THE IOWA SILENT READING TEST'S COMPREHENSION SECTION: LOCAL NORMS AND PREDICTIVE VALIDITY FOR USABILITY STUDIES

James R. Lewis
International Business Machines, Corp.
Boca Raton, FL

The Iowa Silent Reading Test (ISRT), Level 2, Form E, Test 2 (Reading Comprehension) Part A was used to assess 70 participants' reading comprehension skills prior to performing a sequence of tasks in a series of seven printer usability studies (10 participants per printer). This part of the ISRT measures a person's ability to answer questions based on short passages to which the participant has ready access, and requires 26 minutes to administer. Data were collected to provide local norms for future studies, and for use as a covariate if necessary. The participants for these studies were obtained through a temporary employment agency. The distribution of scores for these participants is provided in this paper to describe local norms for this type of population, commonly used in such industrial usability studies. The predictive validity of this test score for the overall percentage of tasks successfully completed was found to be significant ($r = .25$, $p = .035$). The predictive validity of the ISRT score for ratings of satisfaction with user documentation was also significant ($r = .35$, $p = .0028$). Analysis of variance results suggested that participants with better reading comprehension skill generally performed better and were more critical of the user documentation.

INTRODUCTION

Many product usability studies require participants to use product documentation to accomplish specified tasks. In such studies, a participant's ability to read and understand the product documentation may have a substantial effect on his or her task performance, such as the rate of successful task completion. The purpose of this report is to provide data from a series of printer usability studies in which a portion of the Iowa Silent Reading Test was used to assess participants' reading comprehension skill. The use of this data for local norms and as a potential statistical covariate will be discussed.

The Iowa Silent Reading Test

The current version of the Iowa Silent Reading Test (ISRT) was published in 1973, and has been most recently reviewed in the Eighth Mental Measurements Yearbook (Buros, 1978). The reviews were generally favorable. For example:

> Given the range of tests available, the ISRT is an outstanding candidate and can be recommended for use. (Filby, 1978, p. 1197)

> The ISRT is likely the best test series currently available for the assessment of reading skills at the grade 6 level and up. (Hakstian, 1978, p. 1199).

> In terms of technical competence, the publishers have maintained the highest measurement standards in its preparation. (Hunter and Hoepfner, 1978, p. 1201).

The ISRT consists of several subtests, such as vocabulary, comprehension, directed reading, efficiency and power. The reliability coefficients for these subtests were reported to range from .77 to .90. Three levels of the ISRT have been published, Level 1 for grades 6 to 9, Level 2 for grades 9 to 14, and Level 3 for academically accelerated grades 11 to 12 and college students. However, no normative information is given for Level 3. Norms for Levels 1 and 2 are restricted to students from the appropriate grades. No information is given regarding validity studies, other than a brief appeal to content validity.

The ISRT Manual of Directions (1973) suggests that administration of the entire test battery requires 2 hours and 16 minutes. The time required to administer Reading Comprehension Part A (the subtest of interest in this paper) is 26 minutes.

Reading Comprehension and the Assessment of Product Usability

Product usability is frequently assessed by having participants perform realistic tasks using the product. In some cases, competitive products or alternative designs may be studied, and compared against various performance and preference measures taken during the usability test. In other cases, a product may be iteratively tested, and the usability data may be used to determine the effectiveness of design changes. When a participant is expected to use an instruction manual as an aid to performing tasks in such tests, it is important to measure the individual's ability to comprehend what is read.

Although random assignment of participants to products can be expected to result in equal average reading comprehension among experimental conditions, true random assignment may not be possible in some industrial situations. When iterative testing is conducted, the quality of participants may change over time. If competitive products are being evaluated, the evaluations may be conducted in different places at different times. A measure of reading comprehension can be used to determine

whether the experimental groups were composed of participants with approximately equal reading comprehension abilities. If the groups are found to differ in this ability, then the reading comprehension score can be used as a covariate in an analysis of covariance (ANCOVA) (Myers, 1979), allowing the effect of this individual difference to be removed from the other usability measures.

In some cases, the reading comprehension score may be used to select or categorize participants. For example, it may be reasonable to study only participants with reading comprehension skills in the lowest quartile in order to ensure that the product can be used by someone with below average reading comprehension skills. Such selection or categorization cannot be done unless appropriate norms exist against which to judge a reading comprehension score.

The Need for Local Norms

In industrial usability studies, it is a common practice to hire participants from temporary employment agencies. Since the ISRT Manual of Directions only provides normative information on grade and secondary school students, and only for the Reading Comprehension Test Parts A and B combined, local norms need to be established for participants hired from temporary employment agencies. Anastasi (1976, p. 92) points out:

> An approach to the nonequivalence of existing norms . . . is to standardize tests on more narrowly defined populations, so chosen as to suit the specific purposes of each test. In such cases, the limits of the normative population should be clearly reported with the norms. Thus, the norm might be said to apply to "employed clerical workers in large business organizations" or to "first-year engineering students." For many testing purposes, highly specific norms are desirable.

THE PRINTER USABILITY STUDIES

The data presented in this report were collected during a competitive evaluation of seven table-top, dot-matrix printers. The printers were evaluated consecutively, as they became available to the author, over about an 18-month period. Thus, it was believed to be possible that the average reading comprehension skills might differ over time. It was also believed that reading comprehension skill might influence the value of the usability measures being collected.

Participants

Seven printers were studied, with 10 different participants per study. The 70 participants were hired from local temporary employment agencies (Southeast Florida). The sample was composed of 16 (23 percent) males and 54 (77 percent) females. Twenty-four (34 percent) of the participants had graduated from high school only, and 46 (66 percent) had graduated from high school and had attended some college. None of the participants had a college degree. Participants were asked to give their age within a 10-year range. The median age range was 30-39 years old, with the central 50 percent of the participants ranging from 20 to 49 years old.

Materials and Apparatus

Each study required a printer, two sets of the user documentation (one for the participant and one for the monitor), two sets of task descriptions (one for the participant and one for the monitor), 10 sets of observation forms, 10 sets of user rating forms, 10 background questionnaires, 10 sets of the Iowa Silent Reading Test Level 2 Form E, 10 picture release forms, and video recording equipment.

Level 2 of the ISRT was chosen for the study because the participants were not expected to have been in the public school system recently, and were not expected to have completed a college degree. Level 2 was developed to assess reading skills for grades 9 to 14. The reading comprehension skills assessed by the ISRT Test 2 are, for Part A, "the student's ability to answer questions based on short passages to which he has ready access" (ISRT Manual of Directions, 1973, p. 10) and, for Part B, the "ability to answer questions based on a longer, essay-type passage" (ISRT Manual of Directions, 1973, p. 10). It seems clear that of these two subtests, Part A is the one most appropriate for testing the reading comprehension skills that a participant is expected to use in this type of usability study. While there might have been some value in administering the entire ISRT to participants, for practical reasons it was necessary to complete the usability study with one day per participant. Investing 26 minutes in an assessment of reading comprehension was reasonable, but we were not able to invest 2 hours and 16 minutes.

Procedure

Participants were observed individually. At the start of a session, the participant completed a background questionnaire and a picture release form, and was then brought into the laboratory. The tasks were presented to each participant in the same order. Each task was presented three times with minor variations. The tasks were selected to be typical of printer operation, such as loading forms, setting the top of forms, clearing a forms jam, changing the ribbon, and running the printer's self-test.

To start a task, the monitor gave the participant a copy of the task scenario. The participant then performed the task up to the point of actually running a print job. The printers were not connected to a computer, so print jobs were not run. The monitor was trained in problem identification, and was able to judge whether a subsequent print job would succeed or fail. After performing the task, the participant was given a user rating form that asked the user to rate the printer on a 5-point scale regarding satisfaction with task time,

perceived ease-of-use, and satisfaction with the user documentation. For this scale, "1" was the most favorable rating and "5" was the least favorable.

Participants were asked to use the instructions and operator manuals to complete the tasks. If they encountered a problem that they could not solve, they asked the monitor for assistance.

For assists during the first trial, the monitor gave the participants no more information than absolutely necessary for the participant to overcome the immediate problem. This was done to ensure proper assessment of the adequacy of the documentation. During subsequent trials, the monitor provided more comprehensive instruction to the participant if assistance was requested or if problems were observed at the end of the task. In this way, the participants learned more rapidly, and their performance on the third trial reflected as much learning as possible.

The monitor used unobtrusive observational techniques, including the use of one-way glass and video recording equipment. The monitor recorded observations on forms designed for that purpose.

After the tasks were completed, the Iowa Silent Reading Test Level 2 Form E Test 2 Part A was used to assess reading comprehension skill.

Dependent Measures

The following measures were recorded for each task and each trial.

Task time. Task time was defined as the duration of time from the moment that the participant first made a machine manipulation until the participant indicated that the task was completed.

Number of assistance calls. An assistance call occurred whenever the participant was unable to perform a portion of the task and asked for assistance.

Number of problems. The number and type of problems were recorded. Two types of problems were defined. If a participant indicated that a task was complete, but the printer was not properly prepared and the print job would have failed (as judged by the monitor), then these problems were categorized as Severity 1. If a participant experienced difficulty with a portion of the task, but managed to solve the problem before indicating that the task was complete and without asking for assistance, then the problem was categorized as Severity 2.

Successful task completion. A task was considered to be successfully completed if the participant made no calls for assistance and experienced no Severity 1 problems.

User ratings. User ratings of Time to Complete the Task, Ease of Completing the Task, and Satisfaction with the User Documentation were collected by giving the participants a rating form after each task. Ratings were based on a 5-point scale, with "1" the most favorable rating and "5" the least favorable.

The variables primarily used to assess the different printers were the successful task completions and the user ratings. The descriptions of assistance and problems were used to determine how to improve a printer.

RESULTS

Distribution of the Reading Comprehension Scores

The distribution of the reading comprehension scores is shown in Table 1. The highest possible score is 38. As shown in the table, the mean was 25 with a standard deviation of 8.4. The median was 26 with an interquartile range of 13. The central 50 percent of the scores fell within the range of 19 to 32, and the central 90 percent of the scores fell within the range of 8 to 36. A graph of this distribution is shown in Figure 1 (at the end of this paper). The normal distribution with mean 25 and standard deviation 8.4 is laid over the obtained distribution. The obtained distribution is clearly non-normal, skewed to the right.

Table 1. Distribution of Reading Comprehension Scores

Score	Cumulative Frequency	Cumulative Frequency	Percent	Percent
6	1	1	1.43	1.43
7	1	2	1.43	2.86
8	2	4	2.86	5.71
10	1	5	1.43	7.14
13	2	7	2.86	10.00
14	3	10	4.29	14.29
16	2	12	2.86	17.14
17	2	14	2.86	20.00
18	2	16	2.86	22.86
19	3	19	4.29	27.14
20	5	24	7.14	34.29
22	2	26	2.86	37.14
23	2	28	2.86	40.00
24	1	29	1.43	41.43
25	4	33	5.71	47.14
26	3	36	4.29	51.43
27	3	39	4.29	55.71
28	7	46	10.00	65.71
29	2	48	2.86	68.57
30	2	50	2.86	71.43
32	3	53	4.29	75.71
33	2	55	2.86	78.57
34	4	59	5.71	84.29
35	3	62	4.29	88.57
36	5	67	7.14	95.71
37	3	70	4.29	100.00

Mean = 24.957
Median = 26.000
Standard deviation = 8.412
Variance = 70.766

Correlation between Successful Task Completion Rate and the Reading Comprehension Score

Over all printers and tasks, the product-moment correlation between the successful task completion rate and the reading comprehension score was significant ($r = .25$, $n = 70$, $p = .035$). This result shows a modest relationship between the variables, indicating that participants with higher reading comprehension scores tended to complete more tasks successfully.

Correlation between Reading Comprehension Score and the Average Rating of Satisfaction with the User Documentation

Over all printers and tasks, the product-moment correlation between the participants' average ratings of satisfaction with the documentation and their reading comprehension scores was significant ($r = .35$, $n = 70$, $p = .003$). Because a greater reading comprehension score is associated with superior reading comprehension skill, and a greater average rating of satisfaction is associated with less satisfaction, this positive correlation shows a relationship between the variables, indicating that participants with higher reading comprehension scores tended to be more critical of (or less satisfied with) the documentation.

Analysis of Variance for Successful Task Completion Rates

Successful task completion rates were calculated for each participant. A two-way (Printer, Gender) analysis of variance (ANOVA) was conducted. The main effect of Gender and the Printer by Gender interaction were not significant, but the main effect of Printer was highly significant ($F(6,56) = 10.585$, $p < .0001$).

Analysis of Variance for Ratings of Satisfaction with User Documentation

The mean rating for satisfaction with the user documentation was calculated for each participant. A two-way (Printer, Gender) analysis of variance was conducted. The results were similar to those for the successful task completion rates. The main effect of Gender and the Printer by Gender interaction were not significant, but the main effect of Printer was significant ($F(6,56) = 2.614$, $p = .026$).

Analysis of Variance for the Reading Comprehension Scores

A two-way (Printer, Gender) analysis of variance was conducted on the participants' reading comprehension scores. There were no significant main effects or interactions. The main effect of Printer was clearly not significant ($F(6,56) = .703$, $p = .65$).

(For more details, see Lewis, 1990.)

DISCUSSION

These results show that the participants' performance and satisfaction differed significantly as a function of the printer they used during the usability studies. The participant groups were not significantly different in reading comprehension skill as measured by the ISRT Level 2 Form E Test 2 Part A. Therefore, it is reasonable to believe that the usability assessment of the printers was uninfluenced by differences in reading comprehension skill among the participant groups. The importance of measuring the reading comprehension skill is shown by the significant correlations between participants' reading comprehension scores and the dependent variables of correct task completion rate and satisfaction rating of the documentation. Although the correlations were not of tremendous magnitude, they were large enough to be considered useful for predictive validity. Nunnally (1978) states that predictive validity coefficients as low as .30 indicate that a test can be successfully used in some situations. Another point to keep in mind is that the influence of reading comprehension skill on the successful task completion rate may have been overshadowed by the extreme usability differences (as measured by the successful task completion rate) among the printers. In other words, the design of the printers and the quality of the accompanying documentation may have had more to do with successful performance than the individual difference of reading comprehension skill. Also, participants may not have relied upon the documentation (at least, not to a large extent) when performing the tasks. Finally, the ability to generalize these findings is enhanced since the data were collected over seven different printers with substantially different designs.

It is fortunate that the participant groups did not differ in reading comprehension skill. However, if they had differed, the reading comprehension score could have been used as a covariate to remove the effect of this individual difference. When this study was conducted, there was a distinct possibility that the groups would differ in this attribute. Temporary employment agencies would be expected to try to fulfill a request for study participants with the most qualified persons first. When a study is conducted over an extended period of time, there may be definite differences between the initial and final participants. Collecting information on individual differences that might be expected to influence participant performance or satisfaction during the study is analogous to paying for an insurance policy. An investment is required (because the time must be spent to collect the information). You might never collect on the policy (if the participant groups are essentially equal). However, it is important to have if you need it (if the participant groups can be shown to be unequal in the relevant attribute).

Finally, the reading comprehension scores reported here may be useful as local norms for this portion of the ISRT. Tests of individual differences would not usually be used to exclude potential participants from a usability test, although it may occasionally be necessary to do so. Also, when conducting specific problem analyses from a usability test, it may be very useful to know the relative reading comprehension skill of the person experiencing the

problem. This knowledge may influence the determination of the type of problem being experienced and reasonable design changes to prevent the problem from occurring again. The sample size of this study is too small to allow strong normative statements to be made. However, the information provided may be useful for approximating the percentile standing of adults hired from temporary employment agencies for the purpose of participating in usability studies.

REFERENCES

Anastasi, A. (1976). *Psychological testing.* New York: Macmillan.

Buros, O.K. (1978). *The eighth mental measurements yearbook.* Highland Park, NJ: Gryphon Press.

Filby, N.N. (1978). Review of the Iowa Silent Reading Test. In O.K. Buros (Ed.), *The eighth mental measurements yearbook* (pp. 1196-1197). Highland Park, NJ: Gryphon Press.

Hakstian, A.R. (1978). Review of the Iowa Silent Reading Test. In O.K. Buros (Ed.), *The eighth mental measurements yearbook* (pp. 1197-1199). Highland Park, NJ: Gryphon Press.

Hunter, R. and Hoepfner, R. (1978). Review of the Iowa Silent Reading Test. In O.K. Buros (Ed.), *The eighth mental measurements yearbook* (pp. 1200-1201). Highland Park, NJ: Gryphon Press.

Iowa silent reading tests manual of directions. (1973). Harcourt Brace Jovanovich.

Lewis, J. R. (1990). *Using the reading comprehension section of the Iowa Silent Reading Test as a Usability Study Pretest: Local norms and predictive validity* (IBM Tech. Report 54.537). Boca Raton, FL: IBM Corp.

Myers, J. (1979). *Fundamentals of experimental design.* Boston: Allyn and Bacon.

Nunnally, J.C. (1978). *Psychometric theory.* New York: McGraw-Hill.

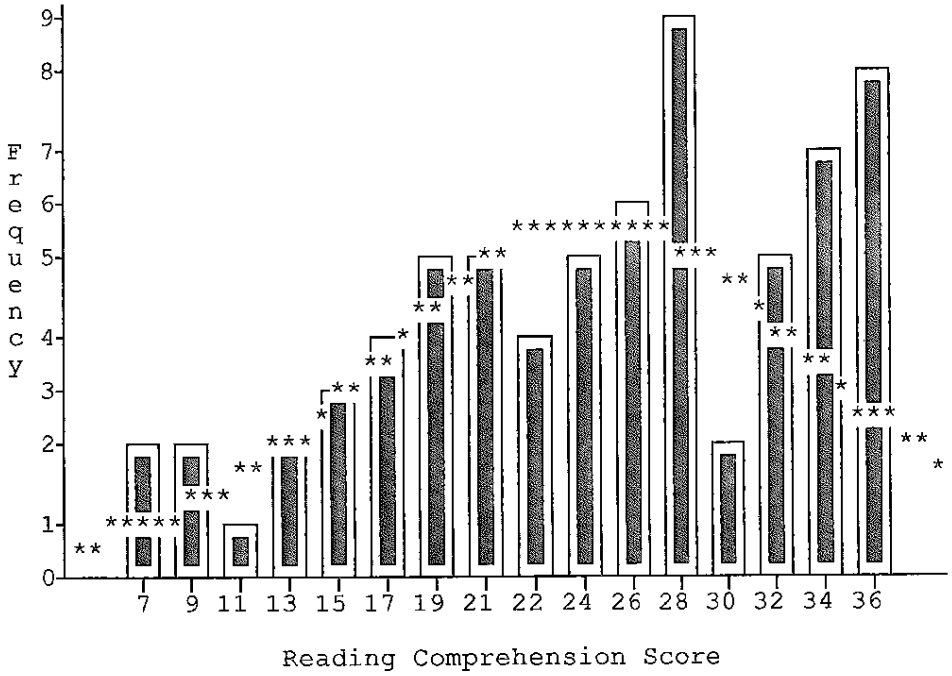

Figure 1. Obtained and Hypothetical Normal Reading Comprehension Score Distributions