Prospective memory (PM) refers to the ability to remember to carry out future intentions such as remembering to buy specific items at the grocery store or taking medications at the scheduled time (Einstein & McDaniel, 1990; Thompson, Henry, Withall, Rendell, & Brodaty, 2011; Troyer & Murphy, 2007). Intact PM is necessary for high-performance daily living and contributes to successful independent living in older adults (Schmitter-Edgecombe, Woo, & Greeley, 2009). PM requires the integrated functioning of multiple cognitive abilities that involve a complex network of brain regions including neocortical areas in the prefrontal lobes in addition to other cortical areas such as the mesio-temporal and parietal lobes, and subcortical structures such as the thalamus (Bisiacchi, Cona, Schiff, & Basso, 2011; Burgess, Quayle, & Frith, 2001; Spindola & Brucki, 2011; West, 1996). PM relies upon both retrospective and prospective components (Einstein & McDaniel, 1990). The retrospective component is dependent on episodic memory and is involved in encoding and retrieving the content of the future intentions (Einstein & McDaniel, 1990). By contrast, the prospective component is mediated by executive control functions that enable an individual to retrieve and initiate delayed intentions despite ongoing activities that may be unrelated to the task to be performed (Costa et al., 2010; Spindola & Brucki, 2011; Troyer & Murphy, 2007). Thus, the prospective component is crucial for remembering the delayed intention in the absence of an external prompt (Costa et al., 2010). The retrospective and prospective components of PM, respectively, are associated with the medial temporal lobe and...
prefrontal brain systems (Burgess et al., 2001; West & Krompinger, 2005) that may be compromised in preclinical dementia conditions—the focus of the current research. Unfortunately, the field lacks brief and clinically useful measures of PM for older adults that can be easily incorporated into neuropsychological test batteries.

PM tasks can be categorized into time-versus event-based and short-versus long-term. Event-based PM refers to instances in which a person remembers to carry out a previously planned action when prompted by a specific external event or stimulus (Einstein, McDaniel, Richardson, Guynn, & Cunfer, 1995; Kazui et al., 2005; Tam & Schmitter-Edgecombe, 2013). Examples of event-based PM tasks include remembering to purchase eggs on the way home from work or a train conductor remembering to perform a specific safety procedure before a trip. Time-based PM tasks include taking medication at the correct time each day or turning off the stove after 30 min. Time-based PM has been thought to rely on self-cuing, and event-based processes can be described as environmentally cued; age-related deficits in PM generally are more pronounced on time-based tasks (Delprado et al., 2012; Einstein et al., 1995; Troyer & Murphy, 2007).

PM tasks also can be divided into short- and long-term tasks performed during the assessment (i.e., in the laboratory) or hours or days after the test session, respectively. Long-term PM tasks have been generally considered to be more naturalistic or “ecologically valid” because they occur in everyday contexts where the examiner has little control over distractions or intervening task demands (Rendell & Craik, 2000).

Impairments in PM have been found in patients with mild cognitive impairment (MCI) compared to healthy older adult controls (HC; Blanco-Campal, Coen, Lawlor, Walsh, & Burke, 2009; Chi et al., 2014; Costa et al., 2010; Costa et al., 2011; Delprado et al., 2012; Kazui et al., 2005; Rabin, Chi, et al., 2014; Tam & Schmitter-Edgecombe, 2013; Thompson et al., 2011; Troyer & Murphy, 2007). MCI has been considered to be a preclinical dementia condition marked by both objective and subjective memory (and possibly other cognitive) impairment with relative sparing of activities of daily living and no global cognitive decline (i.e., no dementia; Delprado et al., 2012; Mitchell, Arnold, Dawson, Nestor, & Hodges, 2009; Petersen et al., 2001; Winblad et al., 2004). The various subtypes of MCI including those with amnestic and nonamnestic features can be conceptualized as intermediate condition between normal cognitive aging and frank dementia (Costa et al., 2010; Spindola & Brucki, 2011) based on outcomes from multiple longitudinal studies examining the clinical progression of individuals with MCI (Ahmed, Mitchell, Arnold, Nestor, & Hodges, 2008; Mitchell et al., 2009; Rosenberg, Johnson, & Lyketsos, 2006). Patients with MCI show underlying neurological changes in cerebral structures, markedly in medial temporal and prefrontal areas (Costa et al., 2010; Delprado et al., 2012; Spindola & Brucki, 2011), similar to what is observed in patients with dementia though somewhat attenuated (Mitchell et al., 2009; Petersen et al., 2001). Although rates have varied from one study to another due to substantial variability in methods used to diagnose and assess MCI, some studies have reported annual conversion rates from MCI to Alzheimer’s disease (AD) of 10 to 15% (Mitchell & Shiri-Feshki, 2010; Panza et al., 2005). Given that PM is impaired in older adults with MCI and mild AD (Blanco-Campal et al., 2009; Spindola & Brucki, 2011; Thompson, Henry, Rendell, Withall, & Brodaty, 2010; Troyer & Murphy, 2007; van den Berg, Kant, & Postma, 2012), assessment of PM in these individuals may enhance understanding of the cognitive changes that occur in the early stages of neurodegenerative cognitive decline and improve assessment and diagnostic approaches for preclinical dementia conditions.

Clinical and research interest in PM in older adults, specifically in those with MCI or AD, is fairly recent, and there are a limited number of reliable neuropsychological tests. The most commonly reported measures include the Rivermead Behavioral Memory Test (RBMT; Wilson, Cockburn, & Baddeley, 1985, 2003), Memory for Intentions Screening Test (MIST; Raskin, 2009; Raskin, Buckheit, & Sherrod, 2010), and Cambridge Prospective Memory Task (CAMPROMPT; Wilson et al., 2005). The RBMT offers four alternate versions to mitigate practice effects associated with repeated-measures in longitudinal studies and consists of three short-term, event-based tasks (Wilson et al., 2003). However, a limitation of the RBMT is that it does not contain time-based tasks over a short retention period or event- and time-based tasks over a long-term retention period (Radford et al., 2011; Wilson et al., 2003). Because all types of PM tasks (short-term, event-based; long-term, event-based; short-term, time-based; long-term, time-based) map to common everyday cognitive activities, it is important to measure PM within each
of these domains. Both the CAMPROMPT and MIST include event- and time-based PM tasks carried out in laboratory settings (Raskin et al., 2010; Wilson et al., 2005), and the MIST also includes an optional long-term PM task and two alternate forms (Raskin, 2009; Raskin et al., 2010). However, the lengthy administration times and inclusion of filler tasks (e.g., crossword puzzles, word searches) in the CAMPROMPT and MIST may make them difficult to incorporate into already lengthly assessment batteries. Additionally, the CAMPROMPT lacks alternate forms (Radford et al., 2011; Wilson et al., 2005).

To address the perceived limitations of currently available clinical PM tasks, Radford and colleagues (2011) developed the Royal Prince Alfred Prospective Memory Test (RPA-ProMem). The RPA-ProMem contains four types of PM tasks (i.e., one short-term, time-based task; one short-term, event-based task; one long-term, event-based task; and one long-term, time-based task). The long-term PM tasks are completed outside the laboratory on the same day as the assessment and 1 week after the test date, which enables measurement of PM performance in naturalistic settings. The scoring system is based on how successful the participant was in remembering and carrying out the intended tasks (Radford et al., 2011). The availability of three alternate forms diminishes possible practice effects of repeated testing, and the brief administration time of approximately 10 min allows for relatively easy integration into existing test batteries.

Recently, the psychometric properties of the RPA-ProMem have been validated in a study that examined PM functioning in a small sample (N = 40) of mixed neurological patients and HCs (Radford et al., 2011). The RPA-ProMem demonstrated high interrater reliability, moderate alternate-form reliability, and acceptable levels of both convergent and face validity in neurological patients (Radford et al., 2011). However, the psychometric properties of the RPA-ProMem in measuring PM in older adults have not yet been fully explored. A recent study of the RPA-ProMem by Rabin, Chi, and colleagues (2014) found that cognitively healthy older adults performed significantly better on the RPA-ProMem than those with MCI. Also, long-term PM tasks of the RPA-ProMem were sensitive to older adults with subjective cognitive decline (SCD). SCD is suggested as a possible “pre-MCI” condition because older adults with SCD present higher than normal levels of cognitive complaints despite intact performance on traditional neuropsychological tests as well as displaying biomarker abnormalities also observed in neuropathology of AD (Jessen et al., 2014; Reisberg, Shulman, Torossian, Leng, & Zhu, 2010; Saykin et al., 2006; Visser et al., 2009). Rabin, Chi, and colleagues (2014) suggested that the RPA-ProMem may help detect subtle cognitive changes associated with the earliest stages of dementia, which are difficult to capture with standard episodic memory tests. Furthermore, the RPA-ProMem demonstrated convergent validity with the MIST total scores (Rabin, Chi, et al., 2014).

The current research represented an expansion of the initial Rabin, Chi, et al. (2014) study and aimed to further investigate the psychometric properties of the RPA-ProMem in community-dwelling older adults. Three hypotheses were proposed. First, we expected that interrater reliability for the RPA-ProMem would be strong given the straightforward administration, standardized scoring procedure, and previous results (Radford et al., 2011). Second, internal consistency was predicted to be moderate because, although each RPA-ProMem test item measures the same latent variable of PM, the task was designed to assess different aspects of PM (i.e., construct heterogeneity). Thus, each PM subtask was expected to contribute unique information. Third, we hypothesized that the three forms of the RPA-ProMem would demonstrate parallel forms reliability given the similarity of the tasks across the different test versions. Validation of the psychometric properties of the RPA-ProMem within an older adult population might promote use of this instrument in clinical and research settings.

Methods

Participants
Participants (N = 257) were a subset of individuals recruited into the Einstein Aging Study (EAS), a community-based longitudinal study of older adults ages 70 and above, who reside in Bronx County, NY (Katz et al., 2012; Lipton et al., 2003). Through a systematic sampling process, EAS participants were randomly selected from Medicare recipient or voter registration lists in Bronx County. Exclusion criteria for the EAS included severe audiovisual disturbances, medical or psychiatric conditions, lack of comprehension due to language barrier, and institutionalization or nonambulatory status. For the current study, those diagnosed with dementia under the DSM-IV criteria were also
excluded. Approval to conduct the research was granted by the institutional committee for the protection of human subjects, and each participant provided written informed consent. Participants were offered lunch and car services to and from the testing facility, and were compensated $25 for their participation. Participants were tested over 2 days, separated by a 2-week interval. Participants reported for Day 1 as part of their annual EAS visit, and completed a standard neuropsychological assessment and neurological and physical examinations (Katz et al., 2012). On Day 2, participants completed study-specific measures, which included the RPA-ProMem (Rabin, Chi, et al., 2014; Radford et al., 2011).

Participants’ demographic information by group is displayed in Table 1. Mean age was 80.78 (SD = 5.53), mean years of education was 14.45 (SD = 3.41), and the percentage of women was 67.7%. Race was dichotomized as White or non-White, and most participants identified as White (60.0%). Notably, the sample included Black (30.4%), Hispanic White (5.4%), Hispanic Black (1.5%), and Asian (0.4%) individuals, as well as those who were identified as other (2.3%). Group differences in demographic characteristics were determined using Analyses of Variance (ANOVs) for age and years of education, and Pearson chi-squared tests for sex and race.

Materials

Tests used for participant classification. A group of 13 neuropsychological tests was used to screen each participant for classification purposes: (a) free recall from the Free and Cued Selective Reminding Test (FCSRT; Grober & Buschke, 1987); (b) Logical Memory I subtest of the Wechsler Memory Scale-Revised (Wechsler, 1987); (c) Letter Fluency (Spreen & Strauss, 1998); (d) verbal fluency/naming exemplars from a category—Category Fluency (Rosen, 1980); (e) short form of the Boston Naming Test (Kaplan, Goodglass, & Weintraub, 1983); (f-g) Trail Making Test Parts A and B (Reitan, 1958); and select subtests of the Wechsler Adult Intelligence Scale-Third Edition (Wechsler, 1997) including Block Design, Digit Symbol-Coding, Digit Span, Information, Vocabulary, and Similarities.

The RPA-ProMem. The RPA Pro-Mem (Radford et al., 2011) assesses PM with four tasks of varying types, which are easily incorporated into existing neuropsychological test batteries. Each task is either an event- or a time-based task with short- or long-term retention periods. The short-term tasks are administered and completed in the laboratory, and the long-term tasks are introduced at the end of the test battery and completed outside of the laboratory. During administration of the neuropsychological battery, a digital clock is provided on the desk for participants’ use. Furthermore, the use of any kind of memory strategy is permitted while completing all four RPA-ProMem tasks, and participants in the current study were free to use the provided note pads, writing tools, or other methods of their choosing.

At the beginning of the test session, the examiner informed participants that they would be asked to “remember to do some things at a later stage of the assessment.” The first part of the RPA-ProMem presented a short-term, time-based PM task in which the examiner stated, “In 15 minutes time, I would like you to remind me to move my car so I don’t get a ticket.” Participants were expected to keep track of time and remind the examiner to move the car at the correct time (please note that the exact content of the instructions differs across the three forms of the RPA-ProMem). The second part presented a short-term, event-based task. For example, the examiner stated, “When my cell phone rings, tell me you would like a drink,” which was said following the response for the first part or anytime from 15 to 30 min after the administration of the first task if no response was provided. The phone was set to ring in 30 min, and participants were expected to respond immediately following the stimulus. The third and fourth parts introduced long-term PM tasks toward the end of the testing session. The third task was an event-based task in which participants were instructed to call the examiner upon arriving home following the assessment. Participants were informed that the phone call would be directed to a voicemail box and that they should leave a message (e.g., what time it was when they arrived at home). The fourth task was
Reliability of a Clinical PM Task | Ko and Rabin

time-based. Participants were asked to return a postcard a week from the test date with their name and a description of something written on the postcard (e.g., dinner menu for that night). Participants received the examiner’s contact information on a business card and a stamped, addressed, and labeled postcard was provided after the instructions for Parts 3 and 4 were read aloud by the examiner.

The RPA-ProMem was scored using an objective scoring system with each part given a score ranging from 0 to 3, with 3 indicating the best possible performance. Total scores were calculated as the sum of the individual parts (range 0–12). Scores were awarded dependent on two factors: accuracy in time of the response and correctness of the response. Table 2 presents the scoring criteria, and further details can be found in Radford et al. (2011) because our study followed the original scoring protocol.

Procedure
Classification of participants. We took a psychometric approach to participant classification. Participants were designated as HC (n = 118), SCD (n = 83), or MCI (n = 56) based on their performance on the 13 neuropsychological tests previously mentioned. The mean scores and standard deviations of individuals in the current sample were compared to robust norms derived from a sample of 411 healthy EAS participants who were dementia free for at least 3 years (Rabin, Chi, et al., 2014; Rabin, Wang, Katz, & Lipton, 2014). We evaluated participants’ scores on three cognitive factors, which were identified by subjecting the 13 tests (listed under Materials) to exploratory factor analysis, which resulted in three underlying cognitive factors: (a) global/verbal (Boston Naming, Information, Similarities, Vocabulary, Digit Span, and Letter Fluency), (b) executive/processing speed (Block Design, Digit Symbol-Coding, and Trail Making Test Parts A & B), and (c) memory (FCSRT, Category Fluency, Logical Memory). Cognitive domain factor scores for each participant were calculated as the average z score of each neuropsychological test associated within a given factor. These were derived using means and standard deviations of the robust sample stratified by age group (Rabin et al., 2012; Rabin, Wang, et al., 2014).

Individuals were classified as MCI if at least one of their cognitive domain factor z scores fell considerably lower (> 1 SD) than the mean of the robust sample and they reported cognitive complaints on any of the EAS subjective report questionnaires (Rabin et al., 2012; Rabin, Chi, et al., 2014; Rabin, Wang, et al., 2014). Participants were designated as SCD if none of their cognitive domain factor z scores were considerably lower (> 1 SD) than the mean of the robust normative group but they reported cognitive complaints exceeding an optimal cut-off score as determined by the novel classification system (see Rabin, Wang, et al., 2014, for details). Participants were designated as HC if none of their cognitive domain factor z scores were considerably lower (> 1 SD) than the mean of the robust normative group and their cognitive complaint scores did not exceed the optimal cut-off score. The 10 cognitive complaint items used in this classification approach have reported reliability and predictive validity for dementia (Rabin et al., 2012).

Measuring reliability. For the purpose of measuring internal consistency and interrater reliabilities, only Form 3 of the RPA-ProMem was administered to the initial sample of participants (N = 257) including HC, SCD, and MCI groups. However, for measuring parallel forms reliability, 77 of the 118 HC’s participated in the subsequent annual assessment and were randomly administered one of the three forms of the RPA-ProMem. Form 1 was administered to 24 participants, Form 2 to 30 participants, and Form 3 to 23 participants.

To determine interrater reliability of the scoring criteria for the RPA-ProMem, a subset of the current study’s sample was used. Approximately 200 of the 257 sample protocols (nearly 80% of sample population) were randomly selected to be scored by a second rater. Each rater scored

| TABLE 2 | Guide to Scoring the Royal Prince Alfred Prospective Memory Test |
|---|---|---|---|
| Score | Parts 1 and 2 | Part 3 | Part 4 |
| 0 | no response, or incorrect response made at greater than 2 min delay | no call made within 2 days of test date | no postcard returned up to 2 weeks after test date |
| 1 | correct response made at greater than 5 min delay | call at incorrect time with incorrect message | postcard returned in incorrect day with incorrect information |
| 2 | correct response made at 2 to 5 min delay, or incorrect response made up to 2 min of the correct time | call at incorrect time with correct message, or call at correct time with incorrect message | postcard returned on incorrect day with correct information, or postcard returned on correct day with incorrect information |
| 3 | correct response made within 2 min of correct time | call at correct time with correct message | postcard returned on correct day with correct information |

(See table for scoring criteria.)
Comparisons demonstrated that the MCI group had significantly lower mean years of education than the SCD ($p < .001$) and HC ($p < .001$) groups. Participants in the MCI group also had a significantly higher percentage of non-Whites than those in the SCD ($p < .001$) and HC ($p < .001$) groups. These results are not surprising given other findings that underrepresented ethnoracial groups and groups with low education are at increased risk for dementia (Demirovic et al., 2003; Miles, Froehlich, Bogardus, & Inouye, 2001; Tang et al., 2001).

There was strong interrater reliability for RPA-ProMem total scores ($\rho = .97$, $p < .001$). Additional interrater reliability analyses were conducted for each part of the RPA-ProMem. All four coefficients were high, though slightly lower than for the total scores. For Part 1, ICC = .96 ($p < .001$), for Part 2, ICC = .91 ($p < .001$), for Part 3, ICC = .96 ($p < .001$), and for Part 4, ICC = .96 ($p < .001$). Table 3 shows the mean total scores and mean scores for each part assigned by the two raters as well as corresponding ICC values.

Cronbach’s alpha is a commonly used statistic for internal consistency reliability (Cronbach, 1951). Cronbach’s alpha for the RPA-ProMem was .50. The RPA-ProMem is composed of a relatively small number of items ($n = 4$, divided into 4 parts), which creates difficulty for RPA-ProMem in reaching a strong alpha coefficient. However, Table 4 shows that Cronbach’s alpha for RPA-ProMem would be lowered if each item was deleted from the test, suggesting that all four PM tasks added to the overall internal consistency reliability of the RPA-ProMem.

Parallel forms reliability was determined by establishing congruency between the mean scores of each individual part of the RPA-ProMem. Table 5 shows the mean scores of the four parts of the RPA-ProMem for the three alternate forms. A one-way between-subjects ANOVA was conducted to compare the effect of test form on PM task performance in each part of the RPA-ProMem. $F$ and $p$ statistics are reported under Table 5 with effect sizes ($\eta^2$). There were no significant effects of form on test performance across the four parts of the RPA-ProMem, and effect sizes were small. We also calculated Spearman’s $\rho$ correlation coefficients among the subscores to explore whether they varied consistently across the sample. Results were as follows: Form 1 and Form 2, $\rho = .68$ ($p = .321$); Form 1 and Form 3, $\rho = .87$ ($p = .132$); Form 2 and Form 3, $\rho = .73$ ($p = .267$).
Reliability of a Clinical PM Task | Ko and Rabin

Discussion

PM deficits are commonly observed in older adults with MCI and mild dementia, and are linked to lowered quality of life (Schmitter-Edgecombe et al., 2009; Thompson et al., 2010; Troyer & Murphy, 2007; van den Berg et al., 2012). Despite the importance of PM to everyday functioning and independent living, however, PM tests are rarely utilized in clinical neuropsychological assessments of older adults. Accurate measures of PM could serve as useful indicators in profiling the cognitive abilities of older adults in clinical and research settings. The RPA-ProMem is a recently developed PM measure with multiple advantages such as measuring various components of PM, having three alternate forms, and having a brief administration time (Radford et al., 2011).

Before the RPA-ProMem can be used routinely in assessments of older adults, adequate psychometric properties must be established. Mean scores on the RPA-ProMem in healthy older adults were previously distinguished from those with MCI and from those with SCD (after controlling for relevant covariates of education, race, and depressive symptoms) suggesting that the RPA-ProMem can discriminate among groups of older adults with different levels of cognitive deficits (Rabin, Chi, et al., 2014). This finding has notable importance considering that varying degrees of PM deficits have been identified in those with SCD, MCI, and AD (Blanco-Campal et al., 2009; Spindola & Brucki, 2011; Visser et al., 2009). Construct validity of the RPA-ProMem was previously documented in 20 healthy and 20 neurological patients (Radford et al., 2011). Rabin, Chi, and colleagues (2014) provided further support for the construct validity of the RPA-ProMem in older adults by demonstrating a significant correlation ($p < .001$) between scores on the RPA-ProMem and another widely used PM task. To further validate the usefulness of the RPA-ProMem in measuring PM deficit in older adults, the current study focused on three types of reliability.

Interrater Reliability Results

Strong interrater reliability is essential in yielding generalizable results on a neuropsychological assessment by demonstrating that the scoring approach is objective across multiple raters (Multon, 2010). Specific measures with proven interrater reliability can be utilized in research with large numbers of protocols using multiple raters to score protocols efficiently. As hypothesized, ICC for the RPA-ProMem total score was strong (ICC = .97), consistent with previous findings of Radford et al. (2011). Additional analyses by RPA-ProMem parts showed correlations above .90, indicating that scoring procedures for each part are reliable among the raters and also suggesting that the current scoring method is objective. Referring to Table 2, the objectiveness of the scoring method can be justified by the fact that scores are determined by two factors: (a) correct versus incorrect response and (b) the time in which the response was given. With respect to time, specific timeframes were predetermined for full or partial scoring; use of these criteria by highly trained examiners likely minimized subjectivity in scoring outcomes.

Internal Consistency Reliability Results

It is important to consider the consistency between each item (i.e., item homogeneity) with respect to the psychological construct the test was intended to measure. Internal consistency reliability measures the extent of equivalence, or the degree of content variance, among items (Cronbach, 1951). In other words, internal consistency of test items elucidates whether the content of each test item is related to the remaining items. In the case of the RPA-ProMem (Radford et al., 2011), each of the four PM parts is considered as an individual test item, and the degree of item homogeneity is determined by the interrelationships between the four parts. Cronbach’s alpha coefficient has been regarded as a conservative estimate of reliability, producing lower-bound estimates (Allen & Yen, 1979; Cronbach, 1951). In general, alpha coefficients equal to or greater than .60 are considered acceptable, but alpha coefficients that are too high suggest redundancy in test items (Cronbach, 1951). To some extent, redundancy is necessary for the test to measure a general construct. At the same time,

| TABLE 3 |
| Interrater Reliability for the Royal Prince Alfred Prospective Memory Test |
|---|---|---|
| **Rater 1** | **Rater 2** | **ICC** |
| **M (SD)** | **M (SD)** |  |
| **Total Score** | 7.88 (2.86) | 8.05 (2.75) | .97 |
| **Part 1 (short-term, time-based)** | 1.74 (1.13) | 1.75 (1.14) | .96 |
| **Part 2 (short-term, event-based)** | 2.14 (0.99) | 2.24 (0.98) | .91 |
| **Part 3 (long-term, event-based)** | 2.27 (1.11) | 2.21 (1.09) | .96 |
| **Part 4 (long-term, time based)** | 1.85 (1.17) | 1.94 (1.14) | .96 |

*Note. ICC = intraclass correlation coefficient. ICC significant with $p < .001$. |
redunancy minimizes the opportunity for the test item to gather novel and/or unique information (Cronbach, 1951).

Cronbach’s alpha for the RPA-ProMem was lower than expected. PM is a heterogeneous construct that utilizes multiple cognitive functions and neural networks (Burgess et al., 2001; Costa et al., 2010) and includes multiple subtypes based on cue type and retention period (Delprado et al., 2012; Spindola & Brucki, 2011). Such heterogeneity is important because a person may experience varying levels of difficulty in carrying different types of PM tasks. For example, older adults with SCD experienced more difficulty in completing PM tasks with longer retention periods (completed outside the laboratory) than tasks with short retention periods (Rabin, Chi, et al., 2014). Normal aging processes can affect PM capabilities such as leading to difficulty with time-based PM tasks (Spindola & Brucki, 2011). Because the RPA-ProMem presents four PM tasks of four different conditions (Radford et al., 2011), a lower alpha would not be unexpected. Part 1 is a time-based task carried out in 15 min time while the participant is still in the laboratory. Part 2 is cued by an external stimulus (i.e., phone ring), and the retention period is relatively short. Parts 3 and 4 with event- and time-based cues, respectively, are carried out in a naturalistic environment (Radford et al., 2011). Therefore, it is possible that the ability of the RPA-ProMem to examine various components of PM contributed to the increased construct heterogeneity and resulted in the relatively low alpha coefficient.

For a given test, a range of alpha coefficients is possible regardless of its other types of reliability properties. For example, a widely used battery, the Neuropsychological Assessment Battery (White & Stern, 2003), shows variability in the alpha coefficients of its subtests, ranging from .24 for the Screening Visual Discrimination test to .79 for the Digits Backward test. The wide range of alpha coefficients was explained by differences in levels of item homogeneity and variability of individual test scores (White & Stern, 2003). The interrelationships between items, represented by Cronbach’s alpha, are easily influenced by two factors: (a) the number of scored items and (b) range of possible scores (Cronbach, 1951). Therefore, the observed alpha coefficient of .50 for the RPA-ProMem was understandable given that the RPA-ProMem is comprised of only four items, with only four possible scores (i.e., 0, 1, 2, 3) for each task. Table 4 shows the alpha coefficient of the RPA-ProMem if each item is deleted from the test. The resulting alpha coefficients are lower than the original value, showing that each task of RPA-ProMem Form 3 is interrelated to the other three tasks. Potential ways to improve internal consistency reliability of the RPA-ProMem are discussed below.

### Parallel Forms Reliability Results

Given limitations in gaining access to study participants on multiple occasions, we were unable to follow the conventional method of examining reliability between the three alternate forms of the RPA-ProMem. Typically, parallel forms reliability is measured by directly comparing performances across forms (i.e., alternate forms are administered to the same participant one after the other with an approximate two-week interval in between the test forms; Radford et al., 2011). Instead, we used an independent-groups design in which HCs were randomly assigned one of the three alternate test forms. The three groups of HCs, each receiving a different form, were equivalent with regard to their cognitive capabilities. Because the four PM tasks of the RPA-ProMem each measure a different aspect of PM, an ANOVA comparing mean scores was performed for each individual part. For example, the mean scores of those who received Form 1 was compared to those of who received Forms 2 and 3 and vice versa (Allen & Yen, 1979). The idea behind this approach was that score outcomes for the equivalent groups would not differ significantly if the alternate PM tasks for each of the RPA-ProMem parts were parallel to one another in terms of measuring that specific aspect of PM.

In fact, mean scores for each of the four parts of the RPA-ProMem were not statistically different, suggesting congruency between the alternate forms for each PM task. However, the p value for Part 2 nearly approached significance at

---

**TABLE 4**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Cronbach’s alpha for the RPA-ProMem total score</th>
<th>Cronbach’s alpha for the RPA-ProMem if each part were to be deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part 1 (short-term, time-based)</td>
<td>.41</td>
<td>.41</td>
</tr>
<tr>
<td>Part 2 (short term, event-based)</td>
<td>.50</td>
<td>.45</td>
</tr>
<tr>
<td>Part 3 (long-term, event-based)</td>
<td>.38</td>
<td>.38</td>
</tr>
<tr>
<td>Part 4 (long-term, time-based)</td>
<td>.48</td>
<td>.48</td>
</tr>
</tbody>
</table>
reliability has yet to be determined in an older adult sample. Our results suggested that the RPA-ProMem can be used for multiple assessments of PM as part of longitudinal assessments with minimal measurement error from practice effects.

As reported, internal consistency reliability of the RPA-ProMem fell below the moderate level. Literature has indicated that Cronbach’s alpha is influenced by multiple factors including test length (Cronbach, 1951). Therefore, the addition of PM tasks to the RPA-ProMem might enhance its internal consistency reliability. For example, doubling the number of tasks to eight might be useful, with each of the four additional tasks representing each of the four PM task types. The addition of a second task for each PM type would mean that there would be less variance between the test items (Cronbach, 1951). With the suggested modifications, the scores for each PM task would be more accurate for detecting cognitive deficits in specific aspects of PM. However, the length of administration time would double to roughly 20 min with these changes. If accuracy and depth of the test can be expanded to increase its utility, the addition of tasks at the expense of elongating administration time might be acceptable.

A question remains as to whether the RPA-ProMem is truly a suitable method for measuring PM in older adults, partly due to concerns about the complex PM construct. PM draws upon multiple cognitive systems for successful task execution (Burgess et al., 2001; Spindola & Brucki, 2011; West, 1996), bringing challenges to researchers and clinicians in developing tasks that adequately capture this complexity. A possible future direction for task validation involves functional neuroimaging during task performance, although this would not be feasible in conjunction with naturalistic PM tasks executed outside the laboratory. Investigating the relation of the RPA-ProMem scores to real-world functional outcomes has also been suggested to address issues of external and ecological validity. Finally, it is important to note that the RPA-ProMem was well-tolerated by our study participants in terms of participants’ ability to understand task instructions and carry out the tasks both in the lab and at home without reported difficulty. All participants who were administered the task over a 2-year period (420 individuals) were able to perform. Our examiners also noted the test to be relatively easy to administer and score given clearly articulated and standardized instructions and scoring criteria.

**TABLE 5**
Analysis of Variance Results for Alternate Forms of the Royal Prince Alfred Prospective Memory Test

<table>
<thead>
<tr>
<th>Form 1</th>
<th>Form 2</th>
<th>Form 3</th>
<th>F (2, 74)</th>
<th>n²</th>
</tr>
</thead>
<tbody>
<tr>
<td>M (SD)</td>
<td>M (SD)</td>
<td>M (SD)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Part 1 (short-term, time-based)</td>
<td>1.71 (1.27)</td>
<td>1.43 (1.36)</td>
<td>1.78 (1.31)</td>
<td>.59</td>
</tr>
<tr>
<td>Part 2 (short-term, event-based)</td>
<td>1.79 (1.28)</td>
<td>2.57 (1.01)</td>
<td>2.13 (1.22)</td>
<td>.06</td>
</tr>
<tr>
<td>Part 3 (long-term, event-based)</td>
<td>1.88 (1.42)</td>
<td>2.33 (1.09)</td>
<td>1.83 (1.30)</td>
<td>.27</td>
</tr>
<tr>
<td>Part 4 (long-term, time-based)</td>
<td>1.83 (1.09)</td>
<td>2.03 (1.16)</td>
<td>1.96 (1.15)</td>
<td>.81</td>
</tr>
</tbody>
</table>
Conclusion

Although the psychometric properties of the RPA-ProMem are yet to be fully investigated, the current study offered valuable preliminary information about the reliability of the RPA-ProMem in a community-dwelling sample of older adults. With growing research interest in PM functioning in older adults with varying levels of objective and subjective cognitive deficits, utilization of the RPA-ProMem along with more traditional episodic memory tests seems warranted given the importance of assessing PM during neuropsychological assessments. Overall, the RPA-ProMem has much to offer because it provides a brief approach to measuring PM in both laboratory and naturalistic settings, and it may be sensitive to early neurodegenerative cognitive changes.

References

Reliability of a Clinical PM Task | Ko and Rabin


Author Note. HaYoung Ko, Department of Psychology, Macaulay Honors at Brooklyn College, 2900 Bedford Avenue Brooklyn, New York 11210; Laura A. Rabin, Department of Psychology, Brooklyn College and The Graduate Center of The City University of New York, 2900 Bedford Avenue Brooklyn, New York 11210.

This research was supported by the National Institute on Aging (NIA) and National Institute of General Medical Sciences (SC2AG039255), NIA (AG03949), Alzheimer’s Association (NIRG-11-206369), Czap Foundation, and The Leonard and Sylvia Marx Foundation.

The authors wish to thank Susan Chi, Ashu Kapoor, Mihushka Elbulok-Charcape, Valdiva Da Silva, Tangeria Ramratan, Cuiling Wang, Richard Lipton, and Mindy Katz for their contributions.

Correspondence concerning this article should be sent to HaYoung Ko, 21617 Northern Boulevard 2nd Floor Bayside, New York 11361. E-mail: h.ryu9114@gmail.com