Psychonomic Bulletin & Review
2006, 13 (6), 1028-1032

Improving the reliability of event-based laboratory tests of prospective memory

WILLIAM L. KELEMEN, W. BAILEY WEINBERG,
HANNAH S. ALFORD, EMILY K. MULVEY, and KEVIN F. KAEOCHINDA
California State University, Long Beach, California

Laboratory tests of event-based prospective memory (ProM) require participants to perform actions in response to infrequent cues in a background task. We conducted three experiments to assess and improve the reliability of this popular procedure. In Experiment 1, we tested college students on 2 separate days and found that the alternate-forms reliability of ProM accuracy was quite low ($r = .31$), although general knowledge accuracy was more reliable ($r = .89$). In Experiment 2, a statistically significant difference in reliability emerged between conditions with a low ($n = 6$) versus a high ($n = 30$) number of ProM targets. Finally, lower ProM accuracy increased reliability in Experiment 3. Adopting these straightforward changes may enhance the search for individual differences in ProM.

Prospective memory (ProM) involves remembering to complete intentions for future activities. A popular procedure for assessing ProM has emerged in which participants complete a computerized background task (e.g., answering general knowledge questions) with a small number of embedded ProM target items. The participants are asked to complete the background task as accurately as possible and also to respond in a particular way (e.g., press a certain key) whenever an infrequent ProM target item appears. Marsh, Hancock, and Hicks (2002) labeled this general event-based laboratory procedure the *standard Einstein–McDaniel paradigm*, after the researchers who devised it (Einstein & McDaniel, 1990).

Despite the increasing popularity of ProM research, little is known about some fundamental psychometric properties of these procedures (e.g., reliability). One potential source of concern is the small number of ProM target items included in many event-based ProM laboratory experiments. The original Einstein and McDaniel (1990) study included 3 ProM target items embedded in a series of 42 short-term memory trials. Subsequent researchers have routinely altered the nature of the background task, but they nearly always have measured ProM with a limited number of items, although some researchers have included more (e.g., Ellis, Kvavilashvili, & Milne, 1999, included 5, 10, or 20 ProM target items; Titov & Knight, 2001, used 21 ProM target items). Acceptable levels of reliability can be elusive with a small number of test questions (Cronbach, 1990), and unfortunately, most published research on ProM has not included psychometric properties of the dependent measures.

A few studies have reported the reliability of ProM tasks over time, and the results have been mixed. On the positive side, Titov and Knight (2001) obtained significant alternate-forms reliability ($r = .65$) with a novel procedure that included 21 ProM target actions during a videotaped New Zealand street scene. Conversely, Schmidt, Berg, and Deelman (2001) found low 5-week test–retest reliability ($r = .24$) for a new training program, using nine ProM target items with older adults. In apparent frustration, Schmidt et al. remarked, “In our experience, one of the main problems in the study of prospective memory is the development of reliable . . . measures. Despite repeated attempts to construct reliable prospective memory measures, we have not been able to do so” (p. 474).

The major goals of the present research were to assess the level of alternate-forms reliability in a typical laboratory test of event-based ProM and to develop ways to increase reliability if it was found wanting. As our model, we selected the procedure used in Experiment 3 in Einstein, McDaniel, Richardson, Guynn, and Cunfer (1995), which is one of the most widely cited studies in the field. We asked participants to answer general knowledge questions and to press a certain key as quickly as possible whenever a question pertained to a U.S. president (i.e., a ProM target item). As in most studies, we included a small number of ProM target items ($n = 6$) in Experiment 1. To anticipate our results, the alternate-forms reliability was found to be quite low, and so Experiments 2 and 3 were designed to improve this aspect of performance. Specifically, we hypothesized that increasing the number of ProM targets in Experiment 2 would improve reliability without altering overall levels of accuracy. In Experiment 3, we changed

Portions of the data from Experiments 1 and 2 were presented in 2003 during a poster session at the 44th Annual Meeting of the Psychonomic Society in Vancouver. Portions of the data from Experiment 3 were presented in 2005 at the 85th Annual Convention of the Western Psychological Association in Portland. The authors thank Mark McDaniel and two anonymous reviewers for their comments on a previous version of the manuscript. W.B.W. is now at Claremont Graduate University; K.F.K. is now at the University of California, Riverside. Correspondence concerning this article should be directed to W. L. Kelemen, Department of Psychology, California State University, 1250 Bellflower Boulevard, Long Beach, CA 90840-0901 (e-mail: wkelemen@csulb.edu).

the ProM target items to produce moderate levels of accuracy and, thereby, improve reliability.

EXPERIMENT 1

Method

Participants. A total of 34 students (23 women, 9 men, and 2 unreported) in an undergraduate research methods course volunteered to participate as one option for completing a course assignment. Data from 3 additional participants who failed to complete both sessions were omitted. The mean age of the students was 22.5 years ($SD = 6.4$).

Materials and Procedure. The participants were tested in a computer lab on 2 separate days (48 h apart). On Day 1, the participants were asked to sign a consent form, and then they received instructions. On each day, the participants were instructed to complete a 150-item computerized test of general knowledge as accurately as possible. In addition, the participants were instructed to press the "p" key whenever they saw a question about a U.S. president. If they neglected to press the "p" key immediately, they were asked to press it as soon as possible after they remembered.

Each general knowledge question appeared on the screen for 6 sec, along with four alternatives [e.g., *What is the unit of currency in Italy? (1) Peseta, (2) Escudo, (3) Lira, (4) Drachma*]. After 6 sec, a message appeared instructing the participants to make a selection using the keys 1–4. The participants were allowed 3 sec to enter their response, and then feedback appeared for 3 sec, indicating whether or not their selection was correct, along with the cumulative percent correct. Responses to the ProM target questions were scored as correct if they occurred anytime during the 12-sec interval.

The ProM target items were six questions about U.S. presidents, which were embedded at Positions 23, 41, 61, 95, 124, and 143 (Version A) or 25, 47, 66, 94, 112, and 140 (Version B). Thus, 4% of the questions in each version (6 out of 150) were ProM target items. Three target items included the word *president* [e.g., *Which former president is featured on the $20 bill? (1) Abraham Lincoln, (2) George Washington, (3) Thomas Jefferson, (4) Andrew Jackson*), and the remaining three target items did not (e.g., *What is Bill Clinton's middle name? (1) Jefferson, (2) Kennedy, (3) Franklin, (4) Jackson*]. The participants were informed that some of these target items would include the word *president* and others would not but that they were to press "p" in either case. Alternate versions of the test were created using different general knowledge questions and different ProM target items. The version used each day was counterbalanced across participants.

After receiving their instructions, the participants completed a 30-item test of conscientiousness for 3 min as a distractor activity. The participants then began the general knowledge test. Upon completion, they were instructed to recap the instructions on a sheet of paper, which served as a manipulation check to ensure that the participants had understood the task. Finally, the participants completed a demographic questionnaire. The procedures for the second session were identical, except that an alternate form of the test and an obsessive-compulsiveness questionnaire were used.

Results

Data from the conscientiousness and obsessive-compulsiveness questionnaires were unrelated to ProM accuracy, and so they will not be discussed further. All tests of statistical significance were conducted at $p < .05$ in Experiments 1–3.

ProM performance. ProM accuracy ranged from 0 to 6 each day, and mean proportion correct was high for both versions (see Table 1); performance did not differ significantly between Versions A and B. The participants identified significantly more ProM target items when the word *president* was included [Version A, $t(33) = 6.71$; Version B, $t(33) = 5.22$]. Thus, ProM accuracy was equivalent in both versions, and omitting the word *president* made the prospective task more difficult.

Table 1
Mean Proportions Correct for Prospective Memory Target Items and General Knowledge Questions in Experiment 1 (With Standard Errors of the Means)

	Version A		Version B	
Type of Question	*M*	*SEM*	*M*	*SEM*
All six prospective memory target items	.80	.03	.78	.04
Three targets including the word *president*	.95	.02	.88	.04
Three targets omitting the word *president*	.65	.05	.69	.04
All 150 general knowledge questions	.60	.02	.64	.02

General knowledge performance. Accuracy on the 150 general knowledge questions was moderate (see Table 1). Although questions were assigned to each version arbitrarily, the small difference in accuracy between Versions A and B was statistically significant [$t(33) = -3.21$].

Alternate-forms reliability. The most important aspect of our results concerned whether or not the participants' ProM accuracy was consistent across versions. For ProM accuracy, Pearson's correlation between Version A and Version B was quite low ($r = .31$, which was not significantly greater than 0). Including or omitting the word *president* from the ProM target items did not influence reliability ($r = .20$ and $r = .24$, respectively). In contrast, the students' performance on the general knowledge questions was highly reliable ($r = .89$, which was significantly nonzero).

Discussion

The correlation between alternate forms of a typical ProM test was low, whereas general knowledge accuracy was quite reliable. Mean ProM accuracy was high, but ceiling effects probably do not completely explain the low reliability, because omitting the word *president* from the target items significantly reduced ProM accuracy but had negligible impact on reliability. Another explanation might be the large disparity in the number of items used to compute the correlation between versions: 6 items for ProM versus 150 general knowledge questions. We directly compared ProM reliability with a low versus a high number of target items in Experiment 2.

EXPERIMENT 2

We varied the number of ProM target items (6 vs. 30) in order to test the hypothesis that reliability would be higher with a larger number of targets. This change might have produced unintended consequences, however. For example, Maylor (1998) showed that ProM accuracy in college-aged participants improved over the course of an experiment using 8 ProM targets. Therefore, we analyzed whether ProM accuracy in our 30-item condition increased due to practice. We also explored the possibility that including 30 target items changed the nature of our task, from one emphasizing ProM to one requiring only vigilance. Brandimonte, Ferrante, Feresin, and Delbello (2001) compared performance on a single task labeled as either a ProM or a vigilance test, and they found that response latencies in the ongoing "cover" task were longer and accuracy was higher in the vigilance condition. Thus, we also tested for differences in these variables in order to infer whether or not the participants treated the 6- and the 30-item conditions differently.

Method

Participants. A total of 87 students (70 women and 17 men) enrolled in either a research methods course or a cognition course volunteered and completed both testing sessions. The average age of these participants was 22.0 years ($SD = 4.1$). None of these students had participated previously.

Materials and Procedure. The methodology in the first experiment was modified in the following ways. The number of general knowledge test questions in Experiment 2 was expanded to 200. Upon arrival, the participants were randomly assigned to a condition with either 6 or 30 questions about presidents. Thus, the 6-target condition contained 3% ProM target items, whereas 15% of the questions were ProM targets in the latter condition. Five different versions of the 6-target condition were constructed, allowing all of the 30 prospective memory target items to be used across participants. These different versions were counterbalanced. All ProM target items included the word *president* in Experiment 2. Finally, the obsessive-compulsiveness questionnaire was omitted.

Results

ProM performance. To facilitate comparisons across conditions, the proportion correct for ProM accuracy is reported in Table 2. Version (A or B) was a within-subjects manipulation, and condition (6 vs. 30 target items) was a between-subjects manipulation. Mean performance tended to be higher in Version A than in Version B and also higher in the 30-target condition than in the 6-target condition. However, a 2 × 2 mixed ANOVA revealed no significant main effects (ps = .09 and .11), nor was the interaction significant ($p = .97$). Thus, the participants' ProM accuracy did not change significantly according to version or condition.

To test for practice effects in the 30-item condition, we examined ProM accuracy in blocks of 6 items (see Table 2). Accuracy did not increase monotonically across blocks in either version, as might have been expected had practice played a major role. In fact, ProM accuracy in Version A was nearly identical in the first and the last blocks (Ms = 0.89 and 0.90, respectively). The omnibus F test across blocks was significant [$F(4,172) = 2.90$], but Block 1 ProM accuracy did not differ significantly from that in any of the subsequent blocks, according to a series of t tests. For Version B, the omnibus F statistic was significant [$F(4,172) = 11.85$], and some signs of practice emerged, as suggested by three significant t tests [Block 1 vs. 2, $t(43) = -4.13$; Block 1 vs. 4, $t(43) = -2.83$; Block 1 vs. 5, $t(43) = -5.52$].

General knowledge performance. Accuracy tended to be lower for Version B of the general knowledge test (see Table 2). A 2 × 2 mixed ANOVA confirmed a statistically significant main effect of version [$F(1,85) = 137.85$]. No significant effect of condition and no interaction emerged.

Table 2
Mean Proportions Correct for Prospective Memory Target Items and General Knowledge Questions by Condition in Experiment 2 (With Standard Errors of the Means)

	Version A		Version B	
Condition	*M*	*SEM*	*M*	*SEM*
Prospective memory accuracy				
6-target condition	.82	.04	.76	.05
30-target condition	.89	.02	.83	.04
Targets 1–6	.89	.03	.76	.04
Targets 7–12	.85	.03	.88	.04
Targets 13–18	.92	.03	.79	.04
Targets 19–24	.89	.03	.84	.04
Targets 25–30	.90	.03	.90	.04
General knowledge accuracy				
6-target condition	.70	.02	.61	.02
30-target condition	.71	.02	.63	.02

If the participants viewed the 6- and 30-item conditions differently (i.e., the former reflecting ProM and the latter reflecting vigilance), we should have replicated Brandimonte et al.'s (2001) findings of improved accuracy and increased response latencies for vigilance in the background task of our 30-item condition. However, no significant difference in accuracy emerged between conditions. Mean response latencies also did not differ significantly from each other, and they were not even in the predicted direction (Ms = 993 and 1,013 msec [SDs > 190] for the 6-item condition; Ms = 974 and 977 msec [SDs > 240] for the 30-item condition). Thus, we observed no evidence that the participants had treated the two conditions differently.

Alternate-forms reliability. The participants' general knowledge performance again showed high levels of reliability ($r = .88$ in the 6-target condition and $r = .86$ in the 30-target condition), which were both significantly nonzero. For ProM accuracy in the 6-target condition, Pearson's correlation between Version A and Version B was $r = .12$, which was not significantly greater than 0. Importantly, the magnitude of the correlation between Versions A and B increased to .62 in the 30-target condition. This correlation was significantly different from 0 with 44 participants. The difference in magnitude of the correlations between the 6- and the 30-item conditions also was statistically significant, using Fisher's r to z transformation ($z = 2.70$, $p < .05$).

Discussion

Experiment 2 supported the hypothesis that increasing the number of ProM target items can improve alternate-forms reliability. Including 30 target items produced a sizable increase in the magnitude of obtained correlations, in comparison with 6 target items. This increase in reliability occurred despite high ProM accuracy and did not significantly influence accuracy. Thus, simply increasing the number of ProM target items can be sufficient to improve test reliability.

This approach can be criticized on the grounds that the processing elicited by the task changed substantially with a larger number of ProM target items. Although our experiment was not designed to test this possibility, some post hoc findings are germane. First, our data did not conform to the pattern proposed by Brandimonte et al. (2001) to differentiate between ProM and vigilance tasks, which argues against the possibility that the nature of the task changed entirely in the 30-item condition. Table 2 also shows that ProM accuracy for the first 6 target items in the 30-item condition was nearly identical to the overall level of ProM accuracy in the 6-item condition, suggesting that the higher ratio of ProM targets in the 30-item condition did not radically alter performance. We did see some evidence of practice effects in one version of the 30-item condition, but not in the other, which should have decreased, rather than increased, reliability in the 30-item condition. Finally, some researchers have argued that including large numbers of targets is acceptable so long as a reasonable time interval (e.g., at least 1 min) elapses between occurrences (Ellis et al., 1999), as was the case in our study.

Nevertheless, ProM may be especially sensitive to changes in procedure, including the number of targets. For example, McDaniel, Guynn, Einstein, and Breneiser (2004) demonstrated that even the first two ProM target items can be sensitive to different manipulations, in comparison with subsequent targets. It remains possible, then, that some subtle differences in processing were induced by the inclusion of 30 target items. Experiment 3 was designed to test an alternative method of increasing reliability without increasing the number of ProM target items. Specifically, we sought to increase the range of responses by reducing ProM accuracy.

EXPERIMENT 3

Ceiling effects are common in the standard Einstein–McDaniel paradigm (Uttl, 2005a, 2005b). For example, our study was based on Einstein et al.'s (1995) Experiment 3, which included six *president* target items. The proportions correct for three different age groups in that experiment were .93, .93, and .86. ProM accuracy also was high in our first two experiments. Because U.S. presidents may have been particularly salient for our students, we changed the ProM target items to questions containing the word *animal*.[1] Familiarity and distinctiveness of target items can influence ProM accuracy dramatically (e.g., McDaniel & Einstein, 1993), and pilot testing suggested more moderate performance for questions about animals than for questions about presidents. We hypothesized that a greater range in the scores associated with nonceiling performance would increase reliability with only 6 target items.

Method

Participants. A total of 33 students (27 women and 6 men) enrolled in a research methods course completed both testing sessions. The average age of these participants was 21.3 years ($SD = 2.7$). All the participants volunteered for testing, and none had participated previously.

Materials and Procedure. Two hundred new general knowledge questions were created for each session, which included 6 target items that contained the word *animal*. Two new questionnaires served as a filler activity between the instructions and the beginning of the task each day: a 45-item depression scale (International Personality Item Pool, 2001) and a 28-item dissociation scale. The depression scale asked the participants to rate certain personal characteristics (e.g., *I often feel blue*) on a 5-point Likert scale. The dissociation scale was unrelated to performance, and it will not be discussed further.

Results

ProM and general knowledge performance. As we hoped, mean ProM accuracy was moderate in Experiment 3 (see Table 3), suggesting that the *animal* target items may have been less salient than the previous *president* target items. A paired samples t test showed that mean ProM accuracy did not differ significantly between Versions A and B [$t(32) = -1.50$]. General knowledge accuracy also was moderate, and the difference in accuracy between Versions A and B was statistically significant [$t(32) = 2.76$].

Alternate-forms reliability. The major question again was whether or not the participants' ProM accuracy was consistent across versions. For ProM accuracy, Pearson's correlation between Versions A and B was $r = .62$. The correlation between versions of the general knowledge tests was $r = .76$. Both of these correlations were significantly greater than 0 with 33 participants.

Depression scores and ProM accuracy. Post hoc analyses were conducted on the participants' depression scores. The mean score was 1.99 ($SD = 0.51$), with higher scores indicating more severe depressive symptoms. A statistically significant negative correlation was observed between depression scores and ProM accuracy measured on the same day ($r = -.42$). For women only, the correlation remained significant ($r = -.48$), whereas for the small number of men the correlation was nonsignificant and tended to be positive ($r = .24$).

Discussion

Including animals as target items yielded a nonzero level of ProM reliability. In fact, the correlation between versions with 6 *animal* questions ($r = .62$) was identical to the correlation obtained in Experiment 2 with 30 *president* questions. Thus, it is possible to achieve statistically significant levels of reliability in the standard Einstein–McDaniel paradigm with a small number of target items. Moreover, we were able to demonstrate a negative correlation between depressive symptoms and ProM accuracy with this task. One suggestive finding was that women's ProM accuracy was negatively impacted by depression

Table 3
Mean Proportions Correct for Prospective Memory Target Items and General Knowledge Questions by Condition in Experiment 3 (With Standard Errors of the Means)

	Version A		Version B	
Type of Question	*M*	*SEM*	*M*	*SEM*
Prospective memory target items	.42	.05	.49	.06
General knowledge questions	.59	.02	.56	.01

(consistent with a ruminating aspect of depression that might impair attention to the prospective task), whereas a positive relationship was the trend for men (consistent with a strategy of distraction from the source of depression that could lead to increased focus on the prospective task).

GENERAL DISCUSSION

The alternate-forms reliability of the standard Einstein–McDaniel paradigm varied substantially across three experiments according to the number of ProM target items included and the overall level of ProM accuracy. Using *presidents* as targets tended to produce high levels of performance and low levels of ProM reliability with only 6 targets (Experiments 1 and 2), but higher ProM reliability did emerge with 30 targets. Using *animal* cues in Experiment 3 also produced good consistency with only 6 targets. These findings suggest that reliable ProM tasks can be achieved but that the magnitude of alternate-forms reliability itself is rather sensitive to changes in stimuli and procedures.

The levels of reliability achieved in Experiments 2 and 3 fell somewhat short of the minimum levels (r = .80–.90 or above) recommended for neuropsychological tests and clinical assessment tools (e.g., Cronbach, 1990). However, the standard Einstein–McDaniel paradigm was not developed for these purposes, and the moderate reliability observed in Experiments 2 and 3 is consistent with that found for other tests of memory. For example, the alternate-forms reliability for Wechsler's original memory scale ranged from .60 to .74 (McCarty, Logue, Power, Ziesat, & Rosenstiel, 1980). Finally, moderate levels of reliability can be acceptable for some purposes, including tests for differences between group means, which can be adjusted for measurement error (Cronbach, 1990).

Given that reliability has not been a major concern in other areas of memory research and that substantial progress has been made in exploring the theoretical mechanisms of ProM, why worry about developing sufficiently reliable tests? There are at least two causes for concern in the case of ProM. First, low reliability can reduce statistical power, and Uttl (2005a) estimated that more than 70% of published ProM studies have been underpowered already, which may account for some of the mixed findings in the literature on aging and ProM. Second, it is important to use reliable tests when examining individual differences, and the relationship between certain personality variables and ProM has drawn increasing interest. Using a moderately reliable ProM task in Experiment 3, for example, we detected a relationship between depressive symptoms and ProM accuracy. Further exploration of individual differences in ProM will require reliable tests of ProM.

In sum, we join the appeal of others (e.g., Uttl, 2005b) for increased work on the improvement of measurement tools in cognitive psychology. In the case of ProM, researchers are well advised to avoid ceiling performance in their tests, while taking care to maintain the construct validity of the ProM task. We hope the present work will lead to further work on measurement issues, including both reliability and validity, in this burgeoning field.

REFERENCES

Brandimonte, M. A., Ferrante, D., Feresin, C., & Delbello, R. (2001). Dissociating prospective memory from vigilance processes. *Psicológica*, **22**, 97-113.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper & Row.

Einstein, G. O., & McDaniel, M. A. (1990). Normal aging and prospective memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **16**, 717-726.

Einstein, G. O., McDaniel, M. A., Richardson, S. L., Guynn, M. J., & Cunfer, A. R. (1995). Aging and prospective memory: Examining the influences of self-initiated retrieval processes. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 996-1007.

Ellis, J., Kvavilashvili, L., & Milne, A. (1999). Experimental tests of prospective remembering: The influence of cue-event frequency on performance. *British Journal of Psychology*, **90**, 9-23.

International Personality Item Pool (2001). *A scientific collaboratory for the development of advanced measures of personality and other individual differences*. Available at ipip.ori.org/.

Marsh, R. L., Hancock, T. W., & Hicks, J. L. (2002). The demands of an ongoing activity influence the success of event-based prospective memory. *Psychonomic Bulletin & Review*, **9**, 604-610.

Maylor, E. A. (1998). Changes in event-based prospective memory across adulthood. *Aging, Neuropsychology, & Cognition*, **5**, 107-128.

McCarty, S. M., Logue, P. E., Power, D. G., Ziesat, H. A., & Rosenstiel, A. K. (1980). Alternate-form reliability and age-related scores for Russell's Revised Wechsler Memory Scale. *Journal of Consulting & Clinical Psychology*, **48**, 296-298.

McDaniel, M. A., & Einstein, G. O. (1993). The importance of cue familiarity and cue distinctiveness in prospective memory. *Memory*, **1**, 23-41.

McDaniel, M. A., Guynn, M. J., Einstein, G. O., & Breneiser, J. (2004). Cue-focused and reflexive-associated processes in prospective memory retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **30**, 605-614.

Schmidt, I. W., Berg, I. J., & Deelman, B. G. (2001). Prospective memory training in older adults. *Educational Gerontology*, **27**, 455-478.

Titov, N., & Knight, R. G. (2001). A video-based procedure for the assessment of prospective memory. *Applied Cognitive Psychology*, **15**, 61-83.

Uttl, B. (2005a). Age-related changes in event-cued prospective memory proper. In N. Ohta, C. M. MacLeod, & B. Uttl (Eds.), *Dynamic cognitive processes* (pp. 273-303). New York: Springer.

Uttl, B. (2005b). Measurement of individual differences: Lessons from memory assessment in research and clinical practice. *Psychological Science*, **16**, 460-467.

NOTE

[1] We are grateful to Lia Kvavilashvili for this suggestion.

(Manuscript received July 25, 2005;
revision accepted for publication April 8, 2006.)