Psychological Methods
2006, Vol. 11, No. 1, 87–105

1082-989X/06/$12.00 DOI: 10.1037/1082-989X.11.1.87

Paper or Plastic? Data Equivalence in Paper and Electronic Diaries

Amie S. Green
New York University

Eshkol Rafaeli
Barnard College, Columbia University

Niall Bolger and Patrick E. Shrout
New York University

Harry T. Reis
University of Rochester

Concern has been raised about the lack of participant compliance in diary studies that use paper-and-pencil as opposed to electronic formats. Three studies explored the magnitude of compliance problems and their effects on data quality. Study 1 used random signals to elicit diary reports and found close matches to self-reported completion times, matches that could not plausibly have been fabricated. Studies 2 and 3 examined the psychometric and statistical equivalence of data obtained with paper versus electronic formats. With minor exceptions, both methods yielded data that were equivalent psychometrically and in patterns of findings. These results serve to at least partially mollify concern about the validity of paper diary methods.

Keywords: diary studies, experience sampling method, ecological momentary assessment, multilevel models

Diary methods, once used primarily in health-related fields, have become increasingly popular in social, personality, developmental, and clinical psychology. The strengths of these methods are clear: They allow sensations, thoughts, and emotions in daily life to be monitored and reported with little retrospection, and they are relatively unobtrusive in individuals' natural settings (see Bolger, Davis, & Rafaeli, 2003, for a general review of diary methodology). However, both users and critics of diary methods have raised questions about the extent to which participants comply with researchers' instructions, particularly with regard to the timing of diary reports (Bolger et al., 2003; Reis & Gable, 2000; Stone, Shiffman, Schwartz, Broderick, & Hufford, 2002). When participants complete diary forms later than required by protocol, they may rely on retrospection, and such retrospection could reintroduce the cognitive biases in self-reporting that diaries were initially designed to circumvent. Authors addressing these topics have warned that participants often forget to complete entries, and that they might at times deliberately falsify reports of when they completed the entries.

Recent technological advancements have enabled researchers to ascertain exactly when participants complete their diaries. Electronic devices such as palm-top computers and Web-based surveys can be used to administer questionnaires and to record automatically the date and time of completion. Some devices provide additional features, such as alarm prompts and records of response times to individual items. Together, these new data-collection features bring many benefits: They eliminate the need for costly and error-prone data entry, help remind participants of scheduled times for responses, and can prompt participants at fixed or random intervals with signals for responses. Most important, they provide a means for verifying compliance independently of participants' reports.

Several diary studies have used such electronic devices to assess compliance rates (Broderick, Schwartz, Shiffman,

Amie S. Green, Niall Bolger, and Patrick E. Shrout, Department of Psychology, New York University; Eshkol Rafaeli, Department of Psychology, Barnard College, Columbia University; Harry T. Reis, Department of Psychology, University of Rochester.

Niall Bolger is now at the Department of Psychology, Columbia University.

Study 1 was assisted by the Fulbright Foundation and the Institute for Psychosocial and Socioecological Research, Rijksuniversiteit Maastricht, the Netherlands. Study 2 was supported in part by a dissertation year fellowship and grant from Northwestern University to Eshkol Rafaeli. Study 3 was supported by Grant MH60355 from the National Institute for Mental Health. We thank the members of the New York University Couples Lab for their helpful comments on a draft of this article. We thank Philippe Delespaul and Sandra Gransier for their invaluable assistance with Study 1.

Correspondence concerning this article should be addressed to Amie S. Green, Department of Psychology, New York University, 6 Washington Place, New York, NY 10003, or to Eshkol Rafaeli, Department of Psychology, 415 Milbank Hall, Barnard College, Columbia University, 3009 Broadway, New York, NY 10027. E-mail: asg228@nyu.edu or erafaeli@barnard.columbia.edu

Hufford, & Stone, 2003; Hank & Schwenkmezger, 1996; Hyland, Kenyon, Allen, & Howarth, 1993; Stone & Shiffman, 2002). Using time stamps, these studies investigated several patterns of poor compliance. The simplest pattern is when participants forget to complete entries or miss some items within an entry. Often, participants attempt to make up for these by completing the forgotten or missed entries when they are next completing the diary. Sometimes referred to as *hoarding* or *backfilling,* this phenomenon occurs, for example, when participants in a daily diary study complete 3 days' entries in a row after missing 2 of the preceding days. The frequency with which this occurs is unknown, though in one report (Gable, Reis, & Elliot, 2000, Study 2), two thirds of the respondents were found to have completed at least two diaries simultaneously. Thus, diary hoarding may be quite common.

An alarming recent report by Stone et al. (2002) provided further data on this issue. In this study, which used a time-based diary design, participants were asked to complete three surveys each day (at 10:00 a.m., 4:00 p.m., and 8:00 p.m.). Some did this with an electronic diary, whereas others completed entries on paper, unaware that their diary logbook was equipped with a small light-sensitive chip that logged the openings and closings of the logbook, thereby allowing the researchers to determine when entries could, or could not, have been made. Defining compliance as the completion of entries within a window of 15 min before or after the targeted time, the authors found 94% of the electronic responses to be compliant, compared with only 11% of the paper entries. Furthermore, although 89% of the paper responses were deemed noncompliant, few instances of noncompliance were acknowledged by the participants (in fact, 90% of the paper diaries were reported to be compliant by the participants).

A recent and more detailed report of the same study (Stone, Shiffman, Schwartz, Broderick, & Hufford, 2003) provided insight into why the difference between the paper and electronic responses reported by Stone et al. (2002) was so large. The authors designed the electronic condition to "embody best practice in the use of electronic diaries" (p. 188), but they designed the paper condition to "embody typical practice" (p. 188). Specifically, electronic responders were (a) prompted with audible alarms at the beginning, middle, and end of the diary window; (b) told that compliance would be assessed by the device; (c) given feedback on their compliance on the basis of weekly uploaded electronic data; and (d) not allowed to provide entries outside of the time window. None of these features were present for those responding on paper. Furthermore, in weekly lab visits during the 3-week study, participants were given explicit feedback on compliance, including encouragement or prompts to improve. For the electronic responders, this feedback was based on actual compliance recorded by the device. In the paper condition, participants were given feedback only with regard to the times they had self-reported on the diaries—at no time were they made aware that actual compliance was being monitored. Because feedback was based on self-reported times for paper diary participants, this may have created pressure on participants to claim timeliness even if this was not the case. Thus, the actual compliance difference between the electronic and paper conditions was confounded with differential awareness of being monitored, differential feedback about actual compliance, and likely differences in participant motivation. The performance literature is replete with studies showing that monitoring, accountability, feedback, and motivation influence performance (Greenberg & Baron, 1996). Thus, it seems premature to draw conclusions about mode-of-assessment differences until the possible impact of these confounds has been assessed.

Nonetheless, the research of Stone et al. (2002, 2003) does raise important questions about the validity of data gathered using paper diary methods. Foremost among these, in our view, is whether and to what extent there are differences between paper diary and electronic diary data when both are collected using comparable procedures. If comparable procedures were used, would compliant responses in paper diaries be as infrequent as has been documented? We have noted that Stone et al.'s (2003) study used procedures that were not equivalent across diary conditions. In addition, we note that these studies are based on only one possible diary design. Other designs, such as randomized signal-contingent designs, may be less susceptible to diary hoarding or other forms of noncompliance. It is essential that more investigations of this issue take place before drawing firm conclusions about participant compliance with paper diaries. We present results of three studies that add information about the impact of data-collection methods in diary research.

In Study 1, we investigated compliance by conducting secondary analyses of data gathered with paper diaries in a randomized signal-contingent study with a protocol that made hoarding detectable (Delespaul, Reis, & DeVries, 2004). Participants were asked to provide many reports each day on a randomized schedule, and it was possible to match the time of the signal to the reported time of completion. Compliance could be assessed in this way because, with many randomized reports on any given day, it was implausible that participants could recreate correct completion times retrospectively. Many studies with such reporting schedules exist in the literature, yet a discussion of compliance in these designs has been mostly overlooked.

Not all longitudinal research questions are best answered by randomized signal-contingent designs, however. Often it is desirable for responses to occur at fixed intervals, and with paper diary methods, comparing signal times with reported completion times cannot be used to verify compliance. Therefore, a different way to check the validity of paper diary reports is to compare paper diary reports with equivalent electronic diary reports that are known to be collected at the correct time. If reports are delayed and/or fabricated, and if the biases resulting from this noncompliance are substantial, then the data sets based on paper diaries should be noticeably dissimilar to those collected electronically in terms of central tendency, variability, and patterns of association among measures. In contrast, if the data obtained using each mode are statistically indistinguishable and if we know that the data obtained electronically are compliant, then it seems reasonable to conclude that paper diaries can still be useful.

This issue of data equivalence has not been sufficiently addressed, although one investigation conducted early in the life of electronic data capture found no differences in mean scores or dispersion measures across different modes of data collection (Hank & Schwenkmezger, 1996). To directly and more thoroughly test data equivalence between paper diaries and current electronic modes of data collection, we examined the results of another study and completed a study with fixed-interval designs.

For Study 2, we examined the equivalence of two sets of diary data: one set from a sample of participants completing a pencil-and-paper diary study and a second set from a sample of participants completing the same study but using Palm Pilots with a time-stamp feature. As in Study 1, our analysis is secondary to the original goals of the study (Rafaeli, Rogers, & Revelle, 2005). The comparison does not establish compliance rates, but rather is informative about the importance of monitoring compliance electronically in a specific diary design. In Study 2, responses were requested from both samples of participants every 3 hr while they were awake for 1 week. In examining equivalence, we compared indices of central tendency, variability, and association among variables across both modes.

Finally, Study 3 was designed explicitly to test these two notions of compliance and data equivalence using procedures that embody best practice of diary studies in both the paper and the electronic conditions. Here, participants were requested to complete one lengthy diary entry before bedtime for 2 weeks—1 week with paper diaries and 1 week with electronic diaries. Because of the design of this study, the question of data equivalence could be examined in two ways: One is to use all of the data, that is, without excluding any observations. This examination is of interest because it replicates most existing diary studies, in which no data were filtered out. A second way to examine equivalence is to compare (unfiltered) paper diaries to that subset of electronic diary data that is confirmed to be compliant. This comparison is informative because it contrasts the results that would be obtained using the traditional paper methods with those that make full use of electronic innovations. Taken together, these studies provide information about the impact of different definitions of compliance and the use of varying research protocols on the overall quality of resulting data sets based on electronic and paper diaries when both formats are implemented to provide the best possible data.

Study 1

As part of a study designed to investigate the self-assessment of social interactions (Delespaul et al., 2004), participants at a medium-sized public university in the Netherlands simultaneously completed time-based and event-based self-reports for 1 week. For our purposes, only the time-based reports are of interest. The procedures of the study were designed to maximize participant motivation and responsibility without explicit monitoring.

Methods

Forty-two students (23 women, 19 men; mean age = 20.9 years) were recruited by newspaper advertisements for a study of daily life experience. They were paid about $15 for their participation. Participants were provided with a portable, preprogrammed digital watch (SEIKO RC-1000) for 7 days. Each time the watch delivered a signal, participants were required to complete a structured diary record. For our analysis, we were simply concerned with the self-reported time at the bottom of the diary record form.

A fixed block of 10 preprogrammed signals per day was used. The signals were programmed to beep between 7:30 a.m. and 10:30 p.m., resulting in 10 intervals of 90 min. Within each interval, the timing of signals was random, subject to the constraint that no 2 signals could occur within 15 min of each other. A different random schedule was used for each day. Participants were permitted to turn the beeper off when sleeping and when they did not wish to be disturbed. At the bottom of the response sheet, participants were asked to record the exact time in the following format: "It is now exactly __ hours and ___ minutes." We estimate that the diary record would have taken about 2–4 min to complete.

In small group meetings, participants were informed that timeliness was important so that the study could determine how students distributed their time across different activities. Additionally, each participant met personally with a research assistant who went over the research protocol and emphasized the participant's collaborative role in the research process. The research assistant took care to build rapport with participants, who were invited to contact her if problems arose. Participants were not told that compliance would be scrutinized, nor was there any mention of possible penalties for noncompliance.

For each of the possible 70 requested responses for each participant, we compared the computer-logged time that the signal was sent and the participants' report of the current time at the bottom of their survey.

Results and Discussion

Compliance

Each of the 42 participants could respond to 70 beeps (10 beeps each day for 7 days) for a total of 2,940 responses. On average, participants completed 46.8 valid responses each week, which was 66.4% of the possible responses. Nearly all of the invalid responses were blank, and many of these were missed while participants were still sleeping in the morning. When the first two beeps of the morning (both before 10:30 a.m.) were eliminated, the valid response rate was 75.3% (1,771 out of 2,352 signals).

Table 1 reports the discrepancy between the actual signal time and the time noted at the end of the report sheet. For the purposes of the present research, we used a window of 5 min before to 15 min after the signal as an indicator of a timely response. (The asymmetry allows for rounding error and for time between the occasion of the signal and the completion of the report.) Using this criterion, 9.9% of responses fell outside of this window. For all but 2 participants, the median response delay was between 0 and 5 min (for these 2, it was 9 and 11.5 min).

Another way to examine compliance is on a per-participant basis (inasmuch as it would not be unusual for a small percentage of individuals to be noncompliant). The stem-and-leaf display in Table 2 shows that the percentage of responses falling within the window was 86% or higher for 37 out of 42 cases. If one excludes the 3 extreme cases with compliance under 40% as outliers (as would seem a prudent step in any diary protocol), the percentage of out-of-window responses in the entire sample is only 4.4%.

Using simple procedures to maximize participants' motivation to provide useful data and a diary design with built-in safeguards against diary hoarding, Study 1 showed that only a very small percentage of all completed surveys had times listed by participants that did not reasonably match the recorded times that the signals were known to have been sent. This participation rate was obtained without using sophisticated analyses or technological devices.

Summary

The results of Study 1 suggest that participant compliance need not be as bad as critics of paper diaries have stated. When participant motivation is high, when researcher–participant rapport is good, and when numerous randomized reports are required each day, participant self-reports of time of completion suggest very low rates of faked compliance, such as hoarding. In other words, these results suggest that compliance with the relatively taxing protocol that many diary studies demand may be more a function of the conditions under which a study is presented and run, and hence the participants' motivation and ability to comply, rather than the format of data collection. As is well known across both survey and experimental research, poorly motivated participants are likely to undermine accurate data collection, whereas well-motivated participants are more likely to comply with data-collection instructions.

Table 1
Distribution of Discrepancy Between Actual Signal Time and Time Noted in Study 1

	Responses	
Time discrepancy	No.	%
More than 10 min early	82	4.2
6–10 min early	15	0.8
1–5 min early	375	19.1
0–5 min late	1,138	57.9
6–10 min late	201	10.2
11–15 min late	58	3.0
16–20 min late	28	1.4
More than 20 min late	69	3.5

Table 2
Stem-and-Leaf Display of Compliance on a Per-Participant Basis for Study 1

Stem	Leaf
100	000000000000
90	8888887766666555411000
80	8764
70	
60	8
50	
40	0
30	
20	40

Note. Leaf values correspond to the units digit of each participant's compliance score for the given stem (e.g., there were 6 individuals with a compliance score of 98%).

Because we gave no incentives to participants for actually completing the reports on time, but instead impressed on them the value of telling us accurately what they were doing and when they recorded it, it seems unlikely that participants would have noted the time of the signal but completed the record later in the day. Furthermore, the wording and the placement of the time question reinforced the notion that we were interested in the actual time of completion, not the time of the watch signal, again reducing the likelihood that compliance would be faked. However, because Study 1 involved secondary analysis of existing data, it was not possible to incorporate actual measurements of the times participants completed the written records. Although this is one limitation of the study, it is offset somewhat by the fact that these data are typical of diary results in research that has a primarily substantive, rather than methodological, focus. That is, when substantive questions involve processes distributed within days, randomized report designs are more likely to be used (Reis & Gable, 2000).

Obviously, however, there are times when numerous randomized reports throughout the day are not desired or appropriate. In these cases, relying on self-reported time of completion is not necessarily going to provide accurate data, as some participants will reintroduce retrospective biases without properly alerting researchers to those occasions in which they have reconstructed responses. To what extent might these reports bias the data that are gathered, compared with data that conform to proper schedules of completion?

Study 2 addresses this question through secondary analysis of data from a time-based diary design in which participants were asked to respond every 3 hr while awake, rather than at random intervals. Though exact compliance rates cannot be determined because of the specific instructions given to participants at the time the study was conducted, these data allow a unique comparison of paper and electronic modes of data collection, providing additional information regarding the quality of data collected.

Study 2

As part of a study designed to investigate the structure and circadian rhythms of affect within persons (Rafaeli et al., 2005), undergraduate students at a medium-sized Midwestern university were asked to complete several brief diary entries each day for 1 week. One sample of participants completed paper-and-pencil diaries, while another sample of participants completed electronic diaries. Participants were instructed to complete these throughout the day, beginning with one entry after waking, and continuing with additional ones every 3 hr while awake. Although differing in some respects (outlined below), these two samples provided an opportunity to examine data equivalence between two sets of participants that were both given similar instructions and incentives. Because the study was designed to investigate affective fluctuations and rhythms, the issue of participant compliance was not made any more or less salient in either of the two conditions.

If the mode of data collection had an important effect on data quality, then we would expect to find that the distributions of responses and the patterns of bivariate associations between variables were notably different. However, if the paper-and-pencil approach did not introduce compliance bias and prevarication, then we would expect to find that the distributions and patterns of bivariate associations across the modes were similar. The goal of Study 2 was to examine the level of equivalence so that the relative weaknesses of the paper-and-pencil method could be assessed. In these analyses, the null hypothesis of equivalence was as much of interest as the alternative hypothesis that the two modes of data collection differed. For this reason, we present estimates of confidence intervals on the differences so that the amount of information provided by these data can be evaluated objectively.

Methods

Participants and Procedure

Paper-and-pencil mode (Sample P). Sixty-two introductory psychology students (38 women, 24 men) ranging in age from 17 to 20 years ($M = 18.3$ years) completed the study for course credit. In an initial lab session, conducted with 1 participant at a time, they completed background questionnaires (not discussed here) and were given paper diary packets. Each page in the packet served as a complete diary entry and included visual analog items. The visual analog items were 10-cm lines anchored by the labels *very little* and *very much* at either end. Participants were asked to intersect the line at a place along the continuum representing their mood at the present moment. The length of the line was divided into 10 equal sections of 1 cm, which were used to convert the location of the intersecting line into a score of 0 to 9.

Participants were asked to complete the first diary sheet of each day after waking up and to carry the diary packets with them at all times so they could continue completing diaries every 3 hr or so while awake. They were instructed to write down the date and time of completion on each diary sheet.

Completed diaries were returned every 2 days, either through campus mail or through a collection box in the students' classroom. One participant did not complete any diaries and was therefore excluded from all analyses.

Electronic mode (Sample E). Ninety-six introductory psychology students (59 women, 37 men) ranging in age from 17 to 21 years ($M = 18.6$ years) completed the study for course credit. In an initial lab session, conducted with 1 participant at a time, they first completed background questionnaires (not discussed here). Participants were then provided with Palm Pilot III devices, programmed with PMC-Diary (Rafaeli & Revelle, 1999), a dedicated program for administering electronic diaries. They were trained in using the program, in which items are presented on the screen of the device one at a time and are rated on a 10-point scale, ranging from 0 (*not at all*) to 9 (*very much*). PMC-Diary includes an alarm that prompts participants for a new entry at specified intervals (in this case, 3 hr after completion of the previous entry). However, participants could initiate diary entries at any time, save for a 10-min lock-out period following each completed diary entry. Participants could also set the device to a sleep setting, thereby determining the hour in which the next day's alarms would resume.

Participants were asked to complete the first diary sheet of each day after waking up, and to carry the device with them at all times so they could continue completing diaries every 3 hr while awake. At the end of the day, they were instructed to complete a final entry and to set the device to a sleep setting until a requested time the next day.

Diaries were automatically date and time stamped. Participants returned the devices at the end of the week. Nine participants did not complete any diaries over the week, usually because of device malfunction, and were therefore excluded from all analyses.

Incentives, rapport, and instructions in the two modes. The incentives in both samples were quite limited—only course credit was given—and participants were given this credit regardless of how many diary entries they completed. In place of extrinsic motivation, we attempted to engender a sense of commitment and partnership in the study by giving user-friendly instructions and by maintaining, for each participant, personal contact with one research assistant. To create rapport, the same research assistant who trained a participant remained in contact with him or her during the week of the study, initiating at least two phone calls or e-mails (on the 1st and the 4th days) and fielding any calls or e-mails initiated by the participant.

In both samples, the participants were told to try and maintain a schedule of entries every 3 hr or so while awake. In the electronic sample, this was aided by the program, which beeped at a 3-hr interval after the completion of an entry. However, participants were encouraged to also complete entries at shorter intervals, because the primary goal of this study was the collection of many mood samples at different times of day.

During the training session for the paper-and-pencil mode, the following instructions were given:

> Naturally, we want to get as many completed ratings as possible. But it's extremely important for us that whatever responses we get from you are accurate; accuracy is much more important than sheer number of answers. In a second we'll talk about ways to remind yourself about filling out these questionnaires, but first, I want to emphasize the importance of truthfulness. You will get the credits for this study even if whole days of sheets are missing. It is certainly reasonable to miss some ratings once in a while. What we want to avoid though is filling out questionnaires and writing down erroneous times. Let's say you forgot to fill out the sheets most of a particular day; you might have the urge to just sit down and fill out 4–5 of them at once, retrospectively. Please don't. Just do the best job in completing the questionnaires, and mark the time and day accurately. If you miss some, try harder to remember the subsequent ones, but don't fill them out in retrospect.

During the training session for the electronic mode, the following instructions were given:

> Naturally, we want to get as many completed entries as possible. But it's extremely important for us that whatever responses we get from you are accurate; accuracy is much more important than sheer number of answers. . . . You will get the credits for this study even if some entries are missing. It is certainly reasonable to miss some entries once in a while. What we want to avoid though is filling out the diary without paying attention. Just do the best job you can in filling out the diary. If you miss some entries that's OK: just do the ones you do truthfully, and do as many as you can.

Measures

Positive and negative affect (PA and NA). The 18 mood items were identical to the 16 used by Feldman (1995), with the addition of two items (tense and energetic). On the basis of the results of factor analyses (Rafaeli & Revelle, in press), two scales were constructed: the PA scale (aroused, energetic, peppy, enthusiastic, happy, satisfied, quiet, sleepy, and sluggish, with the last 3 items reverse scored) and the NA scale (nervous, afraid, tense, sad, disappointed, surprised, calm, relaxed, and still, with the last 3 items reverse scored).

Similarity and differences between the two modes. Questions were presented in the same order in both the electronic and paper diaries, and the format of each question was kept as similar as possible. However, the paper diaries presented all items on one sheet and used a visual-analogue scale, whereas the PMC-Diary program was able to present only one question per screen and had a fixed response format ranging from 0 to 9. Additionally, the anchors for the low end of the scale differed between the samples; on the paper questionnaires, the anchor was *very little,* whereas in the electronic diary, it was changed to *not at all.* This change was undertaken to ensure that the scales were interpreted as strictly unipolar, in response to a study by Russell and Carroll (1999), which appeared during the administration of Sample P but before that of Sample E.

Statistical Analysis Issues

The samples corresponding to administration mode were independent, and tests of mode effects were obtained using between-person methods. Most of the comparisons of interest involved an initial step of summarizing responses for each person over time. The means of these distributions were then compared, using either t tests or comparable tests obtained in the context of multilevel models. The validity of these tests is supported by recognition that the central-limit theorem applies to the means and the differences of means. These methods readily allowed the computation of 95% confidence intervals on the differences. The intervals are especially important in cases when we did not find a significant difference between administration modes. They describe the ranges of possible differences in the means that are consistent with the data.

Results and Discussion

Response and Compliance Rates

The total entries numbered 1,428 in Sample P and 2,309 in Sample E. Out of an expected 5 or 6 diary entries per day (presuming that the participant was awake for 15 hr each day), participants in Sample P completed an average of 23.38 diaries during the week (SD = 9.33, range = 4–48), while participants in Sample E completed a slightly higher average of 26.53 diaries over the week (SD = 8.12, range = 6–44), $t(146) = 2.17$, $p < .05$. The 95% confidence bound on this difference is 0.28 to 6.02. This bound both excludes zero (i.e., the difference is statistically significant at the .05 level) and suggests that the data are consistent with a paper-versus-electronic difference as large as 6 excess electronic diaries over the course of the week.

When we considered only those participants who completed at least 3 diary entries a day (corresponding to morning, afternoon, and evening assessments and leading to a total of at least 21 entries), the weekly averages increased to 29.19 for paper and 30.09 for electronic diaries. These latter averages did not differ, $t(104) = 0.68$, but the restriction retained only 62% of Sample P and 78% of Sample E. It is clear that the difference in average response rates in the unrestricted data was due to a substantial subgroup of paper diary respondents who provided a small number of responses. In the following analyses, all available data were included.

Participants were asked to try to respond every 3 hr while awake, but the design (and instructions) allowed the participants to respond at any time. In fact, participants were encouraged to complete entries even if the elapsed time differed from 3 hr. We checked to see if this flexibility had different effects by administration mode. The distribution of elapsed times (see Table 3) reveals a striking similarity between the two conditions. For example, if we define compliance as entries that were completed within 2 hr of the planned time (i.e., within 1 and 5 hr of the previous entry), and if we include the first entry of each morning as compliant by definition, both conditions demonstrated high, and comparable, compliance. In Sample P, 86.5% of 1,428 reported compliance, while in Sample E, 86.1% of 2,309 were compliant. We computed each person's compliance rate and compared rates across samples. In Sample P, the mean compliance was 84.4%, and in Sample E it was 84.8%. The confidence bounds on the difference were −3.7 to 2.9. Note that the compliance rate for Sample P was based on participants' reports, while the rate for Sample E was based on the devices' time stamps. Fewer than 1% of the total entries in either condition were completed within the first hour following the previous entry.

Data Equivalence

Before conducting any further analyses, we examined whether the internal consistency of the affect reports over time differed from one condition to the other. We were interested in variation over occasions, and so we computed separate alphas for each person over time. There were no apparent differences in internal consistency across modes. For PA, alpha averaged .84 for paper diaries and .82 for electronic diaries, $t(146) = 1.10$, *ns*. The 95% confidence interval around the difference of these mean alphas was −0.015 to 0.053. For NA, the average alpha was .68 for paper diaries and .72 for electronic diaries, $t(146) = -1.54$, *ns*, and the confidence interval around the difference was −0.111 to 0.014.

Data gathered in both samples were analyzed using multilevel nested models to determine whether the results obtained in the two conditions differed (Raudenbush & Bryk, 2002). These models allowed the repeated diary reports for each participant to be summarized so that comparisons of the paper and electronic samples that involve between-persons information can be carried out. Analyses were run on the PA and NA scales. We were interested in comparing, across the two conditions, the scales' means, variances, and correlation. In addition, we were interested in evaluating whether electronic data in which compliance had been confirmed differed from paper data in which compliance was assessed merely by self-report.

Means. To compare means across samples, we constructed a simple means-as-outcomes mixed model (Raudenbush & Bryk, 2002). The model was estimated using PROC MIXED in SAS, with restricted maximum likelihood estimation. The intercept was entered as a random effect, allowing individual differences around the group means. Sample was entered as a fixed effect.

For both NA and PA, the mean levels did not differ from the paper to the electronic samples: For NA, the mean was 2.59 for paper and 2.60 for electronic samples, $t(146) = 0.04$, *ns*. For PA, the mean was 3.91 for paper and 4.02 for electronic samples, $t(146) = 0.77$, *ns*. The confidence bounds for these two differences were −0.48 to 0.50 and −0.17 to 0.39, respectively. Given that both measures were on a 0 to 9 scale, we interpret the size of the upper and lower bounds to be rather small. In other words, the data are inconsistent with the proposition that mean differences exceed more than half a point on the 9-point scale. We

Table 3
Distribution of Lag Times Between Diary Entries for Both Samples in Study 2

	Sample P		Sample E		Total	
Elapsed time	Freq.	%	Freq.	%	Freq.	%
First entry	61	4.27	87	3.77	148	3.96
0–1 hr	5	0.35	16	0.69	21	0.56
1–2 hr	19	1.33	53	2.30	72	1.93
2–3 hr	159	11.13	281	12.17	440	11.77
3–4 hr	464	32.49	734	31.79	1,198	32.06
4–5 hr	157	10.99	153	6.63	310	8.30
5–6 hr	62	4.34	79	3.42	141	3.77
6+ hr	182	12.75	388	16.80	570	15.25
Overnight	309	21.64	504	21.83	813	21.75
Longer	10	0.70	14	0.61	24	0.64
Total	1,428	100.00	2,309	100.00	3,737	100.00

Note. Sample P recorded diary entries with paper and pencil; Sample E recorded diary entries electronically. Freq. = frequency.

repeated these analyses using only those data that were deemed compliant using the criteria listed above. The results based on the more strictly compliant data were very similar and showed no difference between the two samples.

Variances. We next explored potential differences in the scales' variability in each of the samples. Using multilevel models, it is possible to test for differences in two kinds of variability: within-person variance (Level 1) and between-person variance (Level 2; for a review, see Bolger et al., 2003). Level 1 variance reflects the degree to which individuals' responses over time vary around their own mean. Would individuals completing paper diaries vary more or less around their own mean than individuals completing electronic diaries? To explore this question, we estimated two separate models and compared these models' goodness of fit. The first model sought to estimate *separate* within-person error structures for each condition (Model S); the second model *constrained* the error structure to be identical across conditions (Model C).

Two approaches were used to determine which model provided a better fit to the data. The first is a deviance index that compares log-likelihood statistics for the two specified models (–2LL). As discussed by Singer and Willett (2003), the difference between the –2LL statistics of two competing models can be tested formally, because it is distributed as a chi-square distribution with degrees of freedom equal to the number of independent constraints imposed. Because Model C has one constraint relative to Model S, we compared the difference of –2LL statistics to a chi-square distribution with 1 degree of freedom. In addition to the likelihood ratio significance test, we computed an approximate 95% confidence interval on the differences in the variances estimated in Model S. When these intervals included zero, we concluded that the more parsimonious Model C is preferred.

For NA, the estimates of within-person variances were 1.61 for the paper and 1.09 for the electronic samples. For PA, these estimates were 2.92 and 1.66, respectively. In both cases, Model S appeared to provide a better fit to the data, suggesting that within-person variability does indeed differ depending on reporting mode. For NA, the Model S –2LL was 11,873.5; the Model C –2LL was 11,939.4; $\chi^2(1) = 65.9$, $p < .0001$. For PA, the Model S –2LL was 13,612.6; the Model C –2LL was 13,748.8; $\chi^2(1) = 136.2$, $p < .0001$. The confidence bounds for both differences also excluded zero: 0.38 to 0.66 and 1.01 to 1.49, respectively. We repeated these analyses with the strictly compliant data set, obtaining an identical pattern of results.

Level 2 variance reflects the degree to which individuals' mean scores vary around the grand mean for the sample, an indication of between-person variability. Would the individual mean scores obtained by participants completing paper diaries vary more or less around the grand mean than the individual mean scores obtained by participants completing electronic diaries? To explore this question, we carried out analyses of two separate models (again called S and C), comparing the degree of fit of each model. One model allowed for separate error structures based on a grouping factor of sample, while the second model constrained the error structure to be identical across the two samples.

In this case, Model C appeared to provide a better fit to the data. The paper and electronic modes produced similar levels of between-person variability around the grand mean for both NA and PA. For NA, the estimates of between-person variances were 0.54 for the paper and 0.72 for the electronic samples in Model S, and the Model C estimate was 0.66. For PA, the Model S estimates were 0.51 for the paper and 0.53 for the electronic samples, and the Model C estimate was 0.52. The fit statistics for NA were as follows: For Model S, –2LL was 11,872.3; for Model C, –2LL was 11,873.5; $\chi^2(1) = 1.2$, *ns*. The fit statistics for PA were as follows: For Model S, –2LL was 13,612.6; for Model C, –2LL was 13,612.6; $\chi^2(1) = 0.0$, *ns*. The confidence bounds on the two differences both included zero: -0.50 to 0.13 and -0.31 to 0.27, respectively. These results held when the analyses were repeated with strictly compliant data.

Correlations. Beyond investigating possible differences between samples in mean levels, and in within and between sources of variance, we also explored whether the association between variables was different in paper and electronic data-collection modes.

First, we examined interitem correlations that were expected to be strongly negative or strongly positive. Strongly positive correlations were expected of items that load in a similar direction on the same scale (e.g., high PA: energetic and enthusiastic; high NA: nervous and tense). Strongly negative correlations were expected of items that load in inverse ways on the same scale (high and low PA: energetic and sleepy; high and low NA: nervous and calm). Next, we examined correlations that were expected to be of moderate magnitude (e.g., happy and sad; see Rafaeli & Revelle, in press; Russell & Carroll, 1999) and ones that were expected to be close to null (e.g., energetic and nervous, as well the scales of PA and NA).

For each, we compared the average within-person correlations across the interview modes by taking the following steps. We first rescaled (standardized) each person's daily scores to have a mean of zero and a within-person variance of one. Then we used multilevel models to regress one standardized variable on the other in the first (within-person) level of the model. In the second (between-person) level of the model, we regressed the random slope on a dummy variable indicating whether the individual was in the paper or electronic sample. Because the within-person variables had been standardized, the slope can be interpreted as a correlation. Using PROC MIXED of SAS, we were able to obtain an appropriate standard error for the contrast of the average correlations in the two samples so that significance tests and confidence intervals could be computed. The average correlations in the two samples, the test statistic comparing the two average correlations, and the confidence interval on the difference are shown in Table 4. None of the contrasts were significant, and the confidence bound for the difference never included a value more extreme than +/–.15. Reviewing Table 4, we can see that the average correlations differed very minimally between the two conditions, consistent with the hypothesis that the modes are equivalent.

Summary

Although the results of Study 2 were mixed, we conclude that they were more consistent with the proposition that paper and electronic modes were equivalent than that they were strikingly different. The response rates (i.e., proportion of completed entries) and the level of compliance in the two conditions were mostly similar, though users of electronic diaries provided slightly more numerous responses. More important, the internal consistency, means, between-person variability, and between-variable associations were comparable.

Several differences did emerge between the two methods. First, a larger proportion of the participants in the paper condition fell short of our threshold of three entries per day. Second, a significant difference across conditions emerged in the amount of within-person variability, with within-person variance for the paper diaries being consistently greater (30% greater on average) than that for the electronic diaries. Recall, however, that the two data-collection modes differed somewhat in the question formats. At the low end of the visual analog scale, paper diaries were anchored with *very little,* whereas electronic diaries were anchored with *not at all.* As such, a response corresponding to *very little* would receive a 0 on the paper measure, but perhaps a 1 or a 2 on the electronic measure. This difference in range of possible responses may have contributed to more within-person variance in the paper mode, although this explanation seems insufficient to account for the 30% inflation for paper relative to plastic responses.

There were several limitations to this study. One is that participants were not randomly assigned a to data-collection mode, but rather were asked to use a specific mode if they were in one or another section of an undergraduate psychology class. Another limitation is that the mood scales used in the two conditions differed somewhat—one used a visual analog scale, and the other asked participants to choose a number between 0 and 9. Paper scales were anchored with *not at all* and *very much,* whereas electronic scales were anchored with *very little* and *very much*. The overall pattern of similarities despite these procedural differences strengthens our confidence in the equivalence of paper and electronic methods.

The between-persons nature of this analysis does not permit an investigation of within-person differences across mode of questionnaire. Additionally, the response rate on the electronic devices might have been elevated over that obtained on the paper diaries because of the alarm function in the PMC-Diary program; there was no such signal for participants completing paper diaries. These issues were addressed in our third study, which was designed and implemented as an experimental test of the possible differences in responses due to reporting mode.

Study 3

We addressed the limitations noted above by conducting a new study that enabled within-person and between-person comparisons, and used a different design and a different electronic diary program. We checked to see if the equivalence of the data-collection modes generalized to once-a-

Table 4
Correlations Among Measures Across Mode for Study 2

	Paper			Electronic					
Measure	*M*	*SD*	*N*	*M*	*SD*	*N*	Difference	Lower bound	Upper bound
Positive pairs									
Energetic–enthusiastic	.62	.03	61	.62	.02	87	.00	−.06	.06
Nervous–tense	.53	.04	60	.49	.03	84	−.04	−.13	.04
Negative pairs									
Energetic–sleepy	−.56	.03	61	−.55	.02	87	.02	−.06	.09
Nervous–calm	−.42	.04	60	−.39	.03	84	.04	−.05	.12
Moderate pair									
Happy–sad	−.36	.03	57	−.36	.03	86	−.00	−.09	.08
Null pairs									
Energetic–nervous	.00	.04	60	.00	.03	84	.00	−.09	.09
PA–NA	−.04	.05	61	.01	.03	87	.05	−.07	.16

Note. Confidence intervals are 95%. PA = positive affect; NA = negative affect.

day diary designs, rather than a design with three reports each day. In Study 3, participants completed end-of-day diaries for 2 weeks, 1 week in each of the two data-collection modes.

Methods

Participants and Procedure

Participants were 42 individuals (21 couples) in committed relationships, who had lived together for at least 6 months. They ranged in age from 20 to 36 years ($M = 26.5$ years). They had been romantically involved for an average of 4.9 years and had cohabited for an average of 2 years.

Couples were screened over the telephone to ensure their eligibility and to sign up for a 1-hr lab visit. After arriving at the lab, participants were given the following information about the study:

> This study explores relationships and coping styles, and is interested in the types of problems you encounter in a typical day, and how different aspects of your day affect your relationship with your partner. We use daily self-reports to obtain information about these issues. Recently, we have become interested in investigating the use of electronic diaries as opposed to paper and pencil diaries to record these self-reports.

They were then informed that they would be spending a week completing questionnaires on Palm Pilots and another week completing paper-and-pencil questionnaires. This information ensured that participants would remain unaware of our specific research question and may have helped to foster interest in the study by suggesting that they were an integral part of testing new technology in social research.

All participants then completed a comprehensive background questionnaire that took approximately 30 min. Following this session, all participants were trained by an experimenter on how to use the Palm Pilot devices. Participants were not informed of which reporting mode they would be given for the 1st week until the session had ended. This was done to ensure that all participants paid close attention to both diary methods and so that all participants would receive exactly the same instructions, regardless of condition.

At the end of the training session, participants were informed of which reporting mode they would have for the first week, and appointments were made for the following week for a lab visit to exchange materials. A randomly assigned half began the study with paper diaries, while the other half began with electronic diaries. All participants completed one daily diary each night for 6 nights. On the 7th night, all participants came to the lab and returned their paper packets or their electronic devices. Those returning paper diaries were given electronic devices for the next week, and vice versa. No training was needed at this time because training took place before random assignment. Participants continued completing daily diaries for another 6 nights. On the 7th night of the 2nd week, all participants returned their materials.

One participant was excluded from psychometric analyses because of language difficulties with some items, and 1 additional participant was excluded from these analyses because of outlying scores on all measures. However, data from both of these participants were included in the analyses on compliance rates.

Incentives, Rapport, and Instructions

Couples were compensated $75 for participation in the study, regardless of the number of questionnaires they completed. Good rapport between experimenter and participant was fostered during the training session. The experimenter carefully explained each type of question and showed the participants both the electronic versions and the paper versions to point out that they were identical. The experimenter also pointed out any questions that required a formatting change on the electronic version (i.e., checklists converted to yes–no items). This careful training session was not necessary given the user-friendly nature of the program, but was held to ensure that participants did not try to create an obscure purpose of the study and so that they were not wary of deceit. Additionally, the experimenter used this opportunity to explain how important compliance is for the researchers running the study. All participants were given the following instructions when they were shown the paper-and-pencil format:

> Each diary has the date stamped on top. It is VERY important to us that you complete the diary at the end of each day, on the correct day, rather than completing the diary the next morning and answering the questions in retrospect. However, if by some chance you do not fill out the diary on the correct day or at the correct time, please indicate on the top of that diary when you actually did fill it out.

No additional mention of compliance was made when the electronic format was demonstrated, other than to suggest that participants begin the questionnaire only when they know they will have time to finish it, as there is no way to turn the device off in the middle of a questionnaire.

Finally, experimenters called each participant the day before their second scheduled lab visit to remind them of their appointments. After arriving in the lab at the halfway point, participants were casually asked if they had any questions or if they encountered any problems. They were then given their materials for the 2nd week and made appointments to return 1 week later. Only one member of each couple was required to attend these lab visits to minimize participant burden.

Measures

Moods. Three- or four-item versions of mood scales from the Profile of Mood States (POMS; McNair, Lorr, & Droppleman, 1992) were constructed. The five scales (and their corresponding items) were as follows: Anxious Mood (anxious, on edge, uneasy), Depressed Mood (sad, hopeless, blue, discouraged), Anger (angry, resentful, annoyed), Fatigue (fatigued, worn out, exhausted), and Vigor (vigorous, cheerful, lively). Participants were asked to indicate their present mood on a 5-point scale, ranging from 1 (*not at all*) to 5 (*extremely*).

Feelings within the relationship. Two-item scales measuring feelings within the relationship were adapted from those used by Thompson and Bolger (1999). The six scales measuring relationship feelings (and their corresponding items) were Contentment (content, satisfied), Passion (excited, passionate), Anxiety (fearful, worried), Depression (sad, depressed), Anger (angry, irritated), and Love (loved, supported). Items were rated on a 5-point scale, ranging from 1 (*not at all*) to 5 (*extremely*).

Checklists. Three checklist questions were included to assess interpersonal tensions, daily troubles or hassles, and coping behaviors. For the Interpersonal Tensions Checklist, participants indicated whether they had any "tensions, disagreements or arguments with any of the following people in the past 24 hours" by checking any box that applied. Five possible responses included: partner, parent, child or children, someone from work or school, and anyone else. The Daily Troubles or Hassles Checklist consisted of 16 troublesome things "that sometimes happen to people." Participants were asked to indicate each one that happened during the past 24 hr. Items included extra work at work or school, financial problem, and sick or injured. The Coping Checklist included 16 items. Participants were asked to check any behaviors they engaged in that day in response to the most difficult or demanding aspect of their day. Coping behaviors included items such as let out my negative feelings and distracted myself. All three checklist scales were summed to form a total score for each person for each day.

Similarity and differences between the two modes. The electronic version of the diary was created using the Experience Sampling Program (ESP; Feldman-Barrett & Barrett, 2001) and was administered on Palm III and Palm 105 devices. Questions were presented in the same order as in the paper diaries, and the format of each question was kept as similar as possible. However, because the ESP program is only able to present one question per screen, some items did require modification. Specifically, checklist items that read "check all that apply" on the paper diaries became forced-choice items presented one at a time with a yes–no format. The remaining items did not differ.

Results and Discussion

Response and Compliance Rates

Data from all 42 participants were included in this analysis. We expected 252 diaries in each mode (42 participants, 6 diaries each week). In the paper mode, all 252 diaries were completed. In the electronic mode, 253 diaries were actually completed: A participant missed one entry on the first evening, while both partners within one couple completed a superfluous diary on the 7th night. These two entries were dropped.

As in the previous study, compliance was more certain when it was monitored electronically than when it was based on participants' self-reports. One estimate of compliance (and the only one available for the paper mode) was the answer to the question, "Are you completing this diary within an hour of going to bed?" Ninety-four percent of the paper diaries and 92% of the electronic diaries were reportedly compliant using this definition. However, the electronic mode provided a better index of compliance by using the time-stamp feature. We considered entries compliant if they were completed any time after 8:00 p.m. (considering 9:00 p.m. to be a reasonable bedtime) but before 5:00 a.m. on the correct night. (This rather early start time did not affect our results, as only nine of the compliant entries were actually completed earlier than 10:00 p.m.). With compliance defined in this manner, a high proportion of the electronic entries (87%) were deemed compliant. Looking at compliance per participant, 24 of the 42 participants demonstrated perfect compliance, whereas some noncompliance was evident for 18 of the 42 participants. However, 12 of these 18 individuals missed only one entry over the course of the week. Therefore, 6 of the 42 individuals in the study accounted for 66% of the noncompliance found.

With electronic time stamping, we could also detect the occurrence of sequences of diaries completed in one sitting. In our data, 10 participants (24%) completed two or more diaries simultaneously, with 2 of them completing three or more at once. It is important to note that our study explicitly instructed participants to complete any missed entry as soon as it was remembered; thus, these responses should not be considered as outright falsified data.

Data Equivalence

Unlike in Study 2, we were unable to compare the internal consistency of measures over time because of the small number of within-person observations for each mode.[1] There was not enough item variation over time, particularly for the Relationship Feelings Scales, which consisted of only two items per scale. Therefore, we computed Cronbach's alpha for each scale at the between-person level, which does not account for within-person dependency, but which is still informative about possible differences across mode.

There were no significant differences in alpha coefficients for the 11 scales measuring mood and feelings in the relationship across mode at the .05 level. Three differences (POMS Anxious Mood and Anger and Anxiety from the scales measuring relationship feelings) approached significance, but the pattern of findings did not suggest a consistent paper–electronic difference. Whereas anxious mood seemed to have a higher alpha in the electronic version ($\alpha_{\text{electronic}} = .85$, $\alpha_{\text{paper}} = .74$), the two relationship feelings scales had higher alphas in the paper version (angry: $\alpha_{\text{electronic}} = .68$, $\alpha_{\text{paper}} = .83$; anxious: $\alpha_{\text{electronic}} = .64$, $\alpha_{\text{paper}} = .82$).

We next checked the equivalence of the two modes in terms of the outcomes on means, variances, and correlations. We carried out analyses of the five subscales of the POMS, the six subscales measuring relationship feelings, and the three checklists (Interpersonal Tensions, Daily Troubles or Hassles, and Coping). We rescaled all of these dependent measures to a 0–100 scale, facilitating the comparison of effect sizes and confidence intervals across these diverse measures (P. Cohen, Cohen, Aiken, & West, 1999).

[1] The results obtained in several daily diary studies (Gleason, Bolger, & Shrout, 2003) suggest that the first daily response tends to show spurious elevation of scores. For this reason, we excluded the data for the first day of each week (i.e., Days 1 and 8) from all psychometric analyses, but not from compliance analyses.

We were interested in comparing these scales' means, their variances, and the correlations among scales across the different modes of data collection. As in Study 2, we were also interested in evaluating whether electronic data in which compliance had been confirmed differed from paper data in which compliance was assessed by self-report.

Means. To compare means across data-collection modes, we constructed a simple means-as-outcomes mixed model. The model was estimated using PROC MIXED in SAS, using restricted maximum likelihood estimation. The intercept and the mode of data collection were entered as random effects, allowing for individual differences, both in deviation from the group means and in the within-person effect of reporting mode.

In an initial analysis, we examined the order effect (paper vs. electronic mode during the 1st of the 2 weeks). Because we found no evidence for this effect, the order term was removed from all further analyses for the sake of simplicity.

Given the relatively small sample size in this study, and the expectation of no differences across condition, it is particularly important to include effect sizes and confidence bounds. The mean differences are summarized in Table 5. There were no significant mean differences across mode in any of the scales measuring mood and relationship feelings. We also present estimates of the confidence intervals for these scales. The largest absolute difference in the bounds for mood and relationship feelings was 10.12 for POMS Fatigue. Our data are consistent with the possibility that ratings of fatigue could be 10 points higher in the paper mode than in the electronic mode. On a 100-point scale, even this largest possible difference appears small. To get another perspective on the size of the effects, we converted each of the upper and lower bounds into effect size measures, by dividing them by the between-person standard deviation. The average lower bound in effect size units in Table 6 was −.14, and the average upper bound was .26. These effects are less than what J. Cohen (1988) called medium effects. For POMS Fatigue, we cannot rule out a medium effect in Cohen's terminology, but the data are inconsistent with large effects.

As mentioned previously, checklist items were somewhat changed from the paper to the electronic mode, as the program we used in this study (ESP) could only present one item at a time. Thus, we expected these questions to exhibit the largest differences across mode, because participants using the electronic version may pay more attention to each individual item. However, as can be seen in Table 5, only one of the three checklist measures, the Coping Checklist, demonstrated a significant mean difference. Confidence bounds for these scales are consistent with the possibility of an 11-point difference for the coping checklist, again on a 100-point scale. This was the largest difference found in all of the analyses and is considered by Cohen to be a medium effect. The average effect size for the three checklist items was −.24. We repeated these analyses using only those data that were deemed compliant using the criteria listed above. The results of the analyses were identical, with a significant mode effect appearing only in the Coping Checklist.

Individual differences in means. Another way to examine the equivalence of data sets resulting from paper versus electronic diaries is to consider each week's diaries as a measure of an individual's tendency to be high or low on POMS, relationship feelings, and checklist items. If the modes of data collection are equivalent, then a person's tendency to be high or low should be the same across both modes. In this case, the average values from one mode should be correlated with the average values from the other mode.

We estimated these associations using a multivariate multilevel model for each outcome scale. The model specified two random effects, one for paper mode and one for electronic mode. The covariance between the random effects can be readily estimated by the MIXED procedure of SAS, and these covariances can be converted into correlations. This analysis takes into account the number of days that are averaged when estimating the correlation.

As can be seen in Table 7, the correlation between individuals' scores across both modes of data collection was above .70 for 11 of the 14 variables of interest. To determine which of these associations were significantly different from zero, we computed confidence intervals around the covariances.[2] Two of the 14 lower bounds included zero, indicating that the association was not statistically significant. However, for the remaining variables, the upper limit on the covariance was larger than the average variance of the random effects. This suggests that for these variables, the data were consistent with a perfect correlation. From a descriptive point of view, we note that the average correlations for relationship feelings (.88) and behavioral checklists (.86) were larger than the average correlation for POMS mood measures (.59).

Variances. Although there was only one difference in scales' means across reporting mode (in the Coping Checklist), it is important to explore potential differences in the scales' Level 1 and Level 2 variability in each of the modes. At Level 1, we were interested in examining whether the variability of individuals' responses completed on paper diaries differed from the variability of their responses completed on electronic diaries. As in Study 2, we ran two separate models and compared these models' goodness of fit. Model S estimated *separate* within-person error struc-

[2] We present the confidence bounds on the covariances because they can be computed with the usual symmetric form of confidence intervals using the estimated standard error of the covariance that is available from PROC MIXED for SAS. There is no statistical result that we know of that provides an accurate 95% confidence interval for a correlation among random effects.

Table 5
Means and Confidence Intervals for Each of the Variables in the Two Modes in Study 3

	Paper		Electronic		Differences		
Measure	*M*	*SD*	*M*	*SD*	*t* value	Lower bound	Upper bound
Scale							
POMS							
Anxious Mood	14.00	18.36	10.72	17.13	1.54	−0.90	7.59
Depressed Mood	9.00	13.84	7.33	12.68	1.17	−1.10	4.38
Vigor	30.92	21.31	27.64	20.55	1.45	−1.12	7.50
Fatigue	37.96	29.94	34.86	28.84	0.89	−3.82	10.12
Anger	6.50	13.19	7.01	15.74	−0.30	−3.67	2.70
Relationship feelings							
Contentment	67.31	23.60	65.16	26.06	0.90	−2.24	6.08
Passion	45.56	28.85	44.06	27.07	0.51	−3.34	5.71
Anger	6.56	13.26	7.24	15.15	−0.48	−3.39	2.05
Depression	4.75	12.44	5.01	11.51	−0.24	−2.71	2.12
Anxiety	5.88	12.08	5.75	10.95	0.10	−2.12	2.35
Love	73.38	22.39	72.22	26.49	0.50	−2.97	5.00
Checklist							
Daily Troubles or Hassles	12.81	12.08	14.42	12.66	−1.63	−3.39	0.32
Interpersonal Tensions	10.63	15.54	12.87	18.70	−1.28	−5.73	1.20
Coping	23.97	15.76	31.65	17.86	−4.57*	−11.08	−4.43

Note. All scales were rescaled to 0–100. Confidence intervals are 95%. POMS = Profile of Mood States.
* $p < .05$.

tures for each mode; Model C *constrained* the within-person error structure to be identical across mode.

Table 8 presents the results of this analysis. Using the –2LL deviance index that compares log-likelihood statistics for the two specified models as described in Study 2, 8 of the 11 scales measuring mood and relationship feelings were better fit by Model C, the simpler model. For the 3 scales better fit by Model S (POMS Anger and the Anger and Love scales in the relationship feelings scales), the within-person variances were larger in the electronic mode. However, this was not generally the case. For 5 of the POMS and relationship feelings measures, the Level 1 variability was greater in the paper mode, whereas for 6 it was greater in the electronic mode. For the three differences that were significant, the upper bound on the difference indicates the data were consistent with roughly a doubling of the within-person variability.

In contrast to the POMS and relationship feelings scales, the checklists consistently showed differences in within-person variability across mode, with Model S providing a significantly better fit for two and a marginally better fit for the third. For all three checklists, Level 1 variance was greater on the electronic mode. Recall that for these scales, items were presented one at a time in the electronic mode, compared with a visual checklist on the paper device. The confidence bounds on the difference suggest that the electronic Level 1 variance could be as much as 2.5 times larger than the paper.

We repeated these analyses on the strictly compliant data. These results led to mostly the same conclusions for the POMS and the relationship feelings scales, although the superiority of Model S for POMS Anger and Anger and Love from the relationship feelings scales was diminished. When only strictly compliant data were included, the dominance of Model S was again diminished for the checklist scales, providing a better fit only for the Coping Checklist. The difference in the fit index for troubles and tensions, distributed on a chi-square distribution with 1 degree of freedom, were no longer significant at the .05 level: For Daily Troubles and Hassles, $\chi^2(1) = 3.0$, *ns*; for Interpersonal Tensions, $\chi^2(1) = 3.5$, *ns*. In summary, Level 1 variance differences observed for fewer than half the measures were modest in size when they existed, and were reduced (almost eliminated) when only strictly compliant data were included.

Level 2 variance reflects the degree to which individuals' mean scores varied around the grand mean, an indication of between-person variability. Again, we ran the analyses with two separate models, comparing the degree of fit for each model. There was no evidence of differences across mode in between-person variability. All of the 11 POMS and relationship feelings scales and the 3 checklists were better fit when modeled without the grouping factor of mode (i.e., by Model C). We computed confidence intervals around the differences (these data are available from the authors) and

Table 6
Effect Size Estimates for Study 3

Measure	Average effect size	Absolute value of effect size	
		Lower bound	Upper bound
Scale			
POMS			
Anxious Mood	.19	−.05	.43
Depressed Mood	.12	−.08	.33
Vigor	.15	−.05	.36
Fatigue	.11	−.13	.34
Anger	.03	−.25	.19
Relationship feelings			
Contentment	.08	−.09	.24
Passion	.04	−.12	.20
Anger	.05	−.24	.14
Depression	.02	−.23	.18
Anxiety	.01	−.18	.20
Love	.04	−.12	.20
Checklist			
Daily Troubles or Hassles	−.12	−.27	.03
Interpersonal Tensions	−.13	−.33	.07
Coping	−.46	−.66	−.26

Note. Confidence intervals are 95%. POMS = Profile of Mood States.

found some to be narrow and some to be rather wide, with no particular pattern of interest.

We repeated these analyses on the strictly compliant data. For the most part, these results led to the same conclusions, with one minor exception. The Interpersonal Tensions Checklist was significantly better fit by Model S. For Model S, –2LL was 3,129.5; for Model C, –2LL was 3,134.2; $\chi^2(1) = 4.7$, $p < .05$.

Correlations. When studying processes with diary methods, it is often of interest to study correlations among variables within person. We next asked whether the correlations obtained using paper and electronic modes were similar. In other words, do the correlations among variables differ depending on the data-collection mode?

To examine the effect of mode on within-person correlations, we chose four of the scales that are of particular theoretical interest: POMS Anxious Mood and Depressed Mood, the Interpersonal Tensions Checklist, and the Daily Troubles or Hassles Checklist. Because we were interested in within-person associations that might have differed from person to person, we used a multilevel approach for these analyses. For each of the 6 pairwise associations, we repeated the following steps: We first standardized each person's daily score on each variable within mode. Next we regressed one standardized variable on another in a multilevel model with the following features. The model omitted the intercept and the main effect of mode, because the variables were centered around zero within mode. The regression coefficient, which can be interpreted as a bivariate correlation, was considered to be random. Finally, the interaction between mode and the explanatory variable was included as a fixed effect. It was this interaction between the explanatory variable with mode that reflected the difference

Table 7
Association Among Individuals' Average Scores Across Both Modes for Study 3

Measure	Correlation	Covariance	Covariance	
			CI lower	CI upper
Scale				
POMS				
Anxious Mood	.46	42.2	−3.4	87.8
Depressed Mood	.74	34.6	8.7	60.5
Vigor	.74	107.1	33.4	180.8
Fatigue	.63	296.9	97.5	496.4
Anger	.40	9.7	−10.4	29.9
Relationship feelings				
Contentment	.91	295.2	140.3	450.1
Passion	.94	338.6	159.8	517.4
Anger	.86	32.2	7.3	57.1
Depression	.73	16.3	0.4	32.2
Anxiety	.89	36.4	13.9	58.8
Love	.97	251.8	119.4	384.2
Checklist				
Daily Troubles or Hassles	.95	65.8	21.0	110.6
Interpersonal Tensions	.83	82.4	40.9	124.0
Coping	.80	106.5	43.9	169.1

Note. Confidence intervals (CIs) are 95%. POMS = Profile of Mood States.

Table 8
Comparisons of Models With Constrained and Unconstrained Level 1 Variances, for All Scales in Study 3

			Variance		CI on difference	
Measure	−2LL difference[a]	Significance[b]	Paper	Electronic	Lower	Upper
Scale						
POMS						
Anxious Mood	0.8	*ns*	241.9	209.3	−102.1	36.8
Depressed Mood	0.6	*ns*	137.5	121.1	−56.3	23.3
Vigor	1.6	*ns*	324.9	267.4	−148.7	33.6
Fatigue	0.2	*ns*	386.8	413.0	−97.2	149.5
Anger	7.2	<.01	148.1	222.8	17.7	131.6
Relationship feelings						
Contentment	2.1	*ns*	261.1	327.9	−24.6	158.2
Passion	1.9	*ns*	471.4	380.1	−223.1	40.5
Anger	7.0	<.01	131.4	197.8	15.7	117.0
Depression	1.6	*ns*	133.0	110.2	−59.9	14.2
Anxiety	0.2	*ns*	85.8	92.0	−21.1	33.6
Love	3.7	.05	289.3	391.2	−4.6	208.4
Checklist						
Daily Troubles or Hassles	7.9	<.01	52.9	82.4	8.4	50.7
Interpersonal Tensions	3.6	.06	181.9	244.0	−4.6	128.6
Coping	21.6	<.01	98.2	205.3	58.1	156.1

Note. POMS = Profile of Mood States.
[a] −2LL difference is the difference between the deviance indices for the two specified models. [b] The difference is distributed as a chi-square distribution with one degree of freedom. Confidence intervals (CIs) are 95%.

in correlation from one mode to the other. As in Study 2, PROC MIXED of SAS provided appropriate standard errors for the contrast of the average correlations that allowed us to test the significance of the difference and to construct the 95% confidence interval for the difference.[3]

The results of these analyses are shown in Table 9. The median absolute value of the difference in the correlations was less than .06. Five of the six correlations were larger in the paper mode, but none of the differences were statistically different from zero. Although these analyses do not reveal any consistent bias associated with mode, the 95% confidence intervals around the difference remind us that small to moderate differences in correlations could exist and that these would also be consistent with the available data.

Summary

Though the overall finding in this study was one of equivalence, checklist measures proved to be somewhat more sensitive than other measures to the mode of data collection. Specifically, the Coping Checklist demonstrated a significant mean difference across mode, and all three checklists differed across mode in within-person variability. The greater sensitivity of the checklists is not surprising, given the format changes they had to undergo between the paper and the electronic versions. When checklists were presented on paper, individuals might have developed a tendency or pattern of checking off certain boxes, while not paying much attention to the entire list of possible responses. On the electronic device, however, the items on the list were presented one at a time, drawing greater attention to each. This may have resulted in more reflection on the day's experiences or in a more extensive memory search, leading to greater variability in responses.

General Discussion

Results from these three studies contribute new evidence that must be considered in the ongoing dialogue regarding the quality of the data that are obtained from paper versus electronic diary methods. In Study 1, self-reported recording times matched the random signal times sufficiently closely that it is implausible that they could have been fabricated. In Study 2 and Study 3, data sets produced by verifiably compliant electronic methods were remarkably similar to those produced by pencil-and-paper methods in terms of means, variances, and patterns of association. The mode differences we found in our samples were sufficiently

[3] Although symmetric confidence limits are not generally appropriate for correlations themselves, because they are bounded by –1 and 1, the confidence limits of differences between correlations are less affected by these bounds.

Table 9
Comparisons of Correlations Among Selected Variables Across Mode for Study 3

Variable	Average correlation	Difference by mode (*SE*)	CI difference	
			Lower	Upper
Anxiety–depression	.59	−.00 (.09)	−.19	.16
Anxiety–tensions	.22	−.06 (.12)	−.29	.17
Anxiety–troubles	.25	−.07 (.11)	−.27	.14
Depression–tensions	.33	.01 (.12)	−.26	.28
Depression–troubles	.26	−.10 (.11)	−.30	.11
Tensions–troubles	.38	−.05 (.11)	−.26	.17

Note. All correlations differ from zero ($p < .05$), but do not differ across mode. Confidence intervals (CIs) are 95%.

minor that we could conclude that any differences in the population were, at most, small. In broad terms, the results of these studies suggest that compliance is much more an issue of study design and participant motivation than it is an issue of whether a dairy is administered in paper-and-pencil form or electronically.

Rather than assuming the presence or absence of compliance in diary studies, we believe that it is important to consider when strict compliance is important and when it is not, and relatedly, when paper diaries are likely to produce valid data and when they are not. It is to these issues that we now turn.

Compliance: How Should It Be Defined and How Can It Be Increased?

Different designs call for different definitions of adequate compliance and response rates. Regardless of the design, when a spirit of collaboration and respect is established between the researchers and the participants, compliance tends to improve. We have found that explicit directions regarding the expectations of the study, coupled with an engaging and collaborative attitude toward our participants, results in greater satisfaction, adequate responses, and more good will toward the study itself.

Though any diary study includes the risk of participants intentionally faking diary entries, our follow-up results show that more often, participants who provide noncompliant responses are doing so for more benign reasons: to be good by filling out entries they had missed. Adequate communication with participants and a clear explanation of the utility or lack of utility of entries provided at the wrong time can dramatically reduce this second risk of faked compliance. Conducting studies in this way is best done when combined with compensation that is not contingent on unreasonably narrow compliance.

When Is Strict Compliance Important and When Is It Not?

In Study 1, which was designed to randomly sample self-assessments of social interactions throughout the day, the time window for a response to be considered compliant was narrow, from 5 min before to 10 min after the signal. In a study of this sort, researchers would prefer omissions to retrospectively reconstructed responses. In Study 2, designed to sample moods throughout the day for circadian rhythm analyses, lags that differed from the fixed 3-hr intervals could potentially affect the precision of estimates of the circadian patterns, but participants were aware that completion of records at any time in the 3-hr window provided usable data. Study 3 involved an end-of-day daily diary to be completed within 1 hr of bedtime, which was designed to obtain reports about the entire day. The 1-hr requirement was flexible in that entries completed 2 hr before bedtime were likely to serve the purpose of the study almost equally well. However, we regarded filling out a questionnaire the next morning and providing retrospective information about the previous day as clearly detrimental to the study's aims.

The differing goals of these three studies point to the various ways in which compliance can be defined. Applying a narrow time window of compliance will have different implications depending on the nature of the research question. Defining compliance as completion of diaries within 10 min of the specified time sets a rather strict limit, whereas defining compliance as any diary completed on the correct day regardless of time may be considered too lenient. As the examples above demonstrate, we argue that study designs that reduce participant demand and broaden the boundaries of what is considered compliant do not necessarily sacrifice the quality of the data and have the potential to dramatically increase the proportion of them that are usable. Rather than concluding that compliance is poor and paper diaries are unreliable, as was suggested by Stone et al. (2002), we argue for a more moderate position—that depending on the design of a study and the variables of interest, researchers can choose the mode of data collection that best suits both their own needs and the needs of their participants.

Taking this argument one step further, it seems plausible that Stone et al.'s (2002) participants had relatively high rates of compliance in the electronic condition precisely because they knew that the researchers were monitoring and evaluating their rates of actual compliance. Participants in the paper condition were told that they needed to be timely, but were not told that their actual timeliness was being monitored. During feedback sessions, participants in the electronic condition learned that they could not get away with lying about completion times, whereas the participants in the paper diary condition learned the opposite.

More generally, methods for ensuring participant compliance must not be so heavy-handed that they are experienced as intrusive or mistrusting (Webb, Campbell, Schwartz, & Sechrest, 1966). It is well known that scrutiny of a person's behavior may interfere with the flow of natural behavior, and procedures that call undue attention to verifying compliance may alter the very behaviors and processes they are designed to assess.

Individual Differences in Compliance

Individual differences in compliance are an important factor to consider, particularly when it comes to the use of advanced technology. Individual differences in preference may affect compliance and response rates. Highlighting the importance of this issue, a simple follow-up questionnaire administered to participants in Study 3 revealed equivalent levels of preference for electronic versus paper questionnaires (47.2% vs. 52.7%, respectively). Furthermore, when participants were asked to indicate how much they liked or disliked each mode of data collection, very few of the participants (11%) felt very positively or very negatively about paper diaries, but electronic diaries elicited stronger responses with 17% liking them very much and 19% outright disliking them. Given that almost one fifth of our participants were not happy using electronic diaries, these results carry important implications for study designs using solely electronic methods.

It is important to consider that the effects of individual preference may be magnified in certain samples, such as aging populations or samples with vision deficits or other disabilities. Seeing a personal digital assistant (PDA) screen clearly or grasping the stylus might pose real obstacles for some participants, whereas paper methods might be more flexible for implementing more suitable formats (e.g., magnified copies or larger paper). Moreover, relying solely on electronic options could create selection pressure for certain types of question formats (e.g., open-ended responses are not always possible on electronic devices, and when they are possible they can be labor intensive for both participants and for researchers) and could indirectly bias study content. Clearly, further development and research is needed before researchers should feel comfortable relying solely on electronic measures.

Suggestions for Improving Diary Studies

Two factors play an important role in the degree to which participants comply with diary study instructions: the burden created by the study, and the degree to which participant motivation is fostered. Research questions that require studies high in participant demand need not sacrifice data quality, provided that researchers create and maintain sufficient motivation among participants. However, mounting a low-burden study does not guarantee that participants will comply with instructions; having sufficient participant motivation is an essential component of all studies involving self-reports.

The studies described here varied in the degree of burden, although all attempted to maximize participant motivation. In Study 1, participants were asked to provide 10 responses each day for 7 days. They were subject to random digital signals in the midst of any activity throughout the day and therefore had to carry materials with them at all times and had to remember to wear the programmed watch. Furthermore, they had to interrupt whatever they were doing when a signal occurred. Even considering the level of demand in this study, participants completed an average of 7.5 responses per day. These results provide a good example of how high demand combined with high motivation can result in high compliance.

Study 2 also involved high demand. Participants were asked to provide an average of six entries per day for 1 week, although intervals were fixed and allowable response times were more flexible than in Study 1. Participants were again expected to carry the diary packets or electronic devices with them at all times. Despite this level of burden, overall good compliance was achieved. Perhaps reflecting the increased burden of toting a paper packet relative to a Palm Pilot, lowered responsiveness was indeed evident in the paper data-collection mode. Nonetheless, the compliance rates for both samples (i.e., the proportion of completed responses that adhered to our instructions) were similar and reasonably high.

Study 3 could be considered the least burdensome, as participants provided only a single entry each day; each entry, however, was considerably longer than any of the entries in Studies 1 or 2, and the study lasted twice as long. An additional burden was one required lab visit at the midpoint of the 2-week period. In this study, participants' partners were also involved in the research and were simultaneously completing diaries, which may have increased motivation or may have facilitated compliance if partners served as reminders for each other. Again, impressive response and compliance rates were achieved, with only small differences between the two data-collection modes. As both electronic and paper diaries were kept at home, and it was recommended that they be kept at the bedside, they did not differ in ease of use.

These three studies provide clear examples of the ways in which participant burden and participant motivation play key roles in diary studies. We believe researcher efforts to improve participant motivation and foster researcher–participant rapport can have the desired effect of improving participant compliance, even in situations where higher levels of participant burden cannot be avoided. Although monetary compensation is not always possible or desired, creating an environment in which participants truly play a participatory role in the research project, and feel invested in the outcomes of the study, will likely increase their engagement in the study. Providing feedback on their progress, encouraging them to ask questions and contact the researchers if desired, creating a sense of personal involvement in the research, and maintaining regular contact to remind participants of deadlines, are all ways to facilitate compliance. All of the studies reported here maximized participant motivation wherever possible, resulting in overall high rates of confirmed compliance in the electronic conditions and comparable data sets in the paper conditions, suggesting similar, if unconfirmed, compliance.

Summary

Because diary methods are used to address a growing range of questions, we believe that researchers should have more, rather than fewer, tools at their disposal. Our results argue for the utility of both paper and electronic diaries as viable tools. We identified several conditions under which we would expect results to be unaffected by the choice of data-collection mode. Specifically, it seems that research focused on mean levels, between-person differences, and correlations among variables will not be greatly affected by the choice of paper or electronic data-collection methods. We also identified some conditions in which results seem more sensitive to the data-collection mode. Specifically, questions pertaining to the within-person (Level 1) variance may find different answers when different data-collection methods are used and particularly when there is some difference in the response format of the items in the different modes. Given the lack of consistency in the direction of within-person variance differences across Study 2 and Study 3, further research is needed to ascertain whether the variance estimates obtained from paper diaries or from electronic diaries are more correct.

When choosing a diary format, researchers should take into account a range of considerations, including several that we did not consider in our studies. For example, studies that place a high premium on equally spaced reports are likely to benefit from features of electronic diary methods that verify the time of completion. On the other hand, studies of special populations with members who are not familiar with computers or PDA devices may find that paper-and-pencil methods produce better data. Clearly, researchers need to review the types of questions they will be presenting and decide which format, paper or electronic, will be most appropriate given their specific study.

Conclusions

Our studies suggest that it is premature to conclude that paper diaries yield data that are consistently misleading relative to data collected with electronic time stamping. However, our studies are themselves insufficient to establish the opposite claim, namely that paper diaries produce data that are equivalent to more modern methods. In our Study 1 and Study 2, we made use of data that were collected for other purposes, and hence we had to make do with design limitations of these studies as they related to the questions we posed. Study 3 was explicitly designed to address the question of data equivalence, but its sample size is too small to be conclusive about the magnitude of differences that could exist between the collection modes. Nevertheless, the pattern of results from these studies, which used different methods, different populations, and different diary designs, strongly suggests that the question of data equivalence is still very much open.

References

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology, 54,* 579–616.

Broderick, J. E., Schwartz, J. E., Shiffman, S., Hufford, M. R., & Stone, A. A. (2003). Signaling does not adequately improve diary compliance. *Annals of Behavioral Medicine, 26,* 139–148.

Cohen, J. (1988). Set correlation and contingency tables. *Applied Psychological Measurement, 12,* 425–434.

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research, 34,* 315–346.

Delespaul, P. A. E. G., Reis, H. T., & DeVries, M. W. (2004). Ecological and motivational determinants of activation: Studying compared to sports and watching TV. *Social Indicators Research, 67,* 129–143.

Feldman, L. A. (1995). Valence focus and arousal focus: Individual differences in the structure of affective experience. *Journal of Personality and Social Psychology, 69,* 153–166.

Feldman-Barrett, L., & Barrett, D. J. (2001). An introduction to computerized experience sampling in psychology. *Social Science Computer Review, 19,* 175–185.

Gable, S. L., Reis, H. T., & Elliot, A. J. (2000). Behavioral activation and inhibition in everyday life. *Journal of Personality and Social Psychology, 78,* 1135–1149.

Gleason, M. E. J., Bolger, N. P., & Shrout, P. (2003, January). *The effects of study design on reports of mood: Understanding differences between cross-sectional, panel and diary designs.* Poster session presented at the annual meeting of the Society for Personality and Social Psychology, Los Angeles.

Greenberg, J., & Baron, R. A. (1996). *Behavior in organizations* (6th ed.). Englewood Cliffs, NJ: Prentice Hall.

Hank, P., & Schwenkmezger, P. (1996). Computer-assisted versus paper-and-pencil self-monitoring: An analysis of experiential and psychometric equivalence. In J. Fahrenberg & M. Myrtek, (Eds.), *Ambulatory assessment: Computer-assisted psychologi-*

cal and psychophysiological methods in monitoring and field studies (pp. 86–99). Seattle, WA: Hogrefe & Huber.

Hyland, M. E., Kenyon, C. A. P., Allen, R., & Howarth, P. (1993). Diary keeping in asthma: Comparison of written and electronic methods. *British Medical Journal, 306,* 487–489.

McNair, D. M., Lorr, M., & Droppleman, L. F. (1992). *EDITS manual for the Profile of Mood States.* San Diego, CA: EDITS.

Rafaeli, E., & Revelle, W. (1999). *Personality, motivation, and cognition (PMC) diary program.* Retrieved September 1, 1999, from http://personality-project.org/pmc/diary.html

Rafaeli, E., & Revelle, W. (in press). A premature consensus: Are happiness and sadness truly opposite affects? *Motivation and Emotion.*

Rafaeli, E., Rogers, G. M., & Revelle, W. (2005). *Affective synchrony: Individual differences in mixed emotions.* Manuscript submitted for publication.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Reis, H. T., & Gable, S. L. (2000). Event-sampling and other methods for studying everyday experience. In T. H. Reis, & M. C. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 190–222). New York: Cambridge University Press.

Russell, J. A., & Carroll, J. M. (1999). On the bipolarity of positive and negative affect. *Psychological Bulletin, 125,* 3–30.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York: Oxford University Press.

Stone, A. A., & Shiffman, S. (2002). Capturing momentary, self-report data: A proposal for reporting guidelines. *Annals of Behavioral Medicine, 24,* 236–243.

Stone, A. A., Shiffman, S., Schwartz, J. E., Broderick, J. E., & Hufford, M. R. (2002). Patient non-compliance with paper diaries. *British Medical Journal, 324,* 1193–1194.

Stone, A. A., Shiffman, S., Schwartz, J. E., Broderick, J. E., & Hufford, M. R. (2003). Patient compliance with paper and electronic dairies. *Controlled Clinical Trials, 24,* 182–199.

Thompson, A., & Bolger, N. (1999). Emotional transmission in couples under stress. *Journal of Marriage and the Family, 61,* 38–48.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences.* Oxford, England: Rand McNally.

Received June 9, 2005
Revision received July 6, 2005
Accepted July 6, 2005 ■