The Effects of a Computer-Assisted Interview Tool on Data Quality

Justus J. Randolph, Marjo Virnes, Ilkka Jormanainen and Pasi J. Eronen
Department of Computer Science, University of Joensuu, PO BOX 111, 80101, Joensuu, Finland
Tel: +358 13 251 7929
Fax: +358 13 251 7955
justus.randolph@cs.joensuu.fi

ABSTRACT

Although computer-assisted interview tools have much potential, little empirical evidence on the quality and quantity of data generated by these tools has been collected. In this study we compared the effects of using Virre, a computer-assisted self-interview tool, with the effects of using other data collection methods, such as written responding and face-to-face interviewing, on data quantity and quality. An intra-sample, counter-balanced, repeated-measures experiment was used with 20 students enrolled in a computer science education program. It was found that there were significantly more words and unique clauses per response in the Virre condition than in the written response condition; however, the odds of avoiding in-depth responding were much greater in the Virre condition than in the written response condition. The effect sizes observed in this study indicated that there was a greater quantity of data and a higher rate of response avoidance during Virre data collection than what would have been expected, based on previous literature, during face-to-face data collection. Although these results statistically generalize to our sample only, these results indicate that audio-visual, computer-assisted self-interviewing can yield as high, or higher, quality of data as other commonly-used data collection methods.

Keywords

Computer-assisted interviewing, Data collection, Data quality

Introduction

Face-to-face interviewing is considered by many to be superior to other types of interview data-collection techniques, such as telephone interviewing and paper-and-pencil interviewing (Dillman, 1978; Rossi, Wright, & Anderson, 1983; Smith, 1987). According to De Leeuw, compared to paper-and-pencil and telephone interviews, face-to-face interviews have more flexibility, higher response rates, and greater potential for studying the general population (1993). However, face-to-face interviews also require more resources to carry out (De Leeuw, 1993).

Much research has been collected over the past two decades on the feasibility of using computer-assisted data collection methods as low-cost supplements or alternatives to face-to-face data collection (see Saris, 1989; 1991). Most of the computer-assisted data collection research concentrates on interviewers' and respondents' attitudes towards computer-assisted interviewing; however, it does not address the quality of data that is collected in computer-assisted interviews. De Leeuw and Nichols (1996), in a review of research on computer-assisted data collection methods, concluded that:

> Computer-assisted data collection has a high potential regarding increased timeliness of results, improved data quality, and cost reduction in large surveys. However, for most of these potential advantages the empirical evidence is still limited. . . . At present there is still little empirical research into the effect of computerized interviewing on data quality. Most studies have investigated the acceptance of the new techniques by interviewers and respondents. (p. 7.1-7.2)

Ten years later, there is still a high potential for computer-assisted personal interviewing, but there is still also a paucity of empirical evidence on the quality of data generated by those methods. Because of a lack of needed empirical evidence in this area, we examined the quantity and quality of data generated from computer-assisted interviewing. Specifically, we investigated the use of an innovative tool called Virre, which is a combination of a computerized self-administered questionnaire and a computerized self-interviewing tool designed to collect the responses of primary-school-aged interviewees. (We classify Virre as a *self*-interviewing tool because only the respondent and Virre are necessary for an interview to be conducted.)

Several stakeholder groups stand to gain from this research. There has been much recent discussion in the field of research and evaluation about harnessing technologies to improve practice (Gay & Bennington, 2000; Love, 2004; Means, Roschelle, Penuel, Sabelli, & Haertel, 2003). Evaluators and researchers may be interested in this article's findings because we investigate the hypothesis that Virre, and tools like Virre, can be used as a low-cost

alternatives to high-quality, but resource-intensive, data collection strategies, such as face-to-face interviewing. Educational practitioners in general may be interested in the findings of this research; Virre was originally designed as an educational tool to motivate students to 'think about their thinking' aloud, a strategy which has known benefits for learning (see Pressley & McCormick, 1995). Computer science educators may be particularly interested in this research because Virre was created by, and for, computer science educators in K-12 computer science education programs. Although our study design only allows us to statistically generalize to the population of responses made by the students involved in this study, it is reasonable to hypothesize that our findings would generalize most accurately to populations of similar students in comparable K-12 computer science education programs.

In the section that follows, we describe a conceptual model of data collection effects and use that model to point out the theoretical similarities between Virre and face-to-face data collection. In the section titled 'Expected Results,' we describe and justify the point estimate and range of values that we expected in our study. In the procedures section, we operationalize our variables, describe Virre in more detail, and document the data collection and data analysis procedures that we used. In the results section, we present the findings of our research in terms of the comparative effects of Virre and written responding on number of words, number of unique clauses, and presence of response- avoidance phrases. In the discussion section, we revisit our research hypotheses and comment on the implications of our findings.

Literature Review

De Leeuw's (1993) conceptual model of data collection effects on data quality, as illustrated in Figure 1, is composed of three types of factors: media-related factors, information transmission factors, and interviewer impact factors. According to De Leeuw (1993), media-related factors concern "the social conventions associated with the medium of communication" (p.14), information transmission factors concern "the more technical aspects of information transmission, not social customs" (p. 16), and interviewer impact factors concern the extent to which the effects of the interviewer's presence and behaviors are restricted.

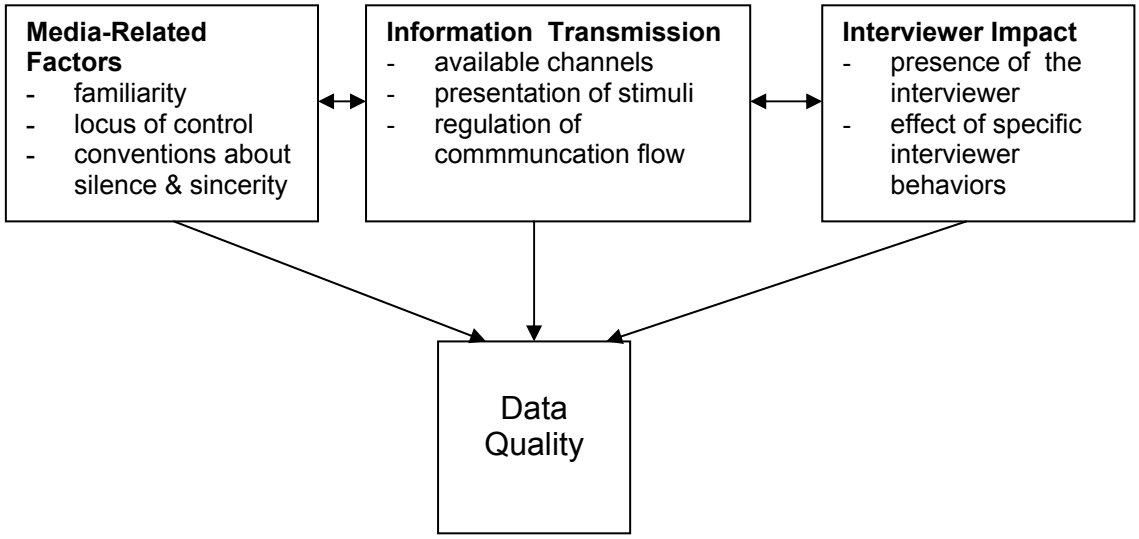

Figure 1. De Leeuw's Conceptual Model of Data Collection Effects on Data Quality
[From De Leeuw, 1993, p. 20, Reprinted with permission]

In terms of media related factors, face-to-face interviewing and Virre methods differ on several counts. Face-to-face interviewing is obviously more familiar to respondents than computer-assisted interviewing. We assume that respondents will have different conventions about silence and evaluate the sincerity of the data collection endeavor differently in face-to-face and Virre conditions. Nevertheless, several studies (e.g., Zandan & Frost, 1989; Witt & Bernstein, 1992) have found that "respondents generally like [computer-assisted interviewing]; they find it interesting, easy to use, and amusing" (De Leeuw & Nichols, 1996, p. 6.3). We suppose that the locus of control in face-to-face interviewing and Virre is comparable.

In terms of interviewer factors, on one hand, Virre is not able to do many of the tasks that a human interviewer can do well, such as motivating respondents, answering respondents' questions, or asking for clarification to respondents' answers. On the other hand, research has shown that participants are much more likely to disclose sensitive information in computer-assisted interviews than in face-to-face interviews -- an effect which has been found, however, to be diminishing over time (Wiesband & Kiesler, 1996). Also, because computer-assisted interviews can be standardized, instrumentation threats to internal validity can be controlled for.

Because of the theoretical similarities between Virre and face-to-face data collection, especially in terms of factors related to information transmission, we hypothesized that the effect sizes of Virre would be similar to the effect sizes found in the research on face-to-face data collection. (The effect sizes found in the research on face-to-face data collection can be found in the following section.)

Expected Results

Because of the lack of previous research on the effects of computer-assisted data collection on data quality, we had to estimate the expected results by proxy; we looked to the research on data comparisons between mail, telephone and face-to-face data-collection methods to put the effect sizes of the current study into context. In a meta-analysis of the effects of mail, telephone, and face-to-face data collection methods on data quality, the two effect sizes, and their ranges, presented below are particularly useful for putting the results of our current study into perspective (De Leeuw, 1993).

Although there is no estimate of the comparative effects of face-to-face surveys and mail surveys on the number of responses to open statements, an estimate does exist between face-to-face surveys and telephone surveys. De Leeuw (1993) found that face-to-face surveys are slightly superior to telephone surveys in terms of the average number of statements to open questions. In a meta-analysis of four studies that examined number of statements to open questions between face-to-face surveys and telephone surveys, De Leeuw reported that the average, weighted *r* (i.e., the Pearson product-moment correlation) across studies was .04, with 95% lower and upper bounds of .02 and .07. (An *r* of .04 in the context of De Leeuw's study is statistically equivalent to 52% of cases having had a greater number of open statements in face-to-face interviews than in telephone interviews. See Rosenthal, Rosnow, & Rubin, 2000, or Randolph & Edmonson, 2005.) Based on De Leeuw's theoretical model (1993), we assume that the effect size between face-to-face surveys and mail surveys would be higher than the effect size between face-to-face surveys and telephone surveys for the number of statements to open questions.

Another important finding from De Leeuw's meta-analysis is that there is a slightly lower rate of nonresponse in face-to-face surveys than in mail surveys. In De Leeuw's meta-analysis, the average, weighted value of r over 8 studies was .03, with 95% lower and upper confidence intervals of .01 and .05 (1993). (An r of .03 in the context of De Leeuw's study is statistically equivalent to 51.5% of cases having had a lower rate of nonresponse in face-to-face surveys than in mail surveys. See Rosenthal, Rosnow, & Rubin, 2000, or Randolph & Edmonson, 2005.)

Because of the similarities between Virre data collection and face-to-face data collection and because face-to-face data collection has been shown to lead to more statements and less nonresponse than written data collection techniques, we drew the following general and specific hypotheses for this study:

1. The Virre responses of students will have more statements and fewer incidences of nonresponse than the written responses of those same students.
2. The magnitude of effects attributed to Virre, in terms of number of statements and incidence of nonresponse, will be near the plausible range of magnitude of effects attributed to face-to-face interviewing for number of statements (i.e., $.02 < r < .07$) and nonresponse [hereafter *response avoidance*] (i.e., $.01 < r < .05$). (Stated in another way, the difference between the number of statements in the Virre condition and the number of words in the written response condition is expected to be approximately the same as the differences between face-to-face interviews and telephone interviews found in the previous literature. Also, the difference between the number of response avoidances in the Virre condition and the number of response avoidances in the written response condition is expected to be approximately the same as the differences in nonresponse between face-to-face interviews and mail surveys found in the previous literature.)

To be able to compare the magnitude of effects of Virre responding and written responding to the magnitude of effects of face-to-face responding and mail responding requires several, admittedly tenuous, assumptions. The first is that mail responding is to face-to-face responding as written responding is to Virre responding. We justify this assumption based on the theoretical similarities between face-to-face and Virre responding. Second, in previous research, two often-used data quality variables were number of statements made and nonresponse.

Because number of statements made is not operationalized in the previous research, we assume that number of words per response and number of distinct ideas per response are valid measures of the construct – number of statements made. Since students were instructed to respond to every item in our study, we assume response avoidance terms such as I don't know or no comment to be analogous to nonresponse in the previous research. (Admittedly, number of statements made and nonresponse are quantitative in nature; however, following the terminology used in previous data collection research [e.g., in De Leeuw, 1993], we used the word quality to refer to these characteristics of collected data.)

Methods

Instruments

The Virre environment contains four parts: the application itself, a webcam, a microphone, and a computer that runs the application. The application, written with Delphi, handles question sets, which can predefined by a Virre operator. For each question in a set, the application plays an audio file in which the question is spoken; it also presents the question as text to the user. While the user responds to the question, the application records audio and video to a file so that the user can later watch and comment on the answer file. All data, such as question sets and comments, are saved to files in XML format.

Virre needs external software for recording and playing question files and for playing back answer files. Since Microsoft's DirectX interface is used to capture and compress Virre's video stream, it is possible to use any DirectX-compatible webcam.

Currently, Virre's components are housed inside a bear with a web cam in its nose and a video screen in its abdomen; however, the Virre components could be installed within different types of exteriors to match different contexts. See Eronen, Jormanainen, and Virnes (2003) for more details. Figure 2 shows a photograph of Virre in its stuffed-bear version, which is designed for elementary-age students.

Participants and Setting

This investigation took place in the context of a computer science education program, called Kids' Club, located at a university in eastern Finland (See Eronen, Sutinen, Vesisenaho, & Virnes, 2002a, 2002b). According to the Kids' Club website (Kids' Club, n.d.):

> Kids' Club is a collaborative research laboratory, where children of age 10-15 work together with university students and researchers of Computer Science and Education. To children, Kids' Club appears as a technology club with an opportunity to study skills of their interests in a playful, non-school-like environment, where there is room for innovative ideation and alternative approaches. (n. p.)

Typical activities in Kids' Club include designing and programming robots and carrying out technology design projects.

The 20 participants in this study self-selected into the Kids´ Club program, which is held at a learning laboratory of a department of computer science. Eight of the participants came from a university practice school; twelve of the students came from a local, public elementary school. The ages of the students ranged from 10 and 12 and their grade levels ranged from 4th grade to 6th grade. As is typical in computer science education programs, the number of female students who self-selected into the program was much lower than the number of males who self-selected into the program (Galpin, 2002; Teague, 1997); only 2 of 20 participants in this study were female participants.

Procedure

In this study the independent variable is response condition (i.e., Virre responding or written responding), the dependent variables are number of words per response, number of unique clauses per response, and the presence of one or more response avoidance phrases as the sole elements of a response. Oral language markers like mmm. . . , hmmm. . . , and oh. . . were removed from the responses before analyzing number of words.

Figure 2. A Student Interacting with an Ursidae Version of Virre

A clause was defined as a subject-predicate combination or implied subject-predicate combination. For example, if a student, in response to the question --"What did you do today?"-- had responded with "Built robots, I built robots today," both of those statements would have been counted as clauses. A unique clause was defined as clause that did not repeat the same information already reported in an earlier clause, in response to the same question. For example, if a student, in response to the question -- "What did you do today?" -- had responded with "Built robots, I built robots today," that response would have been considered to have only one unique clause since the information in the two clauses was essentially the same.

We dichotomously categorized a response as a response avoidance if the students used one or more of the following terms or phrases, translated from Finnish, without commenting further: "nothing," "nothing special," "everything," "can't remember," "nothing new," "all kinds of things," "anything," "something," or "it doesn't matter." For example, if a student, in response to the question -- "What did you do today?" -- answered "All kinds of things," that response would have been counted as a response avoidance; however, if the student had responded with "All kinds of things. We learned programming. We built robots," that response would *not* have been counted as a response avoidance since the student elaborated on what had been done.

Students' responses were collected on four occasions, twice in the Virre condition and twice in the written condition. Prior to this investigation, students had received instruction on how to design, build, and test robots. During all measurement sessions students did the same activities; they built, programmed, and tested their robots. During the last part of the class, students were asked to respond to the following questions, which are translated from Finnish:

1. What did you do in Lego-robot club today?
2. In your opinion, what was hard about today's lesson?
3. In your opinion, what was easy about today's lesson?
4. What did you learn today?

The researchers randomly selected which groups of students would respond using Virre and which would respond in writing. There were no time limits on how long students could take to respond in either condition.

If a student was instructed to use Virre, the student would individually go to Virre, hit the *start* button and would then see the question on the screen and hear the question being spoken by Virre. Then the student would record his or her answer, in private, by speaking to Virre. When the student was done speaking, he or she would tap the start button again to end the recording and move on to the next question until the student had answered all four

questions. If the student was instructed to respond in writing, he or she would be given a form which had the four questions on it. The forms were filled out individually.

The Virre and written response conditions were purposively counterbalanced (i.e., half of the students responded using Virre one measurement session and responded in writing in another measurement session, while the other half did the opposite). After the data had been collected, the responses in the Virre session were transcribed. One rater coded the number of words and number of unique clauses per response. To establish estimates of interrater reliability, a second rater coded a random sample of 30 (about 10%) of the responses.

Data Analysis

We took an intra-sample statistics approach (see Shaffer & Serlin, 2004) to data analysis because this was a preliminary study in which we were primarily interested in making inferences concerning the universe of responses that the participants in our sample might have made. Since we took an intra-sample approach, we treated participants as fixed-effects, used the response as the unit of analysis, and used a factorial, orthogonal, between-subjects univariate design for the dependent variables: number of words and unique clauses. (Had we taken the traditional, extra-sample statistics approach we would have treated the participant as the unit of analysis, used a repeated-measures, two-within factor design where the factors would have been type of treatment and measurement, and made an inference to the universe of participants.) A two-way mixed-effect model, single-measure intraclass correlation coefficient (consistency definition) was used as the estimate of inter-rater reliability. An analysis of how well the data for this study met parametric assumptions and a rationale for using multiple univariate analyses can be found in an online appendix; see Randolph (2006).

We used logistic regression to estimate an odds ratio for the binary, dependent variable: response avoidance. For model checking we used Hosmer-Lemeshow's test (Hosmer & Lemeshow, 1989) and examined classification tables. In these analyses, we used the variables response condition (i.e., Virre responding or written responding) and participants as covariates. To convert the odds ratio to the r metric, via Rosenthal et al.'s equations (2000, p. 24), we used the odds ratio to create a binary effect size display, then calculated r_{besd}. The confidence intervals for r_{besd} were estimated using the variance equations in Shadish and Haddock (1994).

Equation 3.2 of Rosenthal et al. (2000) was used to calculate $r_{contrast}$ for the effect size of words per response and unique clauses per response; confidence intervals for $r_{contrast}$ were calculated based on information given in Shadish & Haddock (1994, Table 16.3). SPSS 11.0.1 was used to conduct the statistical analyses.

Results

Complications

In the Virre condition, 23 student responses were missing because of a faulty design aspect in Virre. (There were 320 responses in the written response condition and 297 responses in the Virre condition.) We found that the students would sometimes repeatedly push the start button and, as a result, would skip one or more questions. We assume that these errors are nonsystematic because we are led to believe that the students pushed the start button repeatedly on accident instead of pushing the button repeatedly as a way to avoid responding. Because this fault has since been corrected, we do not conceptualize this as part of the treatment's construct.

We ran analyses with and without replacement of missing data. (When we replaced missing data, we replaced a missing data point with the data point that corresponded to the data point for the same condition, question, and participant, but in a different measurement.) Since there were only minor differences in the results when using replacement and not using replacement (see the sensitivity analyses section below) and since orthogonal designs are more robust than nonorthogonal designs, we did our parametric analyses on the data with missing values replaced. However, for nonparametric analyses (i.e., for the analyses of response avoidance and number of types of words) the data were not replaced. We also examined the mean differences when response avoidances were treated as missing data.

Interrater Reliability

The intraclass correlation coefficient between the raters was .79. The coefficient had a 95% upper confidence interval of .89 and lower confidence interval of .60.

Number of Words

Table 1 shows that students in the Virre condition responded with 3.39 more words per response than in the written response condition. The ANOVA table of Table 2 indicates that this difference (response condition) was statistically significant, F(1, 280) = 39.01, p < .000. For the difference between conditions, $r_{contrast}$ was .35; the 95% confidence intervals indicate that the plausible values of the parameter should be between .22 and .48.

Number of Unique Clauses

In this sample, there were, on average, 39 more unique clauses per 100 responses in the Virre condition than in the written response condition (see Table 1). The ANOVA table presented in Table 3 indicates that this difference (response condition) was statistically significant, F(1, 280) = 27.91, $p < .000$. For the contrast between response conditions, $r_{contrast}$ was .30 with 95% confidence intervals of .17 and .43.

Response Avoidance

Regardless of whether participants were included as fixed factors, the odds of a student avoiding a response were significantly greater in the Virre condition than in the written response condition. Table 4 shows the crosstabulation of response avoidance by response condition. Not using participants as a fixed factor, the odds of a student avoiding a response was 2.56 times greater in the Virre condition than in the written response condition.

Taking a logistic regression approach and including participants as a fixed-factor, the odds ratio was 4.26 with 99% lower and upper confidence intervals of 1.47 and 12.66. This odds ratio corresponds to an r_{besd} of .35, in the written response direction, and 99% confidence intervals of .22 and .48.

Table 1. Descriptive Statistics for Number of Words and Unique Clauses per Response Condition

Dependent variable	*Response Condition*	*Mean*	*SD*	*Std. Error*	*99% Confidence intervals*	
					Lower	*Upper*
Words[a]	Virre	7.53	7.62	0.05	6.53	8.52
	Written	4.14	4.24	0.05	3.14	5.13
Unique clauses[b]	Virre	1.58	0.95	0.39	1.45	1.70
	Written	1.22	0.47	0.39	1.10	1.34

a. Difference of the mean between response conditions (99% confidence intervals of difference): 3.39 (1.98, 4.79)

b. Difference of the mean between response conditions (99% confidence intervals of difference): 0.39 (0.18, 0.53)

Table 2. ANOVA Table for Words per Response Condition

Source	*Type III Sum of Squares*	*df*	*Mean square*	*F*	*p*
Corrected Model	6,413[a]	39	164.45	6.99	.000
Intercept	10,881	1	10,881.11	462.36	.000
Participant	3,520	19	185.24	7.87	.000
Response Condition	918[b]	1	918.01	39.01	.000
Interaction[c]	1,976	19	103.99	4.42	.000
Error	6,590	280	23.53		
Total	23,844	320			
Corrected total	13,003	319			

a. r^2= .49 (adjusted r^2 = .42), b. $r_{contrast}$ =.35 (99% confidence intervals of contrast = .22, .48). c. The interaction is between response condition and participant

Table 3. ANOVA Table for Unique Clauses per Response Condition

Source	*Type III Sum of Squares*	*df*	*Mean square*	*F*	*p*
Corrected model[a]	86.72	39	2.22	6.11	.000
Intercept	624.40	1	624.40	1,716.15	.000

Participant	45.16	19	2.38	6.50	.000
Response condition[b]	10.15	1	10.15	27.91	.000
Interaction[c]	31.41	19	1.65	4.54	.000
Error	101.88	280	0.36		
Total	813.00	320			
Corrected total	188.60	319			

a. r^2= .46 (adjusted r^2 = .38), b. $r_{contrast}$ = .30 (95% confidence intervals of contrast = .17, .43). c. The interaction is between response condition and participant

Table 4. Cross Tabulation of Response Avoidance and Response Condition

		Response condition		*Total*
		Virre	*Written*	
Response avoidance	Yes	27	14	41
	No	110	146	256
Total		137	160	297

Note. The odds ratio for this table is 2.56 with Mantel-Haenszel 95% lower and upper bounds of 1.28 and 5.11. Participants are not included as a fixed factor in this estimate. Missing data were not replaced in this analysis.

The logistic regression model's overall prediction accuracy was 88.9%. The model's prediction accuracy given that there was no response avoidance was 98.0%; however, prediction accuracy given response avoidance was only 31.7%. The Hosmer-Lemeshow test (with a statistically nonsignificant X^2 of 5.3 with 7 *df*) indicated that this model was appropriate.

Sensitivity Analysis

Table 5 shows the mean differences when no data manipulation was done (i.e., when no data were replaced), when missing data were replaced, and when response avoidances were also treated as missing data. Table 5 indicates that the differences between the data manipulation methods used are negligible. Regardless of the type of data manipulation, $r_{contrast}$ was .35 for number of words and .30 for number of clauses.

Table 5. Mean Differences under a Variety of Data Manipulations

Data Manipulation	*Variable*	*Mean difference*	*99% Confidence interval of difference*	
			Lower bound	*Upper Bound*
No replacement	Words	3.39	1.98	4.79
Replace missing	Words	3.39	2.32	4.46
Replace missing and response avoidance	Words	3.97	2.49	5.44
No replacement	Clauses	0.39	0.18	0.53
Replace missing	Clauses	0.36	0.22	0.49
Replace missing and response avoidance	Clauses	0.42	0.24	0.60

Note. Mean differences are positive in the Virre direction and negative in the written response direction.

Discussion

In summary, these 20 students used markedly more words and more unique clauses in the Virre condition than they used in the written response condition. There were significantly more instances of students avoiding responding in the Virre condition than in the written response condition. However, Table 5 shows that when avoided responses are treated as missing data, in both the Virre and written response conditions the number of words and clauses in each response in the Virre condition was still significantly higher than in the written condition.

In terms of our first research hypothesis, we were correct that there would be more words and unique clauses in the Virre condition than in the written response condition. However, contrary to our hypothesis, there were more incidences of response avoidance in the Virre condition than in the written response condition.

In terms of our second research hypothesis, we were wrong on both counts. Our observed results were far outside of the range of what we had expected. Figure 3 shows the predicted and observed effect sizes (in terms of *r*).

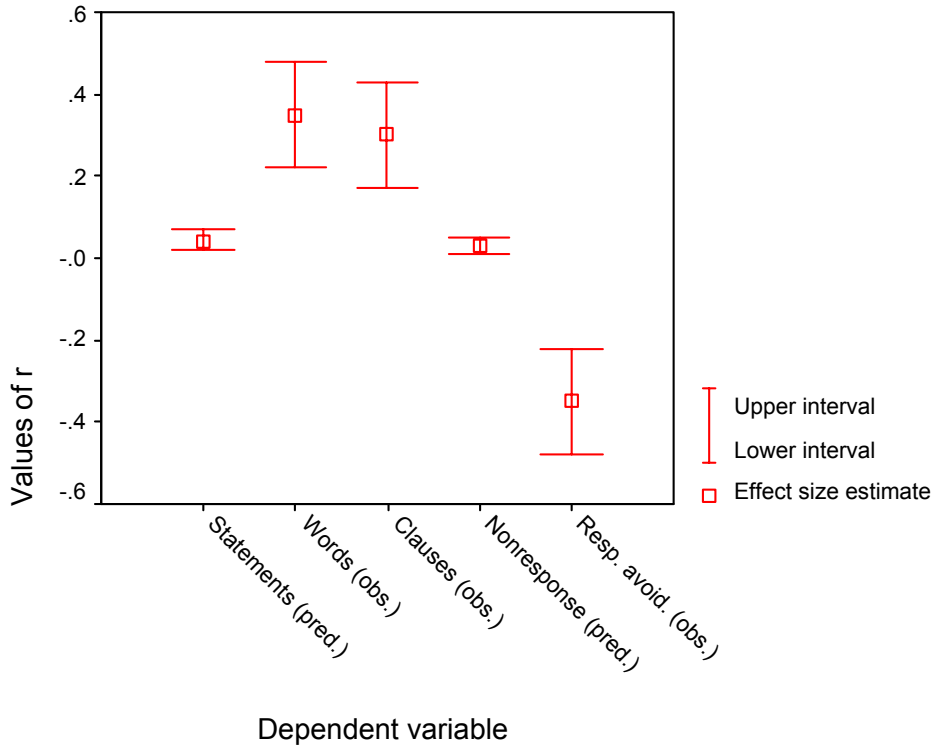

Figure 3. Differences Between Predicted and Observed Effect Sizes

Note. pred. = predicted. obs. = observed. Resp. avoid. = response avoidance. For *Words, Clauses* and *Resp. avoid*., values of *r* are positive in the Virre direction. For *Statements* and *Nonresponse*, values of *r* are positive in the face-to-face interview direction. 95% confidence intervals are shown for predicted variables. 99% confidence intervals are shown for observed variables.

We have several alternative hypotheses for the discrepancy between the expected and observed values:

- The variables that we chose do not correspond directly with the variables in the previous literature. For example, number of words and number of unique ideas may not measure the same construct, in the same way, as number of statements to open-ended questions.
- The analogy that mail responding (or telephone responding in the case of number of statements) is to face-to-face responding as written responding is to Virre responding is strained.
- The discrepancy is an artifact of our self-selecting sample compared to the supposedly random samples selected in the previous literature.
- The discrepancy is due to systematic error in either this study or the meta-analysis of the previous literature or a nonsystematic meta-analytic error as discussed in Briggs (2005).
- In terms of the reason for the higher number of response avoidances in the Virre condition, we suspect that there was a tacit academic convention that each question must not be avoided in the written response condition and that no such convention had been established in the Virre condition. Equally, the greater number of words and unique clauses in the Virre condition may have been due to an experimenter effect.

In actuality, we believe that the discrepancy is probably due to a combination of these hypotheses.

Since we took an intra-sample approach in this study, these results do not statistically generalize to the population of students. The results generalize to the population of responses that these students could have responded with. However, the results do at least provide some support for a revised hypothesis that students when using Virre will use more words and unique clauses, but also have a tendency to 'hedge' more on their answers, compared to when written responding is used. The prevalence of response avoidances in the Virre

condition, however, do not drastically change the fact that there were significantly more words and clauses in the Virre condition than in the written response condition.

Assuming that future generations of Virre studies affirm and generalize our finding that students respond with more words and unique clauses, controlling for response avoidances, there are several practical implications for researchers. If Virre can indeed be used as an alternative to face-to-face interviewing, researchers, evaluators, and educators in similar programs can benefit because Virre, and other computer-assisted self-interviewing tools are more cost-effective in the long run than face-to-face interviewing (De Leeuw & Nichols, 1996) and provide more data than written responses, as was shown in this study.

Conclusion

In this study we compared the effects of using a computer-assisted self-interviewing tool, called Virre, with written responding on the variables: number of words, number of unique clauses, and response avoidance. Taking an intra-sample approach, it was found that students who self-selected into a computer science education program responded to a series of reflection questions with significantly more words and more unique clauses during Virre responding than during written responding. The students were much more likely to use a response avoidance term during Virre responding than in written responding; however, when treating answers that were categorized as a response avoidance as missing data, the number of words and clauses per response was still much higher in the Virre condition. Although these results do not generalize to the population of K-12 students, they do give preliminary evidence that Virre, and computer-assisted self-interviewing tools like Virre, could lead to more data than written responding or face-to-face interviewing. If this theory is supported by further studies, the implications are that researchers, evaluators, and educators could use Virre or Virre-like tools as a supplement or substitute to face-to-face interviewing, which, particularly when large numbers of subjects are to be interviewed, is more expensive and resource-intensive when many interviews need to be done.

Acknowledgement

This research was supported in part by a grant from the Association for Computing Machinery's Special Interest Group on Computer Science Education (SIGCSE).

References

Briggs, D. C. (2005). Meta-analysis: A case study. *Evaluation Review, 29* (2), 87-127.

De Leeuw, E. D. (1993). *Data quality in mail, telephone and face-to-face surveys*, Amsterdam: TT-Publikates (ERIC Document Reproduction Service No. ED 374136).

De Leeuw, E. D., & Nicholls, W. (1996). Technological innovations in data collection. Acceptance, data quality and costs. *Sociological Research Online, 1* (4), Retrieved May 4, 2006, from, http://www.socresonline.org.uk/cgi-bin/abstract.pl?1/4/leeuw.html.

Dillman, D. A. (1978). *Mail and telephone surveys: The total design method*, New York: Wiley.

Eronen, P. J., Jormanainen, I., & Virnes, M. (2003). Virtual reflecting tool - Virre. In *Proceedings of the Third Finnish / Baltic Sea Conference on Computer Science Education*, Finland: Helsinki University of Technology Press, 42-47.

Eronen, P. J., Sutinen, E., Vesisenaho, M., & Virnes, M. (2002a). Kids' Club as an ICT-based learning laboratory. *Informatics in Education, 1* (1), 61-72.

Eronen, P. J., Sutinen E., Vesisenaho, M., & Virnes, M. (2002b). Kids' Club - information technology research with kids. In Kommers, P., Petrushkin, V., Kinshuk, & Galeev, I. (Eds.), *Proceedings of the IEEE International Conference on Advanced Learning Technologies*, Los Almitos, CA: IEEE Press, 331-333.

Galpin, V. (2002). Women in computing around the world. *ACM SIGCSE Bulletin, 34* (2), 94-100.

Gay, G., & Bennington, T. L. (2000). *Information technologies in evaluation: Social, moral, epistemological, and practical implications: New Directions for Evaluation,* New York: Jossey-Bass.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression*, New York: Wiley.

Kids' Club (n. d.). *Welcome to Kids' Club web pages*, Retrieved May 4, 2006, from, http://cs.joensuu.fi/~kidsclub/.

Love, A. (Ed.). (2004). Harnessing technology for evaluation. *The Evaluation Exchange, 10* (3), Retrieved May 4, 2006, from, http://www.gse.harvard.edu/hfrp/content/eval/issue27/fall2004.pdf.

Means, B., Roschelle, J., Penuel, W., Sabelli, N., & Haertel, G. (2003). Technology's contribution to teaching and policy: Efficiency, standardization, or transformation? *Review of Research in Education, 27*, 159-181.

Pressley, M., & McCormick, C. B. (1995). *Advanced educational psychology for educators, research, and policymakers*, New York: Harper Collins.

Randolph, J. J. (2006). *Online appendix to "The Effects of a Computer-Assisted Interview Tool on Data Quality",* retrieved May 4, 2006, from, http://geocities.com/justusrandolph/virre_appendix.pdf.

Randolph, J. J., & Edmondson, R. S. (2005). Using the binomial effect size display (BESD) to present the magnitude of effect sizes to the evaluation audience. *Practical Assessment Research & Evaluation, 10* (14), retrieved May 4, 2006, from, http://pareonline.net/getvn.asp?v=10&n=14.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effects sizes in behavioral research: A correlational approach*, Cambridge, United Kingdom: Cambridge University Press.

Rossi, P. H., Wright, J. D., & Anderson, A. B. (1983). Sample surveys: History, current practice, and future prospects. In Rossi, P. H, Wright, J. D. & Anderson, A. B. (Eds.), *Handbook of survey research*, San Diego: Academic Press, 1-20.

Saris, W. E. (1989). A technological revolution in data collection. *Quality and Quantity, 23,* 333-350.

Saris, W. R. (1991). *Computer assisted interviewing (Quantitative applications in the social sciences, no. 80)*, Newbury Park: Sage.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In Cooper, H, & Hedges, L. V. (Eds.), *The handbook of research synthesis*, New York: Russell Sage Foundation, 261-281.

Shaffer, D. W., & Serlin, R. C. (2004). What good are statistics that don't generalize? *Educational Researcher, 33* (9), 14-25.

Smith, T. W. (1987). The art of asking questions - 1936-1985. *Public Opinion Quarterly, 51*, 95-108.

Teague, J. (1997). A structured review of reasons for the underrepresentation of women in computing. In *Proceedings of the 2nd Australian Conference on Computer Science Education*, New York: ACM Press, 91-98.

Weisband, S., & Kiesler, S. (1996). Self disclosure on computer forms: Meta-analysis and implications. In Tauber, J. M. (Ed.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Common Ground*, New York: ACM Press, 3-10.

Witt, K. J., & Bernstein, S. (1992). Best Practices in Disk-By-Mail Surveys. *Sawtooth Software Conference Proceedings*, Evanston: Sawtooth Software.

Zandan, P., & Frost, L. (1989) Customer Satisfaction Research Using Disk-By-Mail. *Sawtooth Software Conference Proceedings*, Evanston: Sawtooth Software.