Interrater Reliability of the Revised Cognitive Performance Test (CPT): Assessing Cognition in People With Neurocognitive Disorders

Patricia Schaber, Emily Stallings, Collette Brogan, Fouzia Ali

The rigor of occupation-based standardized assessments that rely on observational scoring procedures depends on proven reliability among test administrators. This study measured interrater reliability of the Cognitive Performance Test (CPT), a standardized, occupation-based assessment that measures cognitive–functional capacity in older adults with neurocognitive disorders. To capture a range of experience among test administrators, two sets of raters—four expert and three novice—scored video recordings of 10 patients administered the CPT. Interrater reliability results were strong among all raters (intraclass correlation coefficient [ICC] = .93), with expert raters (ICC = .97) yielding higher coefficients than novice raters (ICC = .93). Spearman’s ρ correlation coefficients were high among all raters (rs = .92–1.00). Practitioners can be confident that results of the CPT give accurate and consistent information to the health care team, family members, and patients when administered with fidelity using standardized protocols.


The Cognitive Performance Test (CPT) is a standardized, performance-based assessment developed by Burns in 1990 as a research instrument and revised in 2002 as a measure of cognitive–functional capacity in older adults with Alzheimer’s-type dementia (Burns, 1990, 2002; Burns, Mortimer, & Merchak, 1994). The instrument is grounded in Allen Cognitive Disability Theory, which asserts that impairment in information processing has a detrimental effect on a person’s ability to function and that this impairment can be described with an ordinal scale (Allen, 1985; Burns et al., 1994). In contrast to assessments that directly measure skills for instrumental activities of daily living (IADLs) and activities of daily living (ADLs), the CPT uses familiar tasks to determine the degree to which particular deficits in information processing compromise performance of everyday activities.

With most neurocognitive disorders, sensory–perceptual memory, working memory, and long-term memory are progressively affected, initially challenging complex–novel tasks and, later, simple–habitual daily tasks (Levy & Burns, 2011). The results of the CPT are a measure of global cognitive capacity used to gauge the ability of the client to successfully manage his or her daily tasks. These results can guide health care management and support to families with older adults with neurocognitive disorders or dementia (Schaber, 2010).

Occupation-based assessments rank high in ecological validity because of the core structure of using real, observable activities in contrast to contrived activities (Burgess et al., 2006). Results are measures of observed, behavioral responses in performance, such as speed, accuracy, and need for assistance to complete the

MeSH TERMS
- activities of daily living
- cognition
- neurocognitive disorders
- reproducibility of results
task. Occupation-based measures that use common tasks prevail in cognitive–functional assessment on the basis of the heightened validity that these measures have in predicting function compared with paper-and-pencil tests or verbally based question-and-answer assessments (Gillen, 2011). The term validity refers to the degree that the assessment measures the skills that it purports to measure in performance of daily activities, thereby increasing representativeness and generalizability (Vaskinn, Sergi, & Green, 2009). Occupation-based measures also provide a rich opportunity to observe the patient’s visual, auditory, motor, and processing skills during the administration process.

Reliability of assessment results depends on skill and accuracy of the test administrator, especially with occupation-based measures that use standardized scoring procedures based on clinical observations. Interrater reliability is defined as the ability of the test to reproduce the same result among two or more test administrators measuring the same participants (Portney & Watkins, 2009). To reduce measurement error in interrater studies, optimally, administration is a single trial with simultaneous and independent ratings (Portney & Watkins, 2009). Video rating allows multiple raters to observe performance independently. The aim is to build trust in the test results by obtaining independent agreement among raters who are representative of a larger pool of test administrators (DiFabio, 2013). Optimally, raters should be selected randomly and should be reflective of the larger test administrative group with a similar range of experience and test training. Random selection heightens the generalizability of the test to all raters.

Cognitive Performance Test

The CPT is organized into six ordinal cognitive levels ranging from normal cognitive functioning (5.6) to unresponsive (1.0). The test measures cognitive decline in predictable levels; each level identifies impact on functional decline: mild (5.0), mild to moderate (4.5), moderate (4.0–3.5), moderate to severe (3.0), and severe (2.5–2.0; Levy & Burns, 2011). The score is based on an average of the ratings on seven tasks: filling a medication box, shopping for a belt, washing hands, making toast, buying paint using a phone, selecting outerwear, and using a map. The tests are administered with standardized verbal instructions and prompts, as needed, to correct errors or to complete the task. Set steps are in place to distract the patient being tested while completing the task. Scoring is determined by the patient’s responses, including accuracy, ability to correct errors, and need for cues for task completion (Burns, 2013).

Advantages of the CPT are that it is largely accepted by patients because the tasks are relatively simple and because the test objects are familiar in the patient’s daily tasks. This acceptance reduces tension around test performance. The CPT score is based on observation of performance (rather than self-report or caregiver report), and the test takes less than 30 min to administer in a clinic. The CPT can be used in home care evaluations using five of the seven tasks; the dress and travel tasks are difficult to transfer to the home setting (Burns, 2013).

Reliability and Validity

Studies on the CPT include use of the assessment as a hospital predischarge measure; the CPT was shown to be an effective tool to predict harm, defined as the time until the patient would experience physical injury to self or property loss or damage (Douglas, Letts, Richardson, & Eva, 2013). Douglas, Letts, Eva, and Richardson (2012) examined concurrent validity with the Standardized Mini-Mental State Exam (SMMSE; Molloy & Clarnette, 1999) and the Process scale of the Assessment of Motor and Process Skills (AMPS; Fisher, 2006). They found significant correlations of the CPT with SMMSE (r = .47, p < .01) and AMPS–Process (r = .53, p < .01). They also found weak correlations with ADLs—FIM™ (Uniform Data System for Medical Rehabilitation, 1997)—r = .32, p < .05—and that age and years of education did not affect assessment results. In an earlier study, Burns et al. (1994) found that the CPT was a valid predictor of the risk for institutionalization on the basis of a generalized Wilcoxon test (p = .003). The test has not been validated with people who do not read or use English as a primary language and has not been adjusted for bias on the basis of age, education, or cultural and socioeconomic status (Lancu & Olmer, 2006).

Interrater reliability studies of the CPT demonstrated strong psychometrics on the 1990 version (Bar-Yosef, Weinblatt, & Katz, 1999; Burns et al., 1994). Burns et al. (1994) used two occupational therapists as raters of test administration 4 wk apart: the author of the CPT test and an occupational therapist. This study revealed intraclass Pearson correlations for differences in mean ratings (intraclass correlation coefficient [ICC] = .91). The 1994 version of the CPT included six tasks; a seventh task was added in 2002. Interrater reliability was reported on CPT total scores, not delineated by task ratings. In the 1990 study, Burns et al. had five occupational therapists view videos of the CPT; two therapists independently rated video performance on each participant (n = 30).
The results were $\kappa$ tests between two raters ($r = .98$, $p < .001$). In this study, the CPT was a modified version of the six-task assessment with three of the tasks altered (toast task was replaced with a coffee task, an eye doctor was substituted for a paint store in the phone task, and shirts were purchased instead of belts in the shop task). These psychometrics of the CPT six-task assessment and modified six-task assessment continue to be cited to support current studies in which researchers have used the revised, seven-task CPT as a study measure (Douglas et al., 2012, 2013). In this study, we aimed to determine the interrater reliability of the current version: the seven-task CPT. We also aimed to increase the rigor of this statistic by selecting raters with a range of experience and training reflective of test administrators in clinical practice.

Method

Ten patients presenting with memory deficits were evaluated in a hospital-based outpatient memory clinic in the Midwest by a comprehensive medical diagnostic team. Under institutional review board approval, 10 patients (7 women and 3 men) and their family members (primary caregivers) were informed and consented to participate in the study, including being videotaped during administration of the CPT. A licensed occupational therapist administered the CPT, and another licensed occupational therapist videotaped the initial evaluation. A profile of the 10 patients indicated that the mean age was 75.2 yr (standard deviation [$SD$] = 9.2). Of the patients, 9 were White, and 1 was Black; 7 were married, 2 were widowed, and 1 was single. The diagnoses included 3 patients with Alzheimer’s disease; 2 with dementia not otherwise specified; and 1 each with vascular dementia, Lewy body disease, cognitive disorder not otherwise specified, probable Alzheimer’s, and frontotemporal–primary progressive aphasia variant. Mini-Mental State Exam (MMSE; Folstein, Folstein, & McHugh, 1975) scores ranged from 9 to 28, with a mean of 20.3 ($SD = 7.1$). Mild cognitive impairment is considered when the patient’s scores on the MMSE are 24–28; patients are further evaluated for dementia with an MMSE score below 24 (Burns, 2013). CPT scores from the initial evaluation ranged from 2.9 to 5.2, with a mean of 4.3 ($SD = 0.78$). This sample was reflective of the larger clinical population seen in the memory clinic on the basis of descriptive statistics of larger studies from the same participant pool. The intent of the study design was to capture patients with a range of cognitive abilities.

In the clinical administration of the CPT, an “O” (observed) score was agreed on for each task of the CPT by the two licensed occupational therapy raters present for the initial assessment: the test administrator and videographer. This O rating was considered the standard score against which the four expert and three novice raters would compare their scores from rating the videos. The videotaped cases were uploaded into a secure, encrypted, password-protected program and were piloted with Rater 1 to test remote access and to add clarifying information needed to score the tasks. For example, additional information was needed for the dressing task to identify whether the raincoat was lined or unlined or gender specific.

Expert video raters were four occupational therapists recruited locally and selected on the basis of formal training on the CPT and years of experience in administering the CPT with older adults. Each expert rater reported administering the CPT more than 300 times and had a range of 5–25 yr of experience in this area of practice. Expert raters were women with an age range of 41–54 yr. Novice raters were three occupational therapy graduate students in their final didactic semester of the program; they had 1 hr of classroom introduction to the test and were given a test manual to read. The raters were granted access to the encrypted URLs and were sent instructions and scoring sheets from the test manual via email. Raters scored each video case independently. Deidentified scoring sheets were returned to the principal investigator.

Using IBM SPSS Statistics (Version 19; IBM Corporation, Armonk, NY), we analyzed the raters’ scores with Spearman’s $\rho$ and ICCs. Spearman’s $\rho$ is used to measure the strength between two ranked variables considering relative position of the scores. ICCs are used to determine consistency or reproducibility of quantitative measurements made by different observers measuring the same quantity (Kielhofner, 2006) and take into account differences in ratings of each case as well as correlation between raters.

Results

The CPT showed high interrater reliability among all raters (Table 1) as measured by ICC (.93). Coefficients were higher with experienced raters (ICC = .97) compared with novice raters (ICC = .93). It is presumed that for clinical measurements, the ICC should exceed .90 for valid measures (Portney & Watkins, 2009). Spearman’s $\rho$ correlation coefficients (Table 2) were high among all raters ($rs = .92–1.00$). The coefficients demonstrated agreement for the total CPT test scores for each case by each rater. This study met established ways to determine agreement among and between raters: The raters agreed...
with the O (clinically observed) rating, and raters agreed with one another. CPT total scores on individual cases revealed a range of 0.0–0.4 for expert raters and 0.0–0.7 for novice raters (Table 3).

CPT scores on individual tasks varied from 0.1 to 0.6 among expert raters and from 0.0 to 0.7 among novice raters. Analysis of the individual task ratings showed that the medication box task and the hand washing task had the highest duplicate scores among raters, with 8 of the 10 cases having identical task scores. The shopping for a belt task and using a map task had the lowest duplicate scores, with only 4 of the 10 cases scoring identically among the raters. The shopping for a belt task had the highest discrepancy in scoring, with a 1.5-point spread among raters.

### Discussion

The CPT showed high interrater reliability when administered to older adults with neurocognitive disorders. This supports prior study results on the CPT demonstrating high confidence in the test scores even with an expansion of the test and substituted tasks (Bar-Yosef et al., 1999; Burns et al., 1994). This study is the third interrater reliability test and substituted tasks (Bar-Yosef et al., 1999; Burns et al., 1994). This study is the third interrater reliability study conducted on the CPT that elevates the category from adequate to excellent in interrater reliability for this psychometric. Replicability of psychometrics of standardized assessments strengthens trust in the test results among many health care disciplines that may use the CPT scores to guide care management.

This study strengthens the evidence on reliability by the increased number of independent raters on a selection of cases representing the spectrum of the disease. This study supports accuracy of test results with patients early in the disease and later in the progression of cognitive decline. The range in background experience and training the raters had with the test demonstrated that even novice practitioners can have some degree of confidence in their skills to administer the test. Although expertise in practice is in the interpretation of the test score and use of the score to guide intervention, a trusted assessment result will provide those working with older adults with neurocognitive disorders a starting point for care management. A limitation of the study is the lack of random selection of raters. To modify the impact of this limitation, we selected raters on the basis of a range of experience, from novice to expert, to reflect the range of experience in a typical clinical setting. A second limitation of the study based on raters’ experience is that video raters had access to view the video of the assessment more than once; in real-time administration and scoring, there is no rewind. Because the test is administered and scored in real time, novice practitioners may need experience to gain keen observational skills as they would with most standardized assessments.

In future studies, researchers could examine specific tasks in the test to improve overall reliability. The task that was the most difficult to rate was the shopping for a belt task, which had identical ratings in only 4 of the 10 cases and a spread of 1.5 points among raters. This could have been due to the complexity of the task and multistep instructions. The task scoring procedure deserves review because the difference between a 3.5 rating and a 5.0 rating indicates a broad range of cognitive functional ability in the patient tested. It should be noted that a single task score when averaged with six other data points mildly skews the total score, resulting in minor impact on the interpretation of the total assessment results (Burns, 2013). Even so, improving the scoring on a single task will improve confidence in the testing results.

Future studies that could increase the rigor of the CPT include replicating test–retest reliability, expanding predictive validity for functional performance in IADLs and ADLs, and conducting longitudinal studies examining the rate of cognitive decline over time. Although the CPT has been studied with older adults with neurocognitive disorders, primarily those with Alzheimer’s disease and related disorders, it has not been validated with patients with stroke, traumatic brain injury, or psychiatric disorders. The CPT should be validated with other populations.

### Table 1. Intraclass Correlation Coefficients of Total Cognitive Performance Test Scores for Expert and Novice Raters Across 10 Cases

<table>
<thead>
<tr>
<th>Rater Group</th>
<th>Intraclass Correlation Coefficient</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>All raters (n = 8)</td>
<td>.929</td>
<td>[.848, .978]</td>
</tr>
<tr>
<td>Expert raters (n = 5)</td>
<td>.967</td>
<td>[.918, .991]</td>
</tr>
<tr>
<td>Novice raters (n = 3)</td>
<td>.931</td>
<td>[.821, .981]</td>
</tr>
</tbody>
</table>

### Table 2. Spearman’s ρ Correlation Coefficients for Total Cognitive Performance Test Scores for Rater “O,” Expert Raters, and Novice Raters for 10 Cases

<table>
<thead>
<tr>
<th>Rater</th>
<th>“O”</th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
<th>E4</th>
<th>N1</th>
<th>N2</th>
<th>N3</th>
</tr>
</thead>
<tbody>
<tr>
<td>“O”</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>E1</td>
<td>.948</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>E2</td>
<td>.985</td>
<td>.952</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>E3</td>
<td>.960</td>
<td>.982</td>
<td>.945</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>E4</td>
<td>.985</td>
<td>.964</td>
<td>.988</td>
<td>.945</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>N1</td>
<td>.942</td>
<td>.976</td>
<td>.915</td>
<td>.985</td>
<td>.927</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>N2</td>
<td>.948</td>
<td>1.000</td>
<td>.952</td>
<td>.982</td>
<td>.964</td>
<td>.976</td>
<td>.939</td>
<td>.964</td>
</tr>
<tr>
<td>N3</td>
<td>.948</td>
<td>.964</td>
<td>.939</td>
<td>.933</td>
<td>.976</td>
<td>.939</td>
<td>.964</td>
<td>—</td>
</tr>
</tbody>
</table>

Note. E = expert; N = novice; “O” = observed (the standard score against which the four expert raters and three novice raters would compare their scores from rating the videos).
experiencing cognitive changes over time before it is used clinically with those patients. Assessments that are altered or revised or used with different populations need repeat studies in reliability and validity to ensure accuracy and rigor.

With a renewed focus on ecological validity, occupation-based assessments are gaining interest across disciplinary lines (Gillen, 2011). Cognitive–functional assessment results that hold strong validity and reliability are the starting point for the health care team in planning and managing care. Health care decisions are determined with information from the CPT assessment, such as safety risks to return home (Douglas et al., 2013) or timing of the need for 24-hr supervision (Levy & Burns, 2011). More studies in which researchers use the CPT in guiding health care management decisions are warranted, such as use of the CPT to gauge ability in driving, financial management, medication management, meal preparation, and home management in older adults with neurocognitive disorders. Those involved in health care decision-making processes need reliable and valid tools to support clinical judgments and recommendations.

**Implications for Occupational Therapy Practice and Research**

The findings of this study have the following implications for occupational therapy practice and research:

- Health care practitioners can be confident that CPT scores by experienced and novice raters will be reproducible, giving accurate and consistent information to the health care team, family members, and patients regarding cognitive–functional capacity. The CPT could be tested with broader populations that are experiencing cognitive decline to validate application of the results in designing interventions or formulating recommendations in older adults with conditions beyond Alzheimer’s disease and associated disorders.
- Predictive validity studies on the CPT and IADL performance would increase application of the assessment to practice.

**Acknowledgments**

We thank the occupational therapy raters for their collaboration and their time and expertise.

**References**


Vaskinn, A., Sergi, M. J., & Green, M. F. (2009). The challenges of ecological validity in the measurement of social perception in schizophrenia. *Journal of Nervous and Mental Disease, 197,* 700–702. http://dx.doi.org/10.1097/NMD.0b013e3181b3ae62