a product from 'worst imaginable' to 'best imaginable' and placed SUS scores along this continuum. Further, they marked the boundaries of this continuum along an acceptability scale, marking items as 'acceptable,' 'marginally acceptable,' and 'unacceptable'. If we look at the average SUS score by textbook title, for STEM and non-STEM textbooks, and plot them according to this acceptability scale, we see that only 32% of the rated STEM textbooks are judged as 'acceptable.' Although non-STEM textbooks fare better, over 35% of them still fall below the 'acceptable' mark (Figure 3). No student should have to learn from a textbook that is not supporting their ability to do so, and these data suggest that too many textbooks fall into this category.

Clearly, the study provides a research design and some promising results for measuring the usability and attributes of STEM and non-STEM textbooks. Future research should involve larger sample sizes, a larger, broader demographic, and a much larger range of titles to approach more definitive claims about the efficacy of the SUS measure for the purpose of rating textbooks, but the current results certainly provide a start.

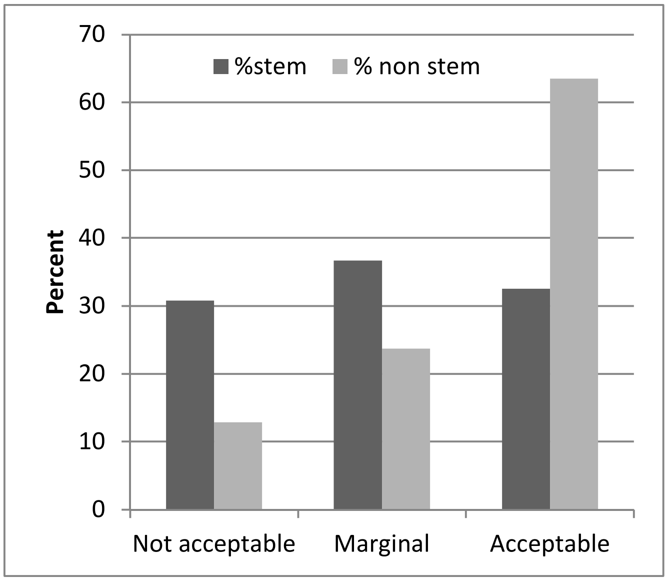

**Figure 3:** Percentage of STEM and non-STEM Textbooks, Categorized by Acceptability (Bangor, Kortum, & Miller, 2009)

## CONCLUSIONS

The System Usability Scale appears to be an interesting and innovative way to examine the goodness of textbooks, using the construct of usability as a basis for such an examination. With sufficient data, students and instructors alike could use the information to choose instructional material that best supported the learning goals of the class. Incorporation of the SUS into more formalized and widespread rating systems (like Amazon.com) would be the best way to collect and disseminate this information. Clearly, much work remains to be done before these kinds of wide-scale efforts should be undertaken; however, this study demonstrates that measuring textbook usability could be a promising avenue in STEM education.

## ACKNOWLDGEMENTS

We would like to thank Jonika Tannous for her assistance in collecting and coding the data.

## REFERENCES

Bangor, A., Kortum, P. & Miller, J. (2008). An empirical evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction, 24*:6,574-594.

Bangor, A., Kortum, P. & Miller, J.A. (2009). Determining what individual SUS scores mean: adding an adjective rating scale. *Journal of Usability Studies*, 4(3), 114-123.

Brooke, J. (1996). SUS: a "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & A. L. McClelland (Eds.),*Usability evaluation in industry*. London: Taylor &Francis.

Byrne, M. D., Greene, K. K., & Everett, S. P. (2007). Usability of voting systems: Baseline data for paper, punch cards, and lever machines. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 171-180). ACM.

Ciborowski, J. (1988). *Improving Textbook Usability.* Final Report.

Dasgupta, N. (2011). Ingroup experts and peers as social vaccines who inoculate the self-concept: The stereotype inoculation model. *Psychological Inquiry, 22*, 231-246.

Dweck, C. S. (2008). *Mindset: The new psychology of success.* NY: Ballantine Books.

Flory, S. M., Phillips Jr, T. J., & Tassin, M. F. (1992). Measuring readability: A comparison of accounting textbooks. *Journal of Accounting Education, 10*(1), 151-161.

Froyd, J. E. (2008). White paper on promising practices in undergraduate STEM education. *Evidence on Promising Practices in Undergraduate Science, Technology, Engineering, and Mathematics (STEM) Education Project, The National Academies Board on Science Education.*

ISO. (1998). *Ergonomic Requirements for office work with visual display terminal (VDT's) – Part 11: Guidance on Usability* (ISO 9241-11(E)). Geneva, Switzerland: International Organization for Standardization.

Kirakowski, J., Claridge, N., & Whitehand, R. (1998). Human centered measures of success in Web design. *Conference Proceedings of the 4th Conference on Human Factors and the Web.*

Kirakowski, J., & Corbett, M. (1993). SUMI: The Software Measurement Inventory. *British Journal of Educational Technology, 24,* 210–212.

Kortum, P. & Bangor, A. (2013). Usability ratings for everyday products measured with the System Usability Scale (SUS). *International Journal of Human Computer Interaction 29*(2), 67-76.

Lepionka, M.E. (2008). *Writing and developing your college textbook: a comprehensive guide to textbook authorship and higher education publishing*. Gloucester, MA: Atlantic Path Publishing.

Lewis, J. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction, 7* (1), 57-78.

Lund, A. M. (2001). Measuring Usability with the USE Questionnaire. STC Usability SIG Newsletter.

Nielsen, J. (1994). Heuristic evaluation. *Usability Inspection Methods, 24*, 413.

Sauro, J. (2011). *Measuring usability with the system usability scale (SUS).*

Sauro, J., & Dumas, J. S. (2009). Comparison of three one-question, post-task usability questionnaires. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1599-1608). ACM.

Spinks, N., & Wells, B. (1993). Readability: A textbook selection criterion. *Journal of Education for Business, 69*(2), 83-87.

Wasem, R. E. (2012, May). *Immigration of foreign nationals with Science, Technology, Engineering, and Mathematics (STEM) degrees.* Congressional Research Service, Library of Congress.

Wharton, C., Rieman, J., Lewis, C., & Polson, P. (1994). The cognitive walkthrough method: A practitioner's guide. In *Usability inspection methods* Jakob Nielsen and Robert L. Mack (Eds.) (pp. 105-140). John Wiley & Sons, Inc.