Neuropsychological Rehabilitation: An International Journal

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/pnrh20

Reliability of the Cooking Task in adults with acquired brain injury

Frédérique Poncet, Bonnie Swaine, Chantal Taillefer, Julie Lamoureux, Pascale Pradat-Diehl & Mathilde Chevignard

a APHP, Service de Médecine Physique et de Réadaptation. Hôpital Pitié-Salpêtrière, Paris, France
b Équipe de recherche ER6 UPMC (Université Pierre et Marie-Curie) Physiologie et physiopathologie de la motricité chez l'homme, Paris, France
c Centre for Interdisciplinary Rehabilitation Research of Greater Montreal (CRIR), Montréal, Canada
d Université de Montréal, Montréal, Canada
e Department of Rehabilitation for Children with Acquired Brain Injury. Hôpitaux de Saint Maurice, Saint Maurice, France

Published online: 10 Nov 2014.

To cite this article: Frédérique Poncet, Bonnie Swaine, Chantal Taillefer, Julie Lamoureux, Pascale Pradat-Diehl & Mathilde Chevignard (2014): Reliability of the Cooking Task in adults with acquired brain injury, Neuropsychological Rehabilitation: An International Journal, DOI: 10.1080/09602011.2014.971819

To link to this article: http://dx.doi.org/10.1080/09602011.2014.971819

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the “Content”) contained in the publications on our platform.
However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions
Reliability of the Cooking Task in adults with acquired brain injury

Frédérique Poncet1,2,3,4, Bonnie Swaine3,4, Chantal Taillefer1, Julie Lamoureux4†, Pascale Pradat-Diehl1,2, and Mathilde Chevignard2,5

1APHP, Service de Médecine Physique et de Réadaptation. Hôpital Pitie-Salpêtrière, Paris, France
2Équipe de recherche ER6 UPMC (Université Pierre et Marie-Curie) Physiologie et physiopathologie de la motricité chez l’homme, Paris, France
3Centre for Interdisciplinary Rehabilitation Research of Greater Montreal (CRIR), Montréal, Canada
4Université de Montréal, Montréal, Canada
5Department of Rehabilitation for Children with Acquired Brain Injury. Hôpitaux de Saint Maurice, Saint Maurice, France

(Received 25 September 2013; accepted 30 September 2014)

Correspondence should be addressed to Frédérique Poncet, Ph.D., Service de Médecine Physique et de Réadaptation, Hôpital Pitie-Salpêtrière, Paris, France. E-mail: frederique.poncet@umontreal.ca

†Current address: Centre for Interdisciplinary Rehabilitation Research of Greater Montreal (CRIR), Montréal, Canada.

The first author received scholarship funding from France Traumatisme Crânien-Société Française de Médecine Physique et de Réadaptation, the Fondation des Gueules Cassées, the Fonds Européen et Francophone pour le Développement de la Recherche en Ergothérapie, the Région Ile de France (France) as well as from the Université de Montréal and the Centre for Interdisciplinary Research in Rehabilitation of Greater Montreal (Canada).

Findings from this paper were presented at the Ninth World Congress on Brain Injury, in Edinburgh, Scotland, in March 2012, published in the proceedings of the Ninth World Congress on Brain Injury (Brain Injury, 2012, p. 606) and presented at the Provincial rehabilitation network’s meeting in Montréal, Quebec, Canada, on 11 May 2012.

This work was performed by clinicians in a clinical setting, and was supported by the Hôpital Pitie Salpêtrière (France). We thank all the occupational therapists of the rehabilitation team who were involved in the project. We thank Heather Owens for the English corrections. And we thank the patients for their valuable participation in the study.

© 2014 Taylor & Francis
Acquired brain injury (ABI) often leads to deficits in executive functioning (EF) responsible for severe and long-standing disabilities in daily life activities. The Cooking Task is an ecological and valid test of EF involving multi-tasking in a real environment. Given its complex scoring system, it is important to establish the tool’s reliability. The objective of the study was to examine the reliability of the Cooking Task (internal consistency, inter-rater and test–retest reliability). A total of 160 patients with ABI (113 men, mean age 37 years, $SD = 14.3$) were tested using the Cooking Task. For test–retest reliability, patients were assessed by the same rater on two occasions (mean interval 11 days) while two raters independently and simultaneously observed and scored patients’ performances to estimate inter-rater reliability. Internal consistency was high for the global scale (Cronbach $\alpha = .74$). Inter-rater reliability ($n = 66$) for total errors was also high (ICC = .93), however the test–retest reliability ($n = 11$) was poor (ICC = .36). In general the Cooking Task appears to be a reliable tool. The low test–retest results were expected given the importance of EF in the performance of novel tasks.

**Keywords:** Acquired brain injury; Ecological assessment; Executive function; Everyday life; Reliability.

**INTRODUCTION**

Acquired brain injury (ABI) often leads to executive function (EF) deficits responsible for severe and longstanding disabilities in daily life activities. These deficits include difficulties in initiating, planning, organising, strategising and managing time and space, particularly when the environment is novel or changing (Mazaux et al., 1997). Brain injury may also cause changes in behaviour and personality which, when combined with problems of EF, can have a negative impact on a person’s activities, participation and family dynamics (Bayen et al., 2012; Brooks & McKinlay, 1983).

Traditionally, EF disorders are assessed by neuropsychological tests with strong psychometric properties (validity and reliability), i.e., Tower of London Test (Shallice, 1982) or the Trail Making Test (Reitan, 1958). Patients are evaluated while performing traditional neuropsychological tests in a quiet office setting, free from distractions. Standardised tests of EF may not always be sensitive enough for cognitive disorders following brain injury (Crépeau & Scherzer, 1993; Hart et al., 2003; Vilkki et al., 1994). Some patients exhibit severe difficulties in daily life activities, although they demonstrate normal performance on cognitive tests of executive functioning (Chevignard et al., 2000; Cripe, 1998; Eslinger & Damasio, 1985; Fortin, Godbout, & Braun, 2003; Goldstein, Bernard, Fenwick, Burgess, & McNeil, 1993; Lezak, 1982; Shallice & Burgess, 1991). Consequently, in some cases, traditional neuropsychological tests may have limited clinical use because
of their inability to reflect the true impact of EF disorders in everyday life (Eslinger & Damasio, 1985; Lezak, 1982; Shallice & Burgess, 1991).

Indeed there is a growing consensus on the importance of assessing the impact of EF in ecologically-valid environments (i.e., naturalistic settings close to real life) and to use a real task of everyday life. The Multiple Errands Test (MET), developed by Shallice and Burgess (1991), is a good example of a test administered in a natural complex setting. The participant must perform a list of specific tasks within a particular timeframe (e.g., purchase a card, buy a stamp and mail an envelope, arrive at a location at a specific time) within a real-life unfamiliar setting. However the MET has limitations in its applicability. For example, execution time is long (i.e., one hour administration plus the time to go to the shopping mall), it must be adapted to each clinical setting based on the surrounding business environment (shops, shopping centres), and it is difficult to administer in rural areas without a shopping mall. Because the test can be administered in different malls, it is difficult to compare results across studies. Other measurement tools use kitchen-based activities to assess EF. However, they are often highly structured and use simple and short activities. For example, in the Executive Function Performance Test (EFPT; Baum et al., 2008), the task involves preparing oatmeal, putting the oatmeal in a bowl and adding hot water. The Naturalistic Action Test (Schwartz, Segal, Veramonti, Ferraro, & Buxbaum, 2002) is also very structured and asks participants to make a slice of toast with butter and jam, and instant coffee with cream and sugar. Further, it is often challenging for the assessor to evaluate planning and on-line monitoring of action (e.g., error detection and correction). Other tools consider EF but assess independence, for example in the EFPT (Baum et al., 2008) the tasks (e.g., paying a bill) have EF components (e.g., initiation, organisation, etc.) but task performance is scored according to levels of assistance needed. The Cooking Task is a standardised and ecological test of EF performed in a real environment, involving multi-tasking, which was developed to address the limits of the neuropsychological assessments (Damasio, Tranel, & Damasio, 1991; Eslinger & Damasio, 1985). It specifically evaluates the consequences of EF disorders during an activity of daily living which is important for independent living. Since this task is required as part of everyday life, it should be highly generalisable. The Cooking Task requires making a cake and an omelette (e.g., different sub-tasks, such as melting butter and chocolate in the microwave) and is thus less structured in its administration procedure than the EFPT or Naturalistic Action Test. Moreover, given the test is administered in an occupational therapy kitchen, the novel environment might make the task more complex and therefore more sensitive to subtle EF deficits than other tests to assess EF during an activity.

Like the Multiple Errands Test (Alderman, Burgess, Knight, & Henman, 2003; Shallice & Burgess, 1991), the Cooking Task must be administered in an environment which is novel for the participant (i.e., the kitchen of an
occupational therapy department), providing a measure of the person’s ability to adapt to an unfamiliar environment (Cohadon, Castel, Richer, Mazaux, & Loiseau, 2008; Dutil, Vanier, & Lambert, 1995) while performing a cooking activity. The Cooking Task is able to detect problems in the on-line monitoring of the task performance, since the participant is “left on his/her own” to accomplish the required task. The Cooking Task is able to identify (even mild) EF disorders in patients with ABI (Chevignard et al., 2000). The error analysis used in the Cooking Task was developed following Schwartz et al.’s model (1998) of error analysis of task performance that considers different types of errors (“Additions”, “Omissions”, etc.), providing pertinent information to clinicians about the choice of rehabilitation intervention. Discriminant and concurrent validity of the Cooking Task have been established previously in two exploratory studies.

In 2000, Chevignard et al. studied EF disorders using the Cooking Task with 11 adult patients with severe ABI, suffering severe EF deficits in daily life activities, i.e., high scores on the Dysexecutive Questionnaire (DEX; Burgess, Alderman, Wilson, Evans, & Emslie, 1996). Patients were compared to a group of 10 control participants, matched for gender, education level and cooking experience/habits. The Cooking Task proved to have excellent discriminant validity, as patients were found to make significantly more errors than controls. Also the average execution time was significantly longer for patients (81.2 minutes versus 51 minutes). Concurrent validity of the Cooking Task was good with respect to the DEX Questionnaire but it correlated poorly with the traditional neuropsychological tests, e.g., Modified Wisconsin Card Sorting Test (Nelson, 1976), Trail Making Test (Reitan, 1958), and the Six Elements Task (Garnier et al., 1998; Shallice & Burgess, 1991; Wilson, Evans, Emslie, Alderman, & Burgess, 1998).

In 2008, Chevignard et al. replicated these results and studied the Cooking Task’s validity in a larger sample of 45 patients (27 men) with significant ABI. In this study, patients were included on the basis of having a moderate to severe dysexecutive syndrome identified using several neuropsychological tests of executive functioning. Patients’ performances in the Cooking Task were compared to that of a small control group (n = 12) matched for age, sex, education level and cooking habits. Discriminant validity of the task was excellent, as patients made significantly more errors (mean total number of errors = 107.5; SD = 74.3 versus 18.3; SD = 7.8) and were significantly slower than controls. In addition, the authors reported that more than half of the patients were unable to achieve the goal and demonstrated dangerous behaviours, such as leaving the stove on when the task was finished, or taking the cake out of the oven with unprotected hands. Considering concurrent validity, the results of the Cooking Task were significantly correlated with the “cognition” factor of the DEX Questionnaire (r = .573, p = .01) and with a number of neuropsychological tests of EF. The results of the regression analysis indicated that
the best predictor of the total number of errors in the Cooking Task was the score on the Six Elements Task, accounting for 26% of the variance (\( R^2 = .262, p = .0003 \)). Also the authors reported good inter-rater reliability (\( r = .862, p < .0001 \)), but without providing the details of the methods used (e.g., number of raters and rater training).

In both studies (Chevignard et al., 2000; 2008), the results demonstrated that the Cooking Task can measure EF in a naturalistic or ecologically-valid environment and appears to be more sensitive in detecting EF disorders than routine neuropsychological tests, thus supporting the discriminant validity of the tool.

Because the Cooking Task also possesses many of the characteristics of clinical tools deemed important, e.g., low cost (Auger, Demers, & Swaine, 2006; Law, Baum, & Dunn, 2001), easy to administer (Conner-Spady, Slaughter, & MacLean, 1999), and high clinical relevance (Auger et al., 2006; Conner-Spady et al., 1999; Law et al., 2001), it is increasingly used in occupational therapy departments throughout France. In the course of the past four years, 80 occupational therapists (OT) have attended four-day training sessions on the use of the Cooking Task to assess EF disorders. Given the popularity of the Cooking Task in France, and the research to date on its validity, it is essential to conduct further research to evaluate the additional psychometric properties of this clinically relevant tool. Indeed the scoring procedure is complex and it is necessary to establish the reliability associated with the tool so clinicians can interpret the data and use them as part of their clinical decision making. More specifically, the internal consistency, inter-rater and test–retest reliability of the Cooking Task should be assessed in order to determine the degree to which groups of items within the Cooking Task measure a single unidimensional construct, and to assess the errors associated with this measure.

The aim of the present study was thus to establish the internal consistency, inter-rater and test–retest reliability of the Cooking Task in a sample of individuals who had sustained an ABI. Given that EF allows adaptation to new situations (Rabbit, 1997), and that novelty is an important feature of EF assessments, we hypothesised that only test–retest reliability would be poor (because the second score would likely be better than the first due to the learning effect), unlike the internal consistency and inter-rater reliability.

**METHODS**

**Context**

This study was approved by the local ethics committee. It was conducted in a Physical Medicine and Rehabilitation department in Paris, France, providing in- and out-patient rehabilitation services to individuals who sustained an
ABI. Services are provided by a multidisciplinary team including OTs who were involved in the development of the Cooking Task and who have been using it in their clinical practice over the past 15 years (e.g., since the tool’s creation). For this study, data relating to the Cooking Task (performed for clinical purposes) were collected retrospectively and prospectively (see results). Patients meeting the inclusion and exclusion criteria (see below) were included. Results of the Cooking Task, demographic, medical and neuropsychological information were collected from the medical files.

Participants

One hundred and sixty patients having the following characteristics were enrolled in the study: (1) ABI involving either the frontal lobes or the frontal subcortical pathways on computed tomography scan and/or magnetic resonance imaging and (2) evidence of EF deficits (moderate to severe) indicated by scores that were more than two standard deviations beyond the mean scores on at least two of nine of the following neuropsychological tests: Wisconsin Card Sorting Test (WCST; Nelson, 1976), Tower of London Task (Shallice, 1982); Trail Making Test (Reitan, 1958); Verbal fluency (Benton, 1968); Digit span (WAIS-III; Wechsler, 2000); Brown-Peterson Paradigm (Van der Linden, Coyette, & Seron, 1992); Test for Attentional Performance (Zimmermann & Fimm, 1992); Rey Figure Copy (Rey, 1959); and the Six Elements Test (Wilson et al., 1998).

Exclusion criteria were (1) pre-existing psychiatric or neuropsychological disorders, or alcohol or drug abuse; (2) intellectual deterioration (performance cut off of 122 on the MATTIS scale, Mattis, 1998); and (3) sensory-motor deficits, aphasia, linguistic deficits or unilateral spatial neglect likely to interfere with the understanding of the instructions and recipe, or with the Cooking Task itself. The Cooking Task was performed by the patients as part of their rehabilitation programme.

Tools and scoring procedure

Participants were asked to perform the Cooking Task as per its standardised procedure. They were to bake a chocolate cake (using a recipe provided) and cook an omelette for two people in the occupational therapy department’s equipped kitchen. Items needed to prepare the food were displayed on a table, always in the same location: (1) salt, pepper and oil; (2) all the utensils and ingredients necessary for the cooking activity; (3) some semantic distracter utensils and ingredients, unnecessary for the required task, but normally present in a kitchen; and (4) a dessert cookbook containing the recipe for the cake.

One or two raters selected from the pool of raters (see below) were present in the kitchen with the patients. One rater showed the patient all the necessary
items in the kitchen and how to use the oven, gas stove and microwave oven. Then, the rater defined the activity with the following statement: “You must bake a chocolate cake and cook an omelette for two people. The cake recipe is available in this book. All the utensils and ingredients you need are available here on the table. You must leave the kitchen the way you found it when you came in. We cannot help you; you must act as if you were alone. Tell us when you have finished.” The rater made sure the participant had understood and memorised the instructions and explained them again if necessary. The participant was also given a cue card with the written instructions and told he/she could refer to it as needed. Neither rater was allowed to answer any question and/or provide help; the rater only intervened in case of danger. One or two raters independently and simultaneously recorded all of the participant’s actions, comments and questions, and classified the errors (see scoring below) observed during the task. The maximum time allowed to complete the task was two hours. After this time, if the participant had not finished, the raters interrupted the task and it was considered as failed.

Raters and rater training
Eight OTs and two physiatrists participated as raters in this study. They were all part of the Physical Medicine and Rehabilitation team in the department where the research was conducted. They each had between 1 and 26 years (X = 12.70, SD = 9.40) of clinical experience in the treatment of patients with a neuropsychological impairment and were familiar with the use of the Cooking Task.

At the time of this study, three therapists were very familiar with EF and the Schwartz model (Schwartz et al., 1998), having participated in the development of the Cooking Task. They were well versed in the use of the tool and its scoring procedure. However, the other five raters were less familiar with the tool, and were required to participate in a brief training session (i.e., 30–60 minutes of discussion alone with one of the authors to understand the scoring system). This enabled them to become familiar with the concept being evaluated, the effect of EF disorders on activities, and to learn how to administer the tool and use its scoring system.

Scoring procedure
Performance on the task is measured by observing the type of errors occurring during the activity. Each rater/therapist individually reports on a standardised form the errors committed by the participant during the task. Each error is categorised as a descriptive error (i.e., “Omissions”, “Additions”, “Commentary/Question”, “Inversions-substitutions”, “Estimation” errors), and then as a neuropsychological error (i.e., “Control errors”, “Context neglect”, “Environmental adherence”, “Purposeless actions”, “Dependency” and “Behavioural
disorders”). See appendix for definitions and further examples of how the errors were qualified. For example, if the patient forgot to add sugar when making the cake (as per the recipe) the rater would record an “Omission” error (descriptive error type) and then also record a “Control” error (neuropsychological error type), because the patient “loses control” of the recipe. “Dependency” and “Behaviour” are complementary error types because, initially, they are both classified as “Context neglect” errors and then as either “Dependency” or “Behaviour” errors. The following example illustrates the complexity of the scoring system. Following the observation of a “Commentary/Questions” error (e.g., the participant asks where the omelette recipe is, and it is not provided), the rater would record a “Context neglect” error (because the participant should imagine he/she is alone while performing the task) and a “Dependency” error (because the participant demonstrates a certain level of dependency upon the clinician). The rater classifies the number of errors in each category (allowing the calculation of the total number of errors) and records the duration of the task (in minutes). A qualitative rating is then also performed, based on whether the participant spontaneously initiated both recipes (yes/no), his/her ability to achieve the goal (yes/no), and the occurrence of dangerous behaviours (yes/no).

For internal consistency, 155 patients were assessed by raters selected from the pool. Five of the 160 patients were not able to complete the task. According to Rabbit (1997) the study of test–retest reliability is exploratory and only 11 patients were assessed by the same rater, an OT with extensive experience with the tool selected from the pool of raters, on two occasions separated by an average of 11 days ($SD = 3$).

For inter-rater reliability, out of the 160 patients included in the study, 66 were assessed by their regular OT and by another rater randomly selected from the rater pool. The two raters simultaneously observed and independently scored each participant’s performance.

Data analysis

The internal consistency (Cronbach alpha statistic) was calculated for three aspects of the tool since the Cooking Task provides a score for each error type (descriptive errors and neuropsychological errors) and a total score for descriptive error types. Alpha coefficients of .70 or higher are usually regarded as indicative of acceptable internal reliability (De Vellis, 1991); values above .80 are conventionally considered high and values above .90 as very high. Values of .60–.69 are considered questionable, values between .50–.59 as poor, and values below .5 unacceptable (George & Mallery, 2003; Gliem & Gliem, 2003).

Test–retest reliability for this study was determined by having the same rater assess the same 12 error types (five descriptive, six neuropsychological
errors and one total error score) on two different occasions with 11 patients. Inter-rater reliability of the Cooking Task was estimated for each of the possible pairs among the 10 raters and is reported as the average values across all possible pairs. For the test–retest and the inter-rater reliability, intra-class correlation coefficients (ICC) were calculated (Dunn, 1989; Shrout & Fleiss, 1979) for each variable of the Cooking Task (the descriptive and the neuropsychological error types) and for the task duration. ICCs (one factor random) were calculated since this particular model assumes that judges are considered a fixed effect (Shrout & Fleiss, 1979) as the selection procedure of raters was random. ICC values of > .75 were considered excellent; values between .40 and .75 were considered moderate to good and values < .40 were considered poor. All analyses were performed with the SPSS 17 for Windows software package.

RESULTS

Sample characteristics

One hundred and sixty patients with ABI (113 men, mean age 37 years, \(SD = 14.3\)) participated in the study. The data for 100 participants were collected retrospectively and extracted from existing clinical forms (reporting the descriptive and neuropsychological errors) inserted in their medical chart following the evaluation of patients participating in our clinical rehabilitation programme. The data for the remaining 66 participants were collected prospectively to estimate inter- and test–retest reliabilities. Fifty-five percent suffered a traumatic brain injury (TBI) and the remainder had sustained other types of brain injury (e.g., stroke, brain tumour). The participants’ demographic variables and injury characteristics are presented in Table 1.

Completion of the Cooking Task

Five participants were unable to perform the task and were excluded from the analyses. Data for 155 participants were thus analysed. Mean duration of the task (\(n = 151/155\)) was 63.3 minutes (\(SD = 20.9\)). One hundred and thirty-eight patients (89%) were able to achieve the goal and 72 (52%) demonstrated dangerous behaviours (e.g., leaving the stove on when not required).

Internal consistency

A total of 152 participants (113 men, mean age 37 years, \(SD = 14.3\)) had complete data allowing calculation of the internal consistency of the total score (11 items or all error types) and the correlation matrix for these items is presented in Table 2. All correlation coefficients were small to
The computation of the Cronbach alpha for the descriptive errors was conducted using the complete data from 155 participants (retrospective and prospective data). Table 3 presents the correlation matrix for the five descriptive error items. The correlation coefficients were moderate and all positive (.15 to .39) except between the “Estimation” errors and the “Commentary/Questions” (.08). The Cronbach alpha for these five descriptive error types was .63.

A total of 152 participants had complete data for the neuropsychological classification of errors. Three participants could not be assessed by the raters due to the raters’ inability to accurately record this type of error. Table 4 presents the correlation matrix for these six items. All coefficients were small to moderate and positive, except between “Control errors” and “Context neglect”. The Cronbach alpha coefficient was .42 (moderate) for this group of items of neuropsychological error types. If the “Context neglect” item was removed from this reliability analysis, the alpha would reach .60, which is relatively high.
<table>
<thead>
<tr>
<th></th>
<th>Omission</th>
<th>Additions</th>
<th>Commentary/Questions</th>
<th>Inversions-substitutions</th>
<th>Estimations</th>
<th>Control errors</th>
<th>Context neglect</th>
<th>Adherence</th>
<th>Purposeless actions</th>
<th>Dependency</th>
<th>Behaviour disorders</th>
</tr>
</thead>
<tbody>
<tr>
<td>Omissions</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Additions</td>
<td>0.32</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Commentary/Questions</td>
<td>0.36</td>
<td>0.48</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inversions-substitutions</td>
<td>0.59</td>
<td>0.44</td>
<td>0.31</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimation</td>
<td>0.41</td>
<td>0.37</td>
<td>0.13</td>
<td>.32</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control errors</td>
<td>0.35</td>
<td>0.25</td>
<td>0.11</td>
<td>.30</td>
<td>0.41</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Context neglect</td>
<td>0.23</td>
<td>0.12</td>
<td>0.17</td>
<td>.19</td>
<td>0.01</td>
<td>-0.01</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adherence</td>
<td>0.33</td>
<td>0.84</td>
<td>0.40</td>
<td>.38</td>
<td>0.41</td>
<td>0.26</td>
<td>0.09</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Purposeless actions</td>
<td>0.27</td>
<td>0.72</td>
<td>0.43</td>
<td>.42</td>
<td>0.23</td>
<td>0.09</td>
<td>0.12</td>
<td>0.40</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dependency</td>
<td>0.43</td>
<td>0.41</td>
<td>0.58</td>
<td>.27</td>
<td>0.13</td>
<td>0.09</td>
<td>0.27</td>
<td>0.39</td>
<td>0.34</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Behaviour disorders</td>
<td>0.51</td>
<td>0.35</td>
<td>0.42</td>
<td>.49</td>
<td>0.26</td>
<td>0.12</td>
<td>0.11</td>
<td>0.33</td>
<td>0.25</td>
<td>0.38</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Inter-rater reliability

Sixty-six participants (43 men, mean age 39 years, $SD = 13$) had complete data for descriptive error types and among them, 62 had complete data for the neuropsychological error types. Four participants could not be assessed by the raters due to the raters’ inability to accurately record this type of error. Inter-rater reliability for descriptive error types (Table 5) ranged from good (i.e., ICC = .65 for “Omissions”) to excellent (i.e., ICC = .95 for “Commentary/Questions”). For neuropsychological error types (Table 6), ICCs ranged from average (“Purposeless actions” ICC = .49) to good (“Environmental adherence” ICC = .79), except for the “Context neglect” (ICC = .05; −.20; .29) and “Dependency” (ICC = .18; −.07; .41], which are not significantly superior to 0.

As for the task duration, the inter-rater ICC was high at .97 ($p = .024$).
Test–retest reliability

For 11 participants (5 men, mean age 32 years, SD = 13.5), the test–retest reliability varied from low (total number of errors ICC = .36) to good for the descriptive error types (ICC = .27 to .65) (Table 5) and for the neuropsychological error types (ICC = .09 to .66) (Table 6). Given the small sample size, the 95% confidence intervals (CI) were rather large (less precise). We noted, however, that 8 out of 11 patients (73%) improved their overall score over the two sessions and one patient made the same number of errors during both testing sessions.

DISCUSSION

The assessment of psychometric properties of outcome measures is an important issue. The Cooking Task was designed to be used as an ecological tool for identifying the consequences of disorders of EF after ABI. This study
examined the internal consistency, inter-rater and test–retest reliability of the Cooking Task in individuals with ABI.

The internal consistency of the three aspects of the tool appears to be good to very good, supporting the notion that the descriptive and neuropsychological error types are homogeneous (the former more than the latter) and that the tool globally assesses a unidimensional construct. It is not surprising that the internal consistency of the descriptive error type items is higher than that of the neuropsychological error type items because scores within the former group simply represent an observed objective error. Error recording of the neuropsychological type is more complex and includes double recordings for “Dependency” and “Behaviour disorders”. As such, this grouping of items is less homogeneous. Our results for the internal consistency of the total error count are equivalent to those reported for a similar task adapted for children with ABI (Chevignard, Catroppa, Galvin, & Anderson, 2010): the internal consistency of the Children’s Cooking Task was found to be high (.86) when entering all the error types in the analysis. Moreover, the internal consistency of the Cooking Task is comparable to that of the simplified version of Multiple Errands Test ($\alpha = .77$) for use in a hospital setting (Knight, Alderman, & Burgess, 2002). Although both tools are ecologically valid and reliable, we recommend using the Cooking Task because we think it is more applicable to clinical practice where OT kitchens are more readily available/accessible than a shopping mall. Moreover, the Cooking Task is less structured and thus may allow clinicians to better detect executive dysfunction.

Inter-rater reliability of the tool is acceptable and comparable to that of the Baycrest version of the Multiple Errands Test (ICC = .71–.88) (Dawson et al., 2009). The neuropsychological error types showed lower reliability estimates than the descriptive error types and this is probably due to the subjectivity inherent to the classification of the former. Another probable reason for the lower reliability for the neuropsychological error types may be linked to the complexity of the scoring. Indeed, for each error recorded, raters must simultaneously assess the error within the two error classifications, record the descriptive error type first and then rate the neuropsychological type of the error, while still observing the participant for additional errors. Raters less experienced with the tool may have found this difficult. If the assessments had been videotaped as proposed originally, the raters could have stopped the film to accurately assess the participants’ performance and record all errors. We felt it was more appropriate to evaluate the reliability associated with the tool as it was administered in the clinical setting (i.e., not based on video-taped performance). With regard to the error types having the lowest inter-rater reliability (Context neglect and Dependency), it is possible that the complex recording system is associated with an important degree of error. It is clear that clinicians who use this tool need training with the scoring
system (and perhaps a guide with examples for each type of error) to ensure the reliability of their measurement.

We acknowledge that the test–retest portion of this study involved only a small sample and should thus be interpreted with caution. If similar test–retest results are corroborated with a larger sample, it could be interesting to create an alternative version of the Cooking Task to limit learning during repeated administrations of the test. The lower test–retest results were expected given the importance of task novelty in EF assessment (Denckla, 1994; Phillips, 1997; Shallice & Burgess, 1991; Wilson et al., 1998). In other words, participants’ performance improved (i.e., was not identical) the second time since neither the task nor the environment were novel any more. Given these circumstances, one might question the value of calculating test–retest for this task. Interestingly, in a study using a similar task adapted for children in a group of children with TBI of similar size ($n = 9$), test–retest reliability was found to be very good for the total number of errors ($ICC = .89$) (Chevignard et al., 2010). This difference could be partly explained by the longer mean delay between the two testing sessions in that study (48 days, $SD = 21$), versus 11 days in our study. Given the effect of task novelty on performance, and the way some patients simply remembered the errors they did on the first occasion, one could expect a decrease of the test–retest effect with a longer delay between two sessions. Moreover in the child study, some children explicitly mentioned errors they had done the first time (such as using the wrong recipe), said they would not repeat them, and indeed they did not but they made numerous other errors during task performance. Therefore the test–retest reliability was good for the total number of errors, but not for every error type.

**Study limitations**

This study has limitations. The use of retrospective data may have caused transcription errors influencing the reliability estimates; however, data transcription was validated by another person with each extraction from the medical chart. The sample size for the test–retest analyses was very small and as such, these results should be considered exploratory, especially given the discrepancy found with a study using a similar task in children. Until additional research is conducted on the Cooking Task’s test–retest reliability with a larger sample and with a slightly longer delay (1 month) to reduce the effect of patients memorising their most striking errors (e.g., uses incorrect recipe), we recommend the use of the tool to detect the consequences of EF disorders on performance of a daily living activity, rather than to demonstrate improvements (i.e., change over time) in patients with ABI. Additional test–retest research could investigate patients tested in two different kitchens, thus ensuring that both testing environments are novel.
Nevertheless, the overall sample size was very large, and inter-rater reliability was performed with more than half of the sample, which strengthens the results. It may be of interest to examine the inter-rater reliability of the Cooking Task following an improved training of the use of this tool. Finally, developing an alternative form of the test could help distinguish clinical improvement from practice effects in the context of test–retest reliability.

Conclusion and clinical implications

The results of this study provide some important information regarding the homogeneity of the error types within the Cooking Task and the reliability associated with this tool when administered by trained clinicians to patients with ABI. Firstly, the internal consistency results support the use of the Cooking Task to assess EF disorders. Given the tool’s moderate test–retest reliability, we recommend the tool be used to detect the effects of EF disorders during a daily life activity rather than to demonstrate improvements (i.e., change over time) in patients with ABI. In conclusion, these results allow OTs to better interpret the measurement associated with the Cooking Task in view of its reliability and thus make appropriate intervention choices. Clinicians interested in using the Cooking Task with persons with EF disorders must ensure they have a good understanding of EF disorders in general, and that they are trained in the administration and complex scoring of the tool.

REFERENCES


APPENDIX
TYPES OF ERRORS IN THE ANALYSIS OF THE COOKING TASK

As a first step, the errors were classified at a descriptive level:

1. **Omission**: Any action or sequence of actions necessary to reach the goal which is omitted or incompletely performed, such as not washing hands at the beginning of the activity, forgetting an ingredient or an instruction stated in the recipe, forgetting to turn on the oven, not mixing the ingredients together, or not washing hands even though they are covered with butter.

2. **Addition**: Any action or sequence of actions unnecessary to the completion of the task, such as using distractor ingredients, opening the drawers or closets or the laundry washing machine, picking up an object and putting it down without using it, or any purposeless displacement that are unrelated to the task.

3. **Inversion-substitution**: Any action performed out of the appropriate temporal sequence or any object that is misused or inappropriate to the sub-goal, such as putting a frying pan in the microwave oven, or a salad bowl in the oven, or not following the order of ingredients in the recipe, or putting a spoonful of chocolate into the package of flour.

4. **Estimation error**: Poor estimation of the quantity of ingredients, of the size of an object, of space or time, such as putting too much sugar or too many eggs in the cake, putting the salad bowl on the cookbook, cutting the right amount of butter, then putting the whole plate of butter in the bowl, cooking the cake or the omelette too much or not enough.

5. **Commentary-question**: Any question, remark or joke to the examiners, although the patient had been clearly instructed to act as if he or she were alone, such as asking how to perform an action, or where to find an ingredient or a utensil.
Next, the errors were classified again, this time taking into account the neuropsychological mechanisms underlying the occurrence of errors.

1. **Context neglect**: Failure to respect the instructions or the recipe, poor evaluation of the environment, such as not washing hands at the beginning of the activity, not respecting the quantities in the recipe, not taking the recipe into account, not using the utensils or ingredients prepared for the patient on the working space and searching for them in closets and drawers, or not washing hands even though they are covered with butter, putting a spoonful of chocolate into the package of flour, or all the commentaries and questions, since patients were informed they were supposed to act as if alone.

2. **Environmental adherence**: Inappropriate action induced by the presence of an object, such as washing a utensil or the sink several times, mixing the dough again and again without being able to stop, putting the bowl on the recipe book, cutting the right amount of butter, then putting the whole plate of butter in the bowl, or washing utensils present in the kitchen that have nothing to do with the task.

3. **Purposeless actions and displacements**: Behavioural sequence not contributing to goal achievement, such as not doing anything for a long time, picking up an object and putting it down again without using it, any purposeless displacement, choosing to use an inappropriate item or utensil.

4. **Dependency**: Any question about the way to perform an action or to find an object (this does not include the other commentaries), such as asking how or what to do, where to find the recipe, a utensil, an ingredient or the oven, or how to turn on the gas stove.

5. **Behavioural disorder**: Any socially inappropriate or dangerous behaviour, such as taking the cake out of the hot oven or the omelette from the frying pan with the fingers without protection, or letting the gas run without lighting it.