Center for Demography and Ecology

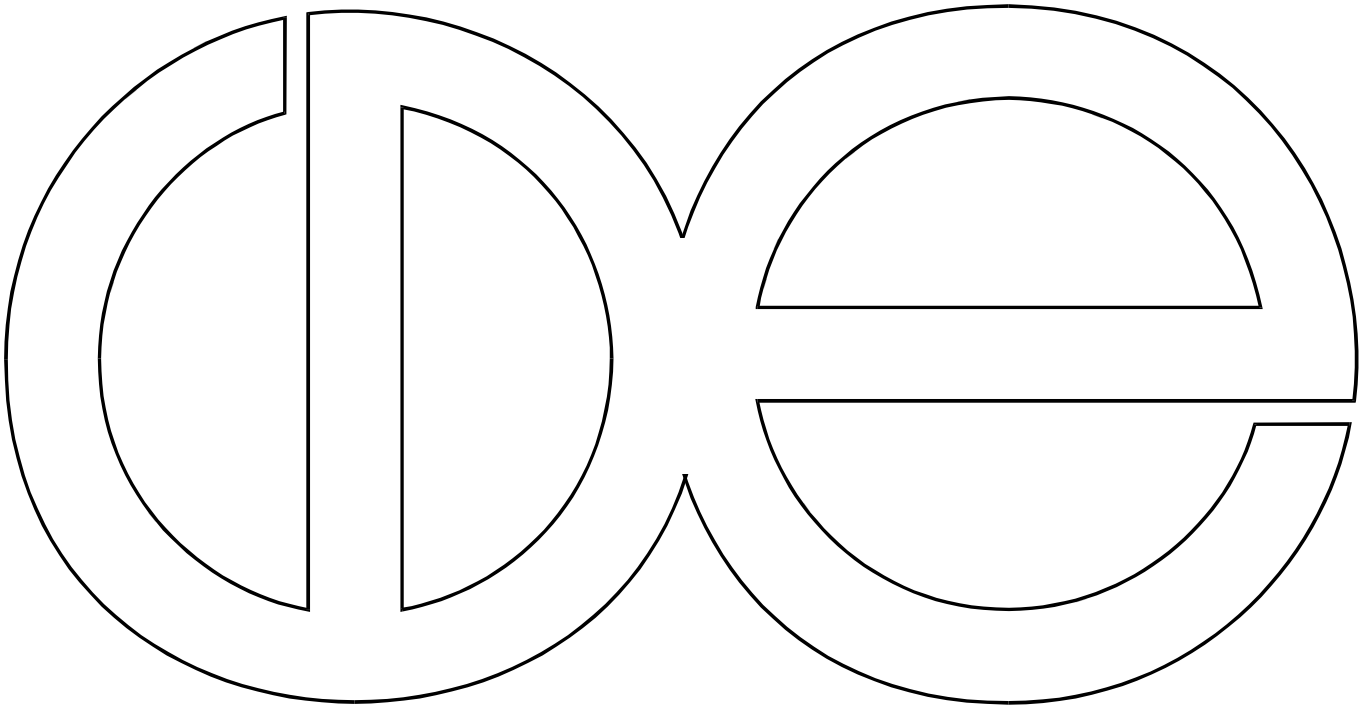
University of Wisconsin-Madison

Paradata Correlates of Data Quality in an SMS Time Use Study: Evidence from a Validation Study

Philip S. Brenner

John D. DeLamater

CDE Working Paper No. 2013-03



Paradata Correlates of Data Quality in an SMS Time Use Study: Evidence from a Validation Study

Philip S. Brenner
University of Massachusetts, Boston

John D. DeLamater
University of Wisconsin, Madison

Abstract

Short Message Service (SMS) text messaging is a ubiquitous technology available on the vast majority of cellphones in use in 2012. It provides a common technological denominator between mobile devices of nearly every make and model, supplying researchers an avenue to collect data without the expense and difficulty of designing specific applications for every cellphone or device on the market. SMS/text messaging was used as a method of data collection using a sample of students from a large, Midwestern university. The procedure adapted conventional time use measurement procedures to fit the device, the sample, and the behavior of interest. After answering questions on a brief Web survey, respondents were asked to text researchers for five days, updating major changes in their activities. Following data collection, data from the text condition was compared to that from a conventional (Web) survey and data from a reverse record check from campus recreation facilities to validate reports of the behavior of interest — physical exercise and activity. Findings suggest that respondents provided consistently high quality data on self-reports of the behaviors of interest. Moreover, paradata measures of text data quality (e.g., number of text messages sent, number of days with messages) predict data quality on the behavior of interest.

Introduction

Short Message Service (SMS) text messaging — or, texting — is widely and frequently used by young adults. In a recent study by researchers at Ball State University, 99 percent of students reported having a cellphone, and virtually all of these students (97 percent) reported sending and receiving text messages (Ball State University 2009). Many of these young adults text prodigiously. A recent study by the Pew Research Center estimated that young adults send an average of 109.5 texts a day.[1] Moreover, the heaviest users of texting prefer text to talk. Over half (55 percent) of adults sending and/or receiving more than fifty text messages a day prefer a text message to a phone call.[2]

More than just the ubiquity and utilization of the technology makes it of interest to social scientists in search of data collection opportunities. Perhaps even more important is the manner of its use. In conjunction with other more recent, web-based social networking technologies and applications (e.g., Instant Messaging [IM], Facebook, Twitter, Foursquare, Google+), texting is used to report current activities and locations to others. All of these technological tools provide researchers with new opportunities for data collection, as well as data mining, to address a wide variety of social science concerns.

However, SMS provides researchers a data collection opportunity not shared by its more recent competitors. SMS is a ubiquitous technology available on nearly all cellphones in use today. It does not require state-of-the-art technology or cutting edge consumer electronics (e.g., a smartphone running the latest version of Google's Android or Apple's iOS) nor does it require

[1] The same study put the median number of texts per day at about 50. The difference between the mean and median shows that the distribution is highly positively skewed, suggesting the presence of some very extraordinary texting outliers.

[2] Notably, this survey was conducted, via voice, to landline and cell numbers. The response rate for the cell sample was 11.5 percent, two points less than that for the landline sample.

additional software development or any intermediary Web-based application (e.g., a Twitter client, a Facebook app, or a custom software application) for data collection (Raento, Oulasvirta, and Eagle 2009). Rather, SMS/text messaging provides a common technological denominator between cellphones of nearly every make and model, supplying researchers an avenue for data collection without the expense and difficulty of designing specific applications for every cellphone on the market.

Moreover, texting provides a perception of privacy and confidentiality unavailable (or not easily available) with Web 2.0-based social networking applications.[3] Other current messaging applications (e.g., Twitter, Facebook, Foursquare) are a one-to-many communication technology by their very nature. Using these services, a user sends a report on one's current activity, location, or state of mind for multiple (or all) other users of the service to see. This default, rightfully (and hopefully) leads to a selection bias of what is shared and what is not (see Patchin and Hinduja 2010). Texting, however, is inherently a one-to-one communication channel and lends itself more naturally to a data collection procedure in which a confidentiality assurance can be implemented.

However, for all of its benefits, text messaging as a data collection mode has drawbacks. For survey researchers specifically, it presents obstacles to the conventional standardized question and response scale paradigm (see Fowler and Mangione 1990). Certainly, standardized questions and response options could be sent by way of text messages to the respondent with instructions for the respondent to select an answer to and respond with the numerical code reflecting their answer back to the researcher. However, texting is, by the nature of the medium,

[3] This is not to say that the transmission of text messages is perfectly confidential. However, texters view their phones as private devices and believe that there is a "widely accepted, unwritten rule" about the confidentiality of text messages (Häkkilä and Chatfield, 2005).

idiomatic. Unlike a Web survey with checkboxes or radio buttons, consistency checks and forced response, there is nothing to prevent the SMS/text respondent from answering how s/he sees fit, regardless of the standardized options. While respondents could potentially be trained to respond with a number associated with a response option, changing the expressive nature of the text message to force it into the standardized questionnaire paradigm fails to capture the strength of the method.[4]

Using SMS for time use data collection

Many SMS-based data collection procedures previously used, even those labeled "diaries," have been somewhat more akin to the Experience Sampling Method (ESM) or conventional survey data collection. For example, in an "SMS Pain Diary" Alfvén (2010) asked respondents to reply to six messages a day using a prearranged coding scheme to report intensity, duration, and results of pain. Similarly, Anhøj and Møldrup (2004) used SMS to send a series of yes or no questions measuring the occurrence of Asthma symptoms and use of medication to respondents at preselected times during the day.

In these examples and other extant work, researchers fail to leverage the strengths of using SMS for diary data collection. The idiomatic nature of SMS/text messaging is a strength of the time diary method of data collection. The strength of chronologically based data collection procedures, like time diaries, is in their ability to avoid the measurement bias that plagues direct survey questions on normative behaviors (Bolger, Davis, and Rafaeli 2003). Like other normative behaviors, physical exercise is widely understood to be overreported in surveys

[4] Verbatim responses are not without their own problems, of course. Each message requires coding; an expensive and time-consuming proposition. Moreover, the nature of text messaging is miserly with time and effort, with a focus on abbreviation. Many of the abbreviations used in texting are now well known and do not necessarily present coding problems, although idiosyncratic abbreviations or acronyms may.

using conventional direct questions (Ainsworth, Jacobs, and Leon 1992; Chase and Godbey 1983; Klesges 1990). Verbatim responses to open-ended questions (i.e., "What did you do next?") allow researchers to avoid direct questions about specific behaviors of interest (i.e., "Did you go to the gym?") (Robinson 1985, 1999; Stinson 1999), thereby avoiding prompting self-reflection on the part of the respondent, and yielding less biased and higher quality data on many normative behaviors (Bolger et al. 2003; Niemi 1993; Zuzanek and Smale 1999.)

Like all data collection procedures, chronologically based data collection procedures also have weaknesses, two of which are pertinent to this conversation. First, respondents may fail to report activities of very brief duration that happen frequently during the day. For example, trips down the hall to use the restroom or to the water fountain are likely to be omitted as respondents tend to focus on the sorts of activities that the day is planned around and that last for hours rather than minutes. Therefore, the focal activities of such a data collection procedure should be activities that last for hours, rather than minutes, and should preferably be the sorts of activities around which individuals plan their day.

A second main weakness of chronologically based data collection procedures is primarily related to the heavy burden they place on respondents. This burden can result in high rates of nonresponse — either through high rates of refusals to participate that yield increased unit nonresponse or in incomplete participation as respondents quit the study or choose to participate intermittently, resulting in partial "interviews" and "item" nonresponse.[5] In order to reduce the burden of the data collection process, diaries can be, either by the researcher's design or by the unilateral decision of the respondent, filled out at the end of the day or at the end of the reference

[5] The AAPOR Standard Definitions and other nonresponse terminology, while still very useful, fit somewhat awkwardly in the case of time use data collection. For example, there are not "items," *per se*, to be skipped, although certainly skipping parts of the data collection process yields a similar outcome.

period. However, shifting the timing of diary completion away from the time of occurrence of individual activities can result in poorer data quality as respondents may introduce errors into the data collection procedure, like forgetting to include events or attributing them to incorrect times.

The SMS procedure offers some promise as it incorporates features that address these weaknesses and may lead to higher quality data. First, respondents can be asked to report on attitudes or behaviors *in situ* and as they occur. This application of a real-time data collection procedure may help to reduce forgetting and other memory problems. Second, the procedure may overcome another problem with retrospective reporting — editing and judging. Without the time to reflect and put activities and feelings in context, an SMS-based reporting procedure may be able to avoid much of the social desirability effect and other sources of bias inherent to standardized survey questions. While perhaps not true for all behaviors and activities, especially contranormative, illegal, or embarassing activities (e.g., illicit drug use or sexual activity), or those of high frequency and brief duration (e.g., using the restroom or getting a drink of water), this procedure should allow a more accurate measurement of the normative activities that are often overreported and that could be considered major activities in a day's schedule (e.g., going to religious services, volunteering, or exercising).

Third, using a technology that young adults find relevant to their daily lives may yield a more representative achieved sample. Commonly used sampling designs, like random digit dialing, typically produce sampling frames that yield undercoverage of the young adult population (Blumberg and Luke 2007; Currivan, Roe, and Stockdale 2008). Making matters even worse, conventional survey modes commonly result in high rates of nonresponse amongst sampled individuals in this age group (Groves and Couper 1998). Combined with an appropriate

sampling design, this adoption and adaptation of a technology used frequently by young adults may provide an additional level of interest to leverage their participation (Groves, Singer, and Corning 2000). In sum, using texting in a manner similar to other diary-like Web-based applications (i.e., Facebook and Twitter), may encourage the participation of young adults, garnering higher rates of cooperation than more conventional data collection methods.

While not a panacea, an SMS-based chronological data collection procedure does offer some promise in reducing these forms of error. However, the promise of this procedure depends on three important considerations. First, the target population must be one that fits well with the method (e.g., a population with a high rate of ownership and use of cellphones, like young adults).[6] Second, the sampling frame needs to contain cellphone numbers or would have to be relatively easily switched between a recruitment mode (e.g., Web, landline phone, or mail) to cellphone number for data collection. Given the requirements of the first point, Web would be the obvious choice here. Third, the research problem or question must be one that fits the method well (e.g., an interest in major activities, rather than very frequent but short-duration activities).

The current project matches these requirements well. This technology was used to obtain reports from a sample of university undergraduates regarding their daily activities. The research was focused specifically on the validity of measurement of physical exercise although this emphasis was not disclosed to respondents. Since this is one of the first attempts to implement this method in a rigorous research project, it is useful to examine these data to determine how well the method worked, the quality of the data it produced, and what can be done to improve

[6] If used in a more general population, adequate funding must be available to purchase text-enabled cellphones for respondents and provide training for their use.

each. To this end, a series of paradata indictors will be used to predict the observed criterion validity of the focal behavior, physical exercise at university recreational sports facilities.

Data and methods

A random sample of 325 undergraduates, stratified by gender and year in school, from a large, Midwestern university were sent an email invitation to participate in the "[University Initials] Student Daily Life Survey" in March and April 2011. The invitation was sent to the student's university email address and included a link to a Web survey. An email reminder to complete the survey was sent three days after the initial invitation, and a final reminder was sent five days after the first reminder email.

The Web survey was comprised of approximately twenty questions about usage of university facilities. While the true purpose of the study was to measure use of university recreation facilities, questions about type and frequency of use of campus libraries, the student union, and other facilities were also asked in order to mask the focus of the study. Respondents were asked about their "typical" use of recreational facilities on campus and their "usual" activities at these facilities (e.g., weightlifting, swimming, aerobics, and cross-training). Respondents received ten dollars upon completion of the Web survey. 124 respondents completed the Web questionnaire yielding a response rate of 38 percent.[7]

The final question of the Web survey was a request to participate in the SMS/text message data collection procedure. Respondents were told that participation in this part of the project entailed sending SMS/text messages to the research team reporting all changes in their

[7] All response rates are computed as AAPOR RR 5, as there are no ineligible cases or cases of unknown eligibility.

major activities for a period of five days. In acknowledgment of their participation, respondents were told that they would receive an additional thirty dollars at the conclusion of their participation. If the respondent was amenable to participating, s/he was asked to enter his or her cellphone number. 87 percent (108 of 124) of the respondents who completed the Web survey agreed to continue into the text component of the study.

Respondents received training materials detailing how and what to report. These training materials were comprised of a single, two-sided document, emailed to the respondent. The first page described the purpose of the study, the tasks required of the respondent, an example of a full day of nine text messages, and instructions on how to text updates to the research staff. Respondents were asked to report all changes in the major daily activities and where they were taking place without reference to particular activities. The second part of the document was a FAQ list, including instructions on how to report late activities and whom to call or email with questions or concerns.

Respondents were then assigned to one of five five-day field periods. Cohorts of text respondents were distributed over a two-week period to ensure coverage of both weekday and weekend days. Respondents were reminded multiple times each day to send messages updating their activities, regardless of how recent their most recent update was. These reminders were more frequent on the first day of their participation (four times, at 10:00 am, 1:00 pm, 5:00 pm, and 8:00 pm) and less frequent on the final days of participation (two reminders, at 10:00 am and 8:00 pm). 81 percent (87 of 108) of the respondents who agreed to participate in the text condition sent at least one text during the field period.

At the completion of the texting component of the study, each respondent was asked for his or her student identification number so that study staff could request records on each respondent's use of campus recreation facilities. These records are the product of the swiping of students' identification cards upon admission to the facilities. This process records the student's identification number and the time and day of admittance to the facility. 77 percent (67 of 87) of the respondents who completed the text condition permitted access to their record data, yielding final effective response rates of 27 percent for all texters and 20 percent for respondents allowing access to verification data.

Measures

Six paradata measures of data quality were observed and will be used as independent variables in the following analyses: (1) the total number of text messages sent, (2) the number of days the respondent sent messages, (3) the percent of messages sent late, (4) the number of days the respondent skipped, (5) the percent of reports that are temporally proximal to a reminder text, and (6) the number of messages that are repeats of prior messages. Two outcome variables will be used in the following analyses: (1) the validity (whether overreported or underreported) of the respondent's claim of the number of days s/he exercised at University facilities, and (2) an indicator of respondent compliance with the record check procedure. Each of these will be described in greater detail.

Total number of messages. The number of messages sent is clearly an indicator of data quality. The fewer messages sent by the respondent, the more poorly these messages will represent the respondent's day.[8] Respondents sent a total of 1904 messages, ranging from 2 to 59

[8] Clearly, this will vary by day of the week. Weekdays tended to have more activity than weekend days, especially Sunday, which elicited the fewest number of messages.

messages per respondent (omitting the respondents who agreed to participate in the texting component of the study but did not send a text.) Respondents averaged 22 messages (s.d. = 10.8) during their assigned field period of five days, or about five messages per day.[9]

Number of messaging days. Respondents were assigned to one of five five-day reporting periods to distribute reporting across the seven days of the week. On average, respondents submitted messages for 5.1 days (s.d. = 1.1), ranging from 2 to 8 days. Most respondents (81 percent) reported activities for at least five days. As this suggests, a number of respondents (31, or 36 percent) reported activities for more than the requested 5 days, while 19 respondents (22 percent) reported on fewer than five days. Clearly, sending messages on fewer than the complete assigned five day period will negatively affect data quality.

Percentage of late messages. Late messages were those flagged by the respondent as reporting on activities occurring prior to the sending of the message. Knowing that respondents would likely forget to report some changes at the time they occurred, respondents were advised that, if necessary, they could report activities late by including a flag (the word "TIME" in all capital letters) and the time of the activity in the report. Respondents averaged five late reports during the field period, ranging from zero to 80 percent of their messages. Approximately 21 percent of reports were sent late (s.d = .23), and more than half (57 percent) of the respondents reported late one or more times. As can be seen from the range, some participants provided many late reports. Thirteen participants, 15 percent, texted more than half of their reports after-the-fact.

[9] In determining date received, messages received after midnight which reported an activity at the end of the day, typically "going to bed," were coded as received the previous day.

Number of skipped days. The integrity and validity of these data depends on every participant reporting each day during their assigned field period. Therefore, the number of skipped days may also be a good indicator of data quality. Skipped days are not just a mathematical function of the number of messaging days and the number of days in the reference period. Some respondents who skipped a day in the middle of their assigned reference period continued to report after their assigned field period had ended, perhaps in an attempt to make up for the missed day. About 30 percent (26 respondents) skipped one or more days. The average number of skipped days was greater than a third of a day (0.40) per respondent, ranging from zero to three skipped days. Sundays were especially likely to be skipped; almost two-thirds of the respondents with skipped days (16 respondents) resulted from a failure to report activities for an assigned Sunday. Since the sample was drawn from an undergraduate student population and many of the provided examples were student-related activities, respondents may have felt it was unnecessary to report Sunday leisure activities.

Percentage of messages proximate to reminders. Another way to measure data quality is to evaluate responses by their proximity to reminder messages. There is no reason to believe that students would be engaging in new activities in any kind of systematic way at 10am, 1pm, 5pm and 8pm, and only at these times. Therefore, a high rate of messages proximate to these reminder messages suggests that the respondent may only be reporting activity changes in reaction to these prompts, resulting in poor data quality. The average rate of messages sent proximate to a reminder was about 19 percent (s.d. = .15), where "proximate" is defined as within thirty minutes following a reminder message. The observed range of proximity is very large, with minimum and maximum values matching theoretical limits: some respondents sent

all of their messages just after a reminder, whereas other respondents did not send any messages proximate to a reminder.

Number of repeated messages. A careful reading of the corpus of messages indicated a small number of cases where the same message was sent twice within a few minutes. A message was considered repeated if two texts reporting the same activity were sent on the same day within 10 minutes of each other. Typically, the activity was reported twice (e.g., "Going to Target.") In two cases, the second message expanded the information contained in the first (e.g., "Going to the grocery store." "The one in [building name].").

Validity of the report of exercise. The first outcome measure, the validity of the reporting on exercise activity, was computed as the difference between the reverse record check and the self-report from the respondent. Reported changes in respondents' major activities were coded for exercise activities and, more specifically, for those that occurred at campus recreational sports facilities. Each day with a report of exercise at a campus recreational sports facility was coded as 1, 0 otherwise. This variable was then summed over the days of the reference period.

Each day during the reference period with record of admittance to a campus recreational sports facility was coded as 1, 0 otherwise. This procedure yielded a series of variables, one for each day, each coded for the presence or absence of an admittance. These were summed to reflect the number of days during the reference period that the respondent used campus recreational sports facilities.

The difference between the self-report and the record variable provided an estimate of the validity of the self-report of exercise. This procedure resulted in a three-category nominal variable: (0) valid reporters, (+1) overreporters, and (-1) underreporters. Due to small cell sizes,

the latter two categories are collapsed in some analyses creating a dichotomous variable for comparison of accurate and inaccurate reports. Notably, these data appear to be of very high quality. About 80 percent of respondents reported accurately, their claims verified by the reverse record check. The remainder of cases were equally split between over- and underreporting suggesting that measurement error was random rather than systematic.

Compliance with record access. Finally, comparisons will be made with respondents for whom these validity data are available and those for whom these data are not available (i.e., those respondents who did not allow access to the record data). It is possible that respondents who disallowed access to their gym access records differ in a systematic way in their data quality from compliant respondents who allow access to these records. This analysis addresses this possibility.

Analysis plan

Two methods were used to examine the quality of these data and the value of the paradata indicators as predictors of the criterion validity of the measure of the focal behavior — physical exercise. The first method applied a cluster analysis to the full dataset (all text respondents, with or without validation data) to generate a typology of respondents in terms of the paradata indicators of data quality. A *k*-means cluster analysis was estimated using a set of paradata variables from the text respondents, including number of days the respondent sent text messages, the total number of messages sent, the number of days the respondent skipped, and the number of late messages sent. These clusters were then compared using two outcomes computed from the validation procedure: (1) rate of inaccurate reporting, and (2) and rates of compliance for the

reverse record check. Comparisons use Fisher's exact text and Cohen's *d* to assess statistical and substantive significance of the predictive value of the paradata indicators as a whole on data quality.

The second analysis uses logistic regression to predict the propensity of respondents to over- or underreport, given these paradata indicators of data quality. This analysis expands on the previous comparisons in two ways. First, it assesses individual paradata indicators of data quality given the criterion measure, discerning those which have predictive validity from those which do not. Second, this analysis permits for separate prediction of both overreporting and underreporting, allowing for a more nuanced understanding of the nature of the error in the self-report of exercise and the effect of the paradata determinants of data quality in the assessment of validity.

Results

Subjective assessment of the results of the cluster analysis suggests that the most parsimonious model allows four clusters of respondents to emerge (see Table 1). For purposes of presentation, these clusters have been given descriptive names: (1) Prodigious texters, (2) Frequent texters, (3) Occasional texters, and (4) Infrequent texters.

[Table 1 about here.]

The Prodigious texters of the first cluster comprised less than ten percent of the achieved sample (8 of 87 respondents). Respondents in this cluster sent an average of 44 messages during the reference period, yielding over eight messages a day on average, with no skipped days.

About thirty percent of their messages were late and about 14 percent of their messages were sent shortly after reminder texts.

The second cluster, Frequent texters, comprised over a third of the achieved sample (31 of 87 respondents). The main difference between the Prodigious and Frequent texters was the number of messages sent: Frequent texters sent about a third fewer messages than the Prodigious texters. The respondents in this cluster sent about 28 messages during the reference period, yielding over five messages a day on average, skipping very few days. Very similar to the Prodigious texters, Frequent texters' messages were late about a quarter of the time and they sent about 15 percent of their messages shortly after reminder texts.

Occasional texters comprised the largest cluster of respondents at nearly 40 percent of the achieved sample (34 of 87 respondents). Occasional texters sent almost half the number of messages than the Frequent texters (approximately 17 messages during the reference period.) The Occasional texters also skipped about a third of a day on average, yielding fewer than 3.5 messages a day. About 22 percent of their messages were sent late and nearly 20 percent were sent shortly after reminder texts.

The final cluster, Infrequent texters, comprised about a sixth of the achieved sample (14 of 87 respondents). The respondents in this cluster sent only about seven messages during their entire reference period, averaging just over one message a day. These respondents shortened the intended reference period by over a day, skipping 1.3 days on average. Infrequent texters, however, sent very few late messages (about three percent). This rate of timeliness is not surprising given how few messages Infrequent texters sent. Moreover, of those messages, almost a third were sent within thirty minutes of a reminder text.

How do these clusters of respondents, generated using the indicators of data quality from the texting paradata, compare given the outcome of the validation procedure? First, consider the distribution of the 20 respondents who did not allow access to their recreational sport facilities admission records. These respondents were evenly distributed across categories: ten were in the top two categories of better respondents and the other ten were in the bottom two categories of poorer respondents. Thus, the respondent's decision to grant access to their record data is not associated with the quality of the respondent's texting performance.

But the more important question is whether these clusters based on paradata have predictive validity. Table 2 compares respondents in these clusters by the outcome of the validation procedure. While cell sizes are small, there appear to be a number of important differences emerging. First, the rates of invalid responses (i.e., under- and overreports) appear to be higher for the Occasional and Infrequent texters. Ten percent of respondents in the Prodigious and Frequent clusters inaccurately report their exercise, but nearly 30 percent of the Occasional and Infrequent texters inaccurately report. While the raw between-group difference is quite large (Δ=20 percentage points; Cohen's d = 0.47), the small effective sample size (N=67) leads it to be just outside of conventional levels of statistical significance using either Fisher's exact test (p = 0.058) or Chi-square (p = 0.064).

[Table 2 about here.]

Logistic regression models predicting over- and underreporting

More direct tests of these potential indicators of data quality can be undertaken to predict the validity of the exercise measure. These tests will allow us to see which of these paradata indicators of data quality have the most purchase in explaining the quality of the exercise data.

In addition, these tests will allow the error in the exercise measure to be separated into its two components: overreporting and underreporting.

Logistic regression models were estimated predicting overreporting and underreporting using each of the indicators of data quality: number of messages, number of days with messages, number of skipped days, number of repeated messages, and percentages of late messages and messages sent following a reminder. Results show an important difference between the two forms of error. While none of these indicators predict overreporting in bivariate models, two bivariate models approach conventional levels of statistical significance when predicting underreporting. Both the number of messages sent ($\beta = -0.08$; $p = 0.09$) and the number of days with messages ($\beta = -0.72$; $p = 0.052$) predict underreporting, although the p-value of these tests is just outside conventional levels of statistical significance. As would be expected, these relationships are negative; each additional message sent yields a reduction in the odds of underreporting by eight percent. Moreover, each additional day of messaging leads to a reduction in the odds of underreporting by 50 percent.

Since these two indictors of data quality are highly correlated ($r = 0.7$), including them both in a multivariate model results in multicollinearity. Therefore, a new variable, average number of messages per day, was computed as the dividend of these two indicators. A similar finding emerged when underreporting was regressed on this new variable. Every unit increase in the average rate of messages per day reduces the odds of underreporting by about a third ($\beta = -.46$; $p = 0.07$). This finding, like those from previous models, is of marginal statistical significance, but suggests that the number of messages and the number of messaging days may be predictive of the validity of key measures.

Discussion

Clearly, the most important paradata indicators in predicting data quality are (1) the number of messages and (2) the number of days with messages. These two indicators vary a great deal from the best cluster of respondents (44 messages over all five field days) to the worst cluster of respondents (7 messages with 1.3 field days missed). The distinction between the best and the worst clusters of respondents is stark — a 20 percentage point difference in the validity of their responses. Moreover logistic regression modeling supports this finding. Both the number of messages and the number of messaging days predict data quality — the more of each, the less likely the respondent is to underreport their exercise.

The strength of the paradata indicators of data quality in predicting only one of the two forms of error in the exercise measure may be explained by understanding the nature of these two forms. Overreporting is an error of commission; the respondent has made a claim that cannot be verified. The inability of the indicators of data quality to predict overreporting is understandable as the method is more prone to errors of omission than commission. In contrast, underreporting is an error of omission. The most likely cause of this error is nonresponse. This could take a couple of guises, like forgetting or intentionally failing to report on an activity, choosing to end participation in the study early, or skipping days in the middle of the reference period.

Surprisingly, this last type of nonresponse — skipped reporting days — does not increase one's likelihood to underreport. This may be due to a problem with nonresponse, typified by many students' Sunday reports. A number of respondents reported very few Sunday activities,

texting only a message like “staying in today” or “at home studying.” It is possible that other respondents with a similar level and type of activity failed to report days in which they did not venture out from home. If this is the nature of a skipped reporting day, it is clear why this indicator of data quality would not predict underreporting of exercise at a campus recreation facility. Clearly, in future applications of this method, researchers must more clearly and carefully specify which types of activities should be reported and emphasize reporting on each day included in the reporting period.

Surprising, at least initially, is the comparability of the rate of late messaging in the two clusters of more conscientious respondents (Prodigious and Frequent texters) with the somewhat less conscientious respondents in the Occasional texter cluster. This, in combination with its weakness as a predictor in the logistic regression models, suggests that lateness, in and of itself, is likely not a good indicator of data quality. Rather, it may be an inevitable result of this sort of *in situ* data collection procedure. The rate of lateness, and the lack of an effect of lateness on data quality, suggests that respondents should be told that, while not ideal, sending late messages is understandable and a process should be created to allow respondents to send researchers late reports of their activities, such as that which was provided.

But can lateness be combated with well-timed reminder messages to prompt respondents to update researchers on their recent activities? Compare the findings on lateness with those on the percentage of messages sent after a reminder. In the two clusters of more conscientious respondents (Prodigious and Frequent texters), this rate is between 14 and 15 percent. This rate increases to 19 percent for the Occasional texters, and to 32 percent for the Infrequent texters. This suggests that poorer respondents are either less likely to remember the task of reporting or

more likely to wait for a reminder whereas better respondents are more proactive in reporting their activities. With that said, the difference between the top three categories is not large. Further research on the role of reminders may help to clarify their role in data quality; that is, do reminders prompt otherwise good respondents to improve the quality of their data, or do they spur poor respondents to give only a barebones effort?

Limitations

Perhaps the largest single problem with this study is the low response rate. In order to meet the requirements of the human subjects review board, the design of the study required multiple requests for participation from respondents, creating multiple opportunities for respondents to decide to discontinue their participation. These include (1) the initial request for participation, (2) the request for the respondent's cellphone number, (3) the instruction to begin the text component of the study, and (4) the request for the respondent's student identification number for the collection of validation data. With each subsequent request, some sample members inevitably failed to continue participation. With that said, additional analyses do not suggest that nonresponse has biased estimates (results not shown). For example, the rate of compliance for the reverse record check does not differ between clusters; 79 percent of the Occasional and Infrequent respondents allowed access to their records compared to 75 percent of the Frequent and Prodigious texters. Future research should attempt to combine these requests or better link each step to the payment of incentives.

Perhaps the second most important problem is the relatively small sample size, exacerbated by the low response rate. The small sample size limits the analyses that can be pursued. For example, in some analyses, over- and underreporting were pooled into a single category for comparison with accurate reports. Moreover, small cell sizes lead to a lack of statistical power. Nevertheless, the findings are suggestive and are meant to spur further research.

The sample used here was of undergraduate students at an elite public Midwestern university. As such, findings are hardly generalizable to a national population or even to the larger population of young adult Americans. This sample was good for testing the main hypothesis (see Brenner and DeLamater 2012). Moreover, this sample was ideal for avoiding the self-selection bias that is inherent in similar studies. All students in the sampling frame automatically have access to the campus recreation facilities without making the effort to join (and pay) for membership. Unlike a sampling frame from a similar organization comprised of members of the general population, (e.g., membership rolls at a YMCA or a for-profit fitness center), the sampling frame from the university registrar or bursar allows a frame of gym members (i.e., all students) without a self-selection bias. Nevertheless, future research should attempt to use a sampling frame from a more varied target population.

Conclusion

A time use study was undertaken, adapting conventional time diary procedures to fit with the mode of data collection — SMS/text messaging. Data collected using this novel mode were compared to that from a reverse record check from campus recreational sports facilities to

validate the behavior of interest — physical exercise and activity. These comparisons suggested that these data were of high quality overall, with 80 percent of cases generating valid data on the variable of interest and the remaining cases equally distributed amongst over- and underreporting, leaving the population estimate unbiased.

A cluster analysis using a set of six paradata indicators predicted nearly 80 percent of the cases with misreported exercise. Moreover, testing the predictive validity of these paradata indicators in a logistic regression model suggested that only two — the number of messages sent and the number of days the respondent sent text messages — are important for distinguishing between cases with valid and invalid data. This finding suggests that improvement to the measurement procedure (i.e., increasing the number of messages sent and ensuring that respondents report on activities during all the days of the field period) may even further improve data quality.

The high quality of these data did not come at a steep price. Costs were limited to incentive payments – forty dollars per completed case. While some of the suggestions made here to further increase data quality (e.g., using an HTTP-to-SMS service that allows automated reminders; increasing incentives to improve the response rate) would increase this cost somewhat, this method would still be cost-effective compared to face-to-face or telephone diary interviews. These suggestions would allow researchers to automate data collection and handle a much larger sample.

Under the right conditions, this method is clearly viable. For an appropriate target population (e.g., one with near saturation of text-capable cellphones, like a college-age sample, young professionals, or teens, among others), and with a suitable sampling frame that

accommodates such a procedure, this method provides another tool in the survey researcher's data collection kit. It allows researchers to use cellphones for data collection without the trouble and expense of providing equipment or specially designed applications to respondents. Moreover, the high rate of cellphone adoption in developing countries (in lieu of landlines) makes this method a possibility for data collection in areas where time use studies would otherwise necessitate personal interviews.

References

Ainsworth, Barbara E., David R. Jacobs, and Arthur S. Leon. 1992. "Validity and reliability of self-reported physical activity status: the Lipid Research Clinics questionnaire." *Medicine and Science in Sports and Exercise* 25:92-98.

Alfvén, G. 2010. "SMS pain diary: a method for real-time data capture of recurrent pain in childhood." *Acta Paediatrica* 99:1047-1053.

Anhøj, J. and Møldrup, C. 2004. "Feasability of collecting diary data from asthma patients through mobile phones and SMS (short message service): Response rate analysis and focus group evaluation from a pilot study." *Journal of Medical Internet Research,* 6(4).

Brenner, Philip S. and John D. DeLamater. 2012. "Overreporting of Exercise in a Self-administered Mode: The Biasing Effect of Identity on Survey Questions." Presented at the 107th Meeting of the American Sociological Association, Denver, August 18.

Ball State University. 2009. "Survey finds smart phones transforming mobile lifestyles of college students." Press Release. Retrieved from http://www.bsu.edu/news/article/0,1370,--61565,00.html.

Blumberg, Stephen. J. and Julian V. Luke. 2007. "Coverage Bias in Traditional Telephone Surveys of Low-Income and Young Adults." *Public Opinion Quarterly* 71:734-49.

Bolger, Niall, Angelina Davis, and Eshkol Rafaeli. 2003. "Diary Methods: Capturing Life as it is Lived." *Annual Review of Psychology* 54:579-616.

Chase, David R. and Geoffrey C. Godbey. 1983. "The accuracy of self-reported participation rates." *Leisure Studies* 2:231-35.

Currivan, Doug., David Roe, and Jason Stockdale. 2008. "The Impact of Landline and Cell Phone Usage Patterns among Young Adults on RDD Survey Outcomes." Presented at the 63rd Annual Conference of the American Association for Public Opinion Research, New Orleans, May 16.

Fowler, Floyd J., Jr. and Thomas Mangione. 1990. *Standardized Survey Interviewing: Minimizing Interviewer-Related Error.* Newbury Park, CA: Sage.

Groves, Robert M., Eleanor Singer, and Amy Corning. 2000. "Leverage-Saliency Theory of Survey Participation." *Public Opinion Quarterly* 64:299-308.

Häkkilä, Jonna and Craig Chatfield. 2005. "'It's like if you opened someone else's letter': User Perceived Privacy and Social Practices with SMS Communication." *MobileHCI '05: Proceedings of the 7th International Conference on Human Computer Interaction with Mobile Devices and Services*: 219-22.

Klesges, Robert C., Linda H. Eck, Michael W. Mellon, William Fulliton, Grant W. Somes, and Cindy L. Hanson. 1990. "The accuracy of self-reports of physical activity." *Medicine and Science in Sports and Exercise*, 22:690-97.

Niemi, Ilris. 1993. "Systematic Error in Behavioral Measurement: Comparing Results from Interview and Time Budget Studies." *Social Indicators Research* 30:229-44.

Patchin, Justin W. and Sameer Hinduja. 2010. "Changes in Adolescent Online Social Networking Behaviors from 2006 to 2009." *Computers in Human Behavior* 26:1818-21.

Pew Research Center. 2011. *Americans and Text Messaging*. Retrieved from http://pewinternet.org/Reports/2011/Cell-Phone-Texting-2011.aspx.

Raento, Mike, Antti Oulasvirta, and Nathan Eagle. 2009. "Smartphones: An Emerging Tool for Social Scientists." *Sociological Methods & Research 37:*426-54.

Robinson, John P. 1999. "The Time-Diary Method: Structure and Uses." in *Time Use Research in the Social Sciences*. W.E. Pentland, A.S. Harvey, M.P. Lawton and M.A. McColl, eds. New York: Klewer/Plenum.

-----. 1985. "The Validity and Reliability of Diaries versus Alternative Time Use Measures." in Time, Goods, and Well-Being, F.T. Juster and F.P. Stafford, eds. Ann Arbor, MI: Institute for Social Research. Stinson, Linda L. 1999. "Measuring How People Spend Their Time: A Time Use Survey Design." *Monthly Labor Review* 122:12-19.

Zuzanek, Jiri and Bryan J. A. Smale. 1999. "Life-Cycle and Across-the-Week Allocation of Time to Daily Activities." in *Time Use Research in the Social Sciences*. W.E. Pentland, A.S. Harvey, M.P. Lawton and M.A. McColl, eds. New York: Klewer/Plenum.

Table 1.
Mean numbers, rates of key independent variables, by cluster

		Mean number of				Mean percentage of messages	
	Respondents	**Messages**	**Days**	**Skips**	**Repeats**	**Late**	**After reminder**
Clusters							
Prodigious	8	44	6.3	0	0.75	31%	14%
Frequent	31	28	5.5	0.13	0.19	25%	15%
Occasional	34	17	5.1	0.38	0.09	22%	19%
Infrequent	14	7	3.7	1.29	0	4%	32%

Table 2.
Validation of exercise reports, by cluster

	Result of the validation procedure				
	Valid		Not valid		Total
Clusters	N	%	N	%	
Prodigious	5	83.3%	1	16.7%	6
Frequent	21	91.3%	2	8.7%	23
Top two	26	89.7%	3	10.3%	29
Occasional	19	70.4%	8	29.6%	27
Infrequent	8	72.7%	3	27.3%	11
Bottom two	27	71.1%	11	28.9%	38

Table 3.
Bivariate logistic regression coefficients from models predicting underreporting and overreporting

	Underreporting			Overreporting		
	coeff.	s.e.	*p*	coeff.	s.e.	*p*
Number of messages	-0.080	0.047	+	-0.007	0.037	
Number of days	-0.716	0.369	+	-0.038	0.387	
Number of messages per day	-0.459	0.258	+			
Number of skips	0.622	0.469		0.123	0.548	
Percentage late	-2.940	2.420		-2.193	2.231	
Number of repeats	-0.387	1.002		-0.387	1.002	
Percentage reminder	2.914	2.005		1.494	2.115	

Note: $^{+}p < .10$; N = 67

Center for Demography and Ecology
University of Wisconsin
1180 Observatory Drive Rm. 4412
Madison, WI 53706-1393
U.S.A.
608/262-2182
FAX 608/262-8400
comments to: delamater@ssc.wisc.edu
requests to: cdepubs@ssc.wisc.edu