Estimating the reliability of self-reported data for Web-based instruction

Chong Ho Yu, Ph.D., Barbara Ohlund, Samuel DiGangi, Ph.D.,
& Angel Jannasch-Pennell, Ph.D.

Arizona State University

Instruction and Research Support
870101 Information Technology
Arizona State University
Tempe AZ 85287-0101

Estimating the reliability of self-reported data for Web-based instruction

Chong Ho Yu, Ph.D., Barbara Ohlund, Samuel DiGangi, Ph.D., & Angel Jannasch-Pennell, Ph.D.

Abstract: Many research studies and evaluations of Web-based instruction are based upon self-reported data, which may be highly inaccurate due to faulty memory and other psychological factors. This paper discusses four ways to estimate the reliability of self-reported data. The four approaches are: Kappa coefficient, Index of Inconsistency, Repeated measures, and correlational/regression analysis.

For research on Web-based instruction, web usage data may be obtained by parsing the user access log, setting cookies, or uploading the cache. However, these options may have limited applicability. For example, the user access log cannot track users who follow links to other websites. Further, cookie or cache approaches may raise privacy issues. In these situations, self-reported data collected by surveys are used. This gives rise to the question: How accurate are self-reported data? Cook and Campbell (1979) have pointed out that subjects (a) tend to report what they believe the researcher expects to see, or (b) report what reflects positively on their own abilities, knowledge, beliefs, or opinions. Another concern about such data centers on whether subjects are able to accurately recall past behaviors. Cognitive psychologists have warned that the human memory is fallible (Schacter, 1999) and thus the reliability of self-reported data is tenuous.

Although statistical software packages are capable of calculating numbers up to 16-32 decimals, this precision is meaningless if the data cannot be accurate at even the integer level. Quite a few scholars had warned researchers how measurement error could cripple statistical analysis (Blalock, 1974) and suggested that good research practice requires the examination of the quality of the data collected (Fetter, Stowe, & Owings, 1984).

Bias and Variance

Measurement errors include two components, namely, bias and variable error. Bias is a systematic error that tends to push the reported scores toward one extreme end. For example, several versions of IQ tests are found to be bias against non-Whites. It means that blacks ands Hispanics tend to receive lower scores regardless of their actual intelligence. A variable error, also known as variance, tends to be random. In other words, the reported scores could be either above or below the actual scores (Salvucci, Walter, Conley, Fink, & Saba, 1997).

The findings of these two types of measurement errors have different implications. For example, in a study comparing self-reported data of height and weight with direct measured data (Hart & Tomazic, 1999), it was found that subjects tend to over-report their height but under-report their weight. Obviously, this kind of error pattern is bias rather than variance. A possible explanation of this bias is that most people want to present a better physical image to others. However, if the measurement error is random, the explanation may be more complicated.

One may argue that variable errors, which are random in nature, would cancel out each other and thus may not be a threat to the study. For example, the first user may over-estimate his Internet activities by 10%, but the second user may under-estimate hers by 10%. In this case, the mean might still be correct. However, over-estimation and under-estimation increases variability of the distribution. In many parametric tests, the within-group variability is used as the error term. An inflated variability would definitely affect the significance of the test. Some texts may reinforce the above misconception. For example, Deese (1972) said,

> Statistical heory tells us that the reliability of observations is proportional to the square root of their number. The more observations there are, the more random influences there will be. And statistical theory holds that the more random errors there are, the more they are likely to cancel one another and produce a normal distribution (p.55).

First, it is true that as the sample size increases the variance of the distribution decreases, it does not guarantee that the shape of distribution would approach normality. Second, reliability (the quality of data) should be tied to measurement rather than sample size determination. A large sample size with a lot of measurement errors, even random errors, would inflate the error term for parametric tests.

After calculating the standardized difference between the two measurements, a stem-and-leaf plot or a histogram can be used to visually examine whether a measurement error is due to systematic bias or random variance.

Remedies

In spite of the threat of data inaccuracy, it is impossible for the researcher to follow every subject with a camcorder and record everything they do. Nonetheless, the researcher can use a subset of subjects to obtain observed data such as user log access or daily hardcopy log of web access. The results would then be compared to the outcome of all subjects' self-reported data for an estimation of measurement error. For example, when the user access log is available to the researcher, he can ask the subjects to report the frequency of their access to the web server. The subjects should not be informed that their Internet activities have been logged by the Webmaster as this may affect participant behavior. Also, the researcher can ask a subset of users to keep a logbook of their Internet activities for a month.

Someone may argue that the log book approach is too demanding. Indeed, in many scientific research studies, subjects are asked for much more than that. For instance, when scientists studied how deep sleep during long range space travel would affect human health, participants were asked to lie in bed for a month. In a study concerning how a closed environment affects human psychology during space travel, subjects were locked in a room individually for a month. It takes a high cost to seek out scientific truths.

After different sources of data are collected, the discrepancy between the log and the self-reported data can be analyzed to estimate the data reliability. At first glance, this approach looks like a test-retest reliability, but it isn't. First, in test-retest reliability the instrument used in two or more situations should be the same. Second, when the test-retest reliability is low, the source of errors is within the instrument. However, when the source of errors is external to the instrument such as human errors, inter-rater reliability is more appropriate.

The above suggested procedure can be conceptualized as a measurement of inter-data reliability, which resembles that of inter-rater reliability and repeated measures. There are four ways to estimate the inter-rater reliability, namely, Kappa coefficient, Index of Inconsistency, repeated measures ANOVA, and regression analysis. The following section describes how these inter-rater reliability measurements may be used as inter-data reliability measurements.

Kappa coefficient

In psychological and educational research, it is not unusual to employ two or more raters in the measurement process when the assessment involves subjective judgments (e.g. grading essays). The inter-rater reliability, which is measured by Kappa coefficient, is used to indicate the reliability of the data. For example, the performance of the participants is graded by two or more raters as "master" or "non-master" (1 or 0). Thus, this measurement is usually computed in categorical data analysis procedures such as PROC FREQ in SAS and "measurement of agreement" in SPSS's StatXact.

It is important to note that even if 60 percent of two datasets concur with each other, it doesn't mean that the measurements are reliable. Since the outcome is dichotomous, there is a 50 percent chance that the two measurements agree. Kappa coefficient takes this into account and demands a higher degree of matching to reach consistency.

In the context of Web-based instruction, each category of self-reported Website usage can be re-coded as a binary variable. For example, when question one is "how often do you use telnet," the possible categorical responses are "a: daily," "b: three to five times per week," "c: three-five times per month," "d: rarely," and "e: never." In this case, the five categories can be re-coded into five variables: Q1A, Q1B, Q1C, Q1D, and Q1E. Then all these binary variables can be appended to form a R X 2 table as shown in the following table. With this data structure, responses can be coded as "1" or "0" and thus measurement of classification agreement is possible. The agreement can be computed using Kappa coefficient and thereby the reliability of the data may be estimated.

Index of Inconsistency

Another way to compute the aforementioned categorical data is Index of Inconsistency (IOI). In the above example, because there are two measurements (log and self-reported data) and five options in the answer, a 4 X 4 table is formed. The first step to compute IOI is to divide the RXC table into several 2X2 sub-tables. For example, the last option "never" is treated as one category and all the rest are collapsed into another category as "not never," as shown in the following table.

Table 1.
2X2 table for IOI.

	Never	Not never	Total
Never	a	b	a+b
Not Never	c	d	c+d
Total	a+c	b+d	n=sum(a-d)

The percent of IOI is computed by the following formula:

IOI% = 100*(b+c)/[(2np(1-p)] where p = (a+c)/n

After the IOI is calculated for each 2X2 sub-table, an average of all indices is used as an indicator of the inconsistency of the measure. The criterion to judge whether the data are consistent is as follows:

- An IOI of less than 20 is low variance;
- An IOI between 20 and 50 is moderate variance;
- An IOI above 50 is high variance

The reliability of the data is expressed in this equation: r = 1 – IOI. Put it simply, reliability is the information without the inconsistent portion.

Repeated measures

The measurement of inter-data reliability can be conceptualized and proceduralized as a repeated measures ANOVA. In a repeated measures ANOVA, measurements are given to the same subjects several times such as pretest, midterm and posttest. In this context, the subjects are also measured repeatedly by the web user log, the log book and the self-reported survey. The following is the SAS code for a repeated measures ANOVA:

```
data one; input user $ web_log log_book self_report;
        cards;
        1   215   260   200
        2   178   200   150
        3   100   111   120
        4   135   172   100
        5   139   150   140
        6   198   200   230
        7   135   150   180
        8   120   110   100
        9   289   276   300
proc glm;
        classes user;
model web_log log_book self_report = user;
        repeated time 3;
    run;
```

In the above program, the number of visited Websites by nine volunteers is recorded in the user access log, the personal log book, and the self-reported survey. The users are treated as a between-subject factor while the three measures are regarded as between-measure factor. Table 2 is a condensed output:

Table 2
Output of repeated measures for inter-data reliability.

Source of variation	df	Mean square
Between-subject (user)	8	10442.50
Between-measure (time)	2	488.93
Residual	16	454.80

Based on the above information, the reliability coefficient can be calculated using the following formula (Fisher, 1946; Horst, 1949):

$$r = (MS_{\text{between-measure}} - MS_{\text{residual}})/(MS_{\text{between-measure}} + (df_{\text{between-people}} \text{ X } MS_{\text{residual}}))$$

Correlational and regression analysis

Correlational analysis, which utilizes Pearson's Product Moment coefficient, is very simple and especially useful when the scales of two measurements are not the same. For example, the web server log may track the number of pages accesses while the self-reported data are likert-scaled (e.g. How often do you browse the Internet? 5=very often, 4=often, 3=sometimes, 2=seldom, 5=never). In this case, the self-reported scores can be used as a predictor to regress against page access.

A similar approach is regression analysis, in which one set of scores (e.g. survey data) is treated as the predictor while another set of scores (e.g. user daily log) is considered the dependent variable. If more than two measures are employed, a multiple regression model can be applied i.e. the one that yields more accurate result (e.g. Web user access log) is regarded as the dependent variable and all other measures (e.g. user daily log, survey data) are treated as independent variables.

Method

A preliminary pilot was conducted to test the discussed procedures, and study the accuracy of self-report data.

Participants

Participants in this preliminary pilot were students enrolled in a required methods course as well as staff in information technologies at a southwestern university (n=18). All participants completed a brief survey; twelve completed log-recordings.

Procedures

Participants completed a brief, three-question survey regarding web-based use, shown in Table 3.

Table 3.
Web use survey

1) In a week, estimate how many times you use email: _____
2) In a week, estimate how many times you use the World Wide Web: ____
3) In a week, estimate how many times you use File Transfer Protocol: ____

Frequency was described to participants as how many times email was accessed during a day over a week period, not as initial access (i.e., opening their email for the first time that day) or how many emails they read or sent. For example, if a user had multiple applications open over a period of time during the day, email frequency was how many times the user went back to view their email, regardless of read or sent email. Accessing the World Wide Web was explained in the same manner. As File Transfer Protocol is fairly discrete, frequency was described as how many times the application was accessed during a week.

Twelve participants completed the log-recording of web-use specific to email, World Wide Web and File Transfer Protocol. Participants were asked to keep a frequency log in the morning, afternoon and evening, over a 5-7 day period. All participants in this phase completed six-seven days of log-data. For data analysis, we used six days of log data from each participant.

Results

Correlational analysis indicated that the measures of reporting usage of FTP and email are consistent, but Web browsing is not (see Table 4)

Table 4
Correlation Coefficient between Daily Log and Survey Data

Usage	*r*	*p*
Email	.79	.0039
Web	.58	.0625
FTP	.91	.0001

Difference scores between two measures for usage of email, FTP, and Web browsing were standardized and plotted in histograms for examining bias and variance. Since bandwidth (number of bins) affects the appearance of the distribution, different number of bins and ticks were tried to gain a thorough view of the data. It was found that the measurement errors of usage of email, FTP, Web browsing were variable errors rather than bias. Figures 1-3 indicate that although most difference scores were centered around zero, the spread went as far as three standard deviations and thus the reliability of self-reported data was still questionable.

Figure 1. Standardized Difference Scores of Usage of Email

Figure 2. Standardized Difference Scores of Usage of FTP.

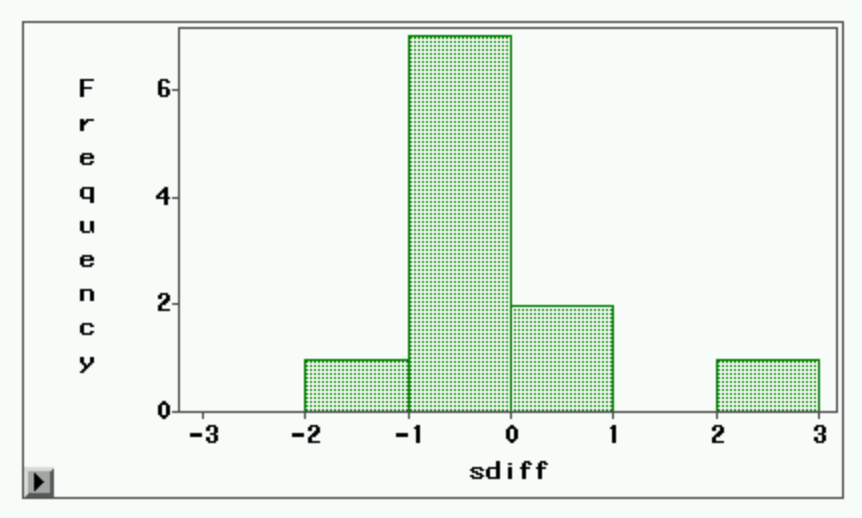
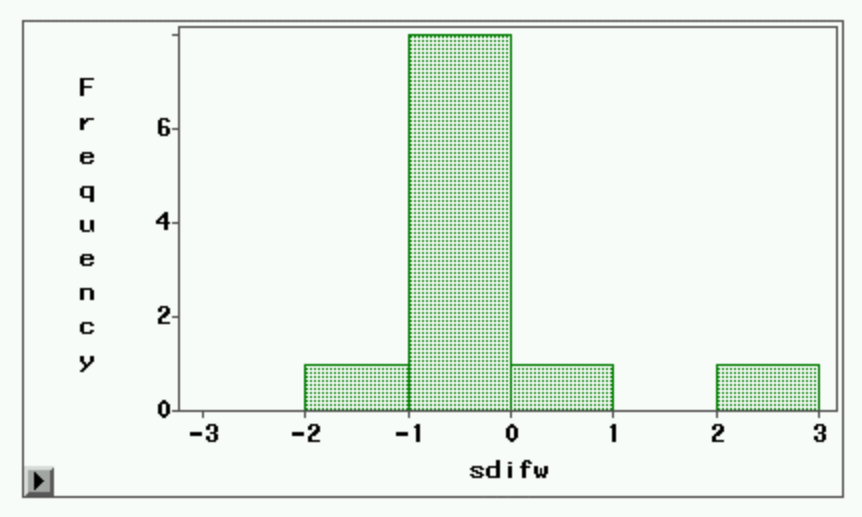


Figure 3. Standardized Difference Scores of Usage of Web browser.



Discussion

Pertaining to reliability estimation, tremendous attention has been paid to internal consistency, which is measured by Cronbach Alpha Coefficient. Internal consistency is just one of several aspects of reliability. Other aspects such as stability over time and equivalence between different measures are considered more important when the accuracy of data is suspect. The approach introduced in this paper addresses the issue of stability and equivalence. Since users are measured in different times, stability is taken into account of reliability estimation. Because different forms of measurement are employed, equivalence is also included as a reliability component.

Since a demanding commitment (writing a log everyday) is required in this study, the number of participants is small and the measurement is as simple as possible (three questions only). -External motivation (e.g. extra bonus points) will be provided in subsequent studies so that more subjects will be obtained and more questions will be asked.

References

Blalock, H. M. (1974). (Ed.) Measurement in the social sciences: Theories and strategies. Chicago, Illinois: Aidine Publishing Company.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues. Boston, MA: Houghton Mifflin Company.

Deese, J. (1972). Psychology as science and art. New York, NY: Harcourt Brace Jovanovich, Inc.

Fetters, W., Stowe, P., & Owings, J. (1984). High School and Beyond. A national logitudinal study for the 1980s, quality of responses of high school students to questionnaire items. (NCES 84-216). Washington, D. C.: U.S. Department of Education. Office of Educational Research and Improvement. National center for Education Statistics.

Fisher, R. J. (1946). Statistical methods for research workers (10th ed.). Edinburgh: Oliver and Boyd.

Hart, W.; & Tomazic, T. (1999 August). Comparison of percentile distributions for anthropometric measures between three data sets. Paper presented at the Annual Joint Statistical Meeting, Baltimore, MD.

Horst, P. (1949). A Generalized expression for the reliability of measures. Psychometrika, 14, 21-31.

Schacter, D. L. (1999). The seven sins of memory: Insights from psychology and cognitive neuroscience. American Psychology, 54, 182-203.

Salvucci, S.; Walter, E., Conley, V; Fink, S; & Saba, M. (1997). Measurement error studies at the National Center for Education Statistics. Washington D. C.: U. S. Department of Education.