Development and Preliminary Reliability of a Multitasking Assessment for Executive Functioning After Concussion

Laurel B. Smith, Mary Vining Radomski, Leslie Freeman Davidson, Marsha Finkelstein, Margaret M. Weightman, Karen L. McCulloch, Matthew R. Scherer

MeSH TERMS

- brain concussion
- executive function
- military personnel
- reproducibility of results
- task performance and analysis

OBJECTIVES. Executive functioning deficits may result from concussion. The Charge of Quarters (CQ) Duty Task is a multitask assessment designed to assess executive functioning in servicemembers after concussion. In this article, we discuss the rationale and process used in the development of the CQ Duty Task and present pilot data from the preliminary evaluation of interrater reliability (IRR).

METHOD. Three evaluators observed as 12 healthy participants performed the CQ Duty Task and measured performance using various metrics. Intraclass correlation coefficient (ICC) quantified IRR.

RESULTS. The ICC for task completion was .94. ICCs for other assessment metrics were variable.

CONCLUSION. Preliminary IRR data for the CQ Duty Task are encouraging, but further investigation is needed to improve IRR in some domains. Lessons learned in the development of the CQ Duty Task could benefit future test development efforts with populations other than the military.

Smith, L. B., Radomski, M. V., Davidson, L. F., Finkelstein, M., Weightman, M. M., McCulloch, K. L., & Scherer, M. R. (2014). Brief Report—Development and preliminary reliability of a multitasking assessment for executive functioning after concussion. *American Journal of Occupational Therapy*, 68, 439–443. <http://dx.doi.org/10.5014/ajot.2014.012393>

Laurel B. Smith, MS, OTR/L, is Captain, U.S. Army, and Research Occupational Therapist, U.S. Army Research Institute of Environmental Medicine, 15 Kansas Street, Natick, MA 01760; laurel.b.smith.mil@mail.mil

Mary Vining Radomski, PhD, OTR/L, is Clinical Scientist, Courage Kenny Research Center, Minneapolis, MN.

Leslie Freeman Davidson, PhD, OTR/L, is Director and Associate Professor of Occupational Therapy, Shenandoah University, Winchester, VA.

Marsha Finkelstein, MS, is Senior Scientific Advisor, Courage Kenny Research Center, Minneapolis, MN.

Margaret M. Weightman, PhD, PT, is Clinical Scientist/Physical Therapist, Courage Kenny Research Center, Minneapolis, MN.

Karen L. McCulloch, PhD, PT, NCS, is Professor, Division of Physical Therapy, University of North Carolina at Chapel Hill.

Matthew R. Scherer, PhD, PT, is NCS Major, U.S. Army, and Chief of Physical Therapy, Andrew Rader U.S. Army Health Clinic, Fort Myer, VA.

Concussion has received unprecedented attention in the military because of the increased incidence in the past decade (Helmick, Baugh, Lattimore, & Goldman, 2012) and has been called the “signature injury” of the conflicts in Iraq and Afghanistan (McCrea et al., 2009, p. 1369). Concussion may result in symptoms including headache, dizziness, nausea, sensitivity to noise and light, slowed thinking and reaction time, memory problems, difficulty concentrating, executive dysfunction, and visual and balance changes (Carroll et al., 2004). Although subtle and sometimes difficult to detect, these multisensory symptoms can negatively affect job performance and safety in servicemembers.

Army occupational therapists play key roles in evaluating servicemembers and making recommendations regarding their ability to return to duty after concussion. Currently, occupational therapy practitioners rely on self-reported symptoms and

vestibular and neuropsychological assessments to determine duty readiness. However, subjective symptom report does not always coincide with clinical recovery (Vagnozzi et al., 2008), and neuropsychological assessment batteries do not always predict real-world functioning, especially after a combat experience (Brenner et al., 2010). Accurate assessment is further limited by measures with ceiling effects or minimal sensitivity to concussion-related deficits.

Multitask assessments may be more sensitive to subtle performance deficits because they replicate the simultaneous cognitive and sensorimotor demands of unstructured, complex real-world activities (Frisch, Förstl, Legler, Schöpe, & Goebel, 2012). Despite the potential benefit of this assessment approach and alignment with priorities for occupational therapy evaluation, few options exist that have satisfactory reliability, validity, and clinical utility (Dawson et al., 2009). The

Multiple Errands Test (MET; Shallice & Burgess, 1991) is an example of a multitask assessment of executive functioning based on five demands of multitasking: (1) performing multiple but discrete tasks that vary in priority, complexity, and length; (2) managing interleaving and dovetailing tasks; (3) performing tasks without feedback; (4) dealing with interruptions, reprioritization, and rule changes; and (5) self-initiating task changes within the activity (Burgess, 2000). The many versions of the MET involve completing at least 10 unrelated tasks while complying with a series of rules in either a shopping mall or hospital lobby setting (Alderman, Burgess, Knight, & Henman, 2003; Cuberos-Urbano et al., 2013; Dawson et al., 2009; Morrison et al., 2013). Although the MET appears to assess “the central aspects of executive functioning in everyday life” (Frisch et al., 2012, p. 257), it has yet to be widely adopted in clinical practice because of site-specific validation requirements, time-intensive administration, and a lack of standardized scoring manuals specific to each site (Radomski & Morrison, 2014).

A team of military and civilian occupational and physical therapists are currently developing a performance-based assessment battery called the Assessment of Military Multitasking Performance (AMMP; Radomski et al., 2013). The AMMP includes six dual- and multitask assessments designed to assess concussion-related deficits. If proven reliable and valid, the AMMP will be used by military occupational and physical therapists to determine duty readiness for servicemembers after concussion.

The Charge of Quarters (CQ) Duty Task (CQDT) was developed as one of the assessments included in the AMMP battery that uses the structure of the MET to assess executive functioning. CQ duty is an additional duty in the military during which servicemembers are responsible for 24-hr supervision and security of a facility; servicemembers on CQ duty are frequently tasked with various assignments that are unstructured and unrelated in nature. This scenario provides a realistic backdrop for the multitask assessment given the reality of task demands and

face validity among servicemembers. This article describes the rationale and development process of the CQDT and presents pilot data from the preliminary evaluation of interrater reliability (IRR).

Description of the Charge of Quarters Duty Task

In the CQDT, as in the MET, participants receive in-depth instructions and a written list of assignments and performance rules. They are required to visit four different hypothetical work areas (marked with duct tape): (1) the CQ desk, (2) the bulletin board, (3) the supply closet, and (4) the assembly area, each containing the information and resources necessary to complete their assignments. They are encouraged to keep transits between work areas to a minimum (seven or fewer) and are told to revisit an area only if necessary to complete the task. Task assignments include reporting a CQ duty shift change, assembling a footstool from PVC pipe, reporting the number of vacant rooms in the barracks (living quarters for servicemembers) using a barracks layout, conducting an inventory of PVC supplies, obtaining the address of the manufacturer of the footstool materials, locating the telephone number of another servicemember using a personnel roster, and locating the room of a specified servicemember using a map of a barracks layout.

During the exercise, participants must adhere to four rules: (1) Assemble the footrest only in the assembly area, (2) bring only the number of PVC parts needed for the footrest to the assembly area, (3) do not move or remove any of the materials from the walls in any of the work areas, and (4) do not speak to the examiners during the assessment. Throughout the task, participants must also deal with interruptions and reprioritization of tasks. Scoring metrics borrowed from the MET include accuracy of task performance (Cuberos-Urbano et al., 2013; Dawson et al., 2009; Morrison et al., 2013), total rule breaks (Cuberos-Urbano et al., 2013; Dawson et al., 2009; Morrison et al., 2013), frequency of rule breaks (Dawson et al., 2009), transits between work areas (Morrison et al., 2013), and total performance time.

Method

Instrument Development

The CQDT was developed as part of the AMMP battery. The initial version of the AMMP included five multitask assessments and three dual-task assessments (Radomski et al., 2013). After initial pilot testing of the AMMP battery, data analysis indicated variable IRR (intraclass correlation coefficients [ICCs] of .45, .37, and .79 for task performance) for the three multitask assessments of executive functioning. Scoring was complicated by errors resulting from simultaneous observation and scoring requirements and by a lack of clearly defined scoring criteria outlining acceptable tolerances for partially accurate task performance. For example, when participants were told to obtain an address, rater disagreements occurred if part of the address was incorrect (e.g., transposed digits, spelling errors); some examiners gave full credit for task completion and others gave no credit. In addition to multiple scoring challenges, test developers indicated substantial test burden from three relatively similar multitask assessments and limited face validity of the tasks as reported by participants. In an effort to improve IRR, face validity, and clinical feasibility, the CQDT was developed to replace the three previous iterations of multitask assessments.

The first step in the development of the CQDT was to reexamine the literature pertaining to current multitask assessments. The team also shared the initial concept, materials, and instructions of the CQDT with a panel of experienced servicemembers who provided recommendations to improve face validity of the task with the target population. On the basis of the definition of multitasking (Burgess, 2000) and feedback from subject matter experts, the team created a list of parameters to be tested.

Once the initial task was developed, test developers practiced administering the task on servicemembers and civilians to observe variations in performance and variations in the interpretation of performance by multiple evaluators. After practice administrations, test developers clarified task instructions and revised the approach

to scoring by creating operational definitions that clarified situations in which no credit, partial credit, or full credit should be given. These operational definitions were included on the score sheet. For example, a participant who reported the incorrect number of barracks rooms would receive partial credit for task performance in that domain as determined by the operational definition for that task. This scoring approach reduced scoring complexity and allowed raters to assign a score quickly upon observation of task completion.

The score sheet was also improved to reduce scoring errors resulting from simultaneous observation and scoring requirements. Many aspects of the CQDT required scoring in real time (i.e., radio communications with various personnel on the correct radio frequency) to determine whether participants completed tasks independently and accurately or required cueing. Raters who were distracted or who failed to score performance on these tasks immediately made scoring errors. To address this issue, task assignments were listed chronologically on the score sheet, and tasks requiring immediate scoring were emphasized with bold font. This design helped cue the evaluators to ensure observation of performance at appropriate times. Last, the score sheet included correct responses for objective performance components (e.g., correct number of vacant barracks rooms to be reported, manufacturer's address), allowing the rater to quickly identify performance accuracy and assign the appropriate score. These additions were implemented to maximize scoring efficiency.

After all modifications were made to the CQDT, test developers piloted the revised multitask assessment in a healthy population to assess IRR. Given the anticipated variability in task performance between healthy servicemembers and those with concussion, evaluation of IRR in healthy servicemembers allowed for subsequent scoring and procedural refinements to be made before evaluating IRR in servicemembers with concussion.

Intrarater Reliability Testing

Preliminary IRR was assessed between 3 (2 trained and 1 novice) raters when measuring individual participant performance

on the CQDT. The two trained raters were involved in test development, and the novice rater was a physical therapist with no prior experience with the CQDT. This design helped determine whether inexperienced providers could easily and accurately score the assessment. Before evaluating participants, the novice rater received a brief orientation (<30 min) to the score sheet, performance metrics, and operational definitions of task performance, rules, and rule breaks. IRR was established for all raters.

Participants

Participants were recruited by convenience sampling from the U.S. Army Research Institute of Environmental Medicine in Natick, Massachusetts. All healthy active-duty servicemembers (active duty, guard, or reserve component) ages 18–42 yr were eligible to participate. Participants were excluded if they reported a history of traumatic brain injury (TBI) or concussion in the previous year, any documented active-duty restrictions (currently on a military profile), any physical or behavioral health condition preventing sustained activity for up to 30 min, history of psychiatric disorder, and uncorrected hearing deficits. All participants gave written informed consent before participation, and the institutional review board at the U.S. Army Research Institute of Environmental Medicine approved the study.

Data Collection

The following components were measured via observation:

- *Task completion* was defined as the extent to which participants independently and accurately completed each assignment. Each assignment was scored 0 (*not complete*), 1 (*partially complete or required cueing to complete*), or 2 (*completed to defined standard independently without cueing*). The test included 17 assignments (some assignments required more than one task), with up to 2 points possible for each, for a total of 34 possible points for task completion.
- *Total rule breaks* for the four rules were operationally defined on the score sheet. Each rule that was broken was recorded.

- *Frequency of rule breaks* was recorded for each rule; it was possible to break the same rule multiple times. No limit was placed on the frequency of rule breaks.
- *Performance time* was defined as the total time to complete the task.
- *Transits* were defined as movements between work areas. Leaving one work area and entering another was considered one transit.

Data Analysis

The ICC was used to quantify preliminary IRR. The Krippendorff (Hayes & Krippendorff, 2007) α macro was run under SPSS Version 18.0 (IBM Corporation, Armonk, NY) to generate the ICCs. Twelve cases provided 95% confidence to measure our objective for an ICC of .90 against a minimum ICC of .70 (Bonett, 2002). For metrics that achieved an ICC of .90, the mean, standard deviation, and range are reported on the basis of the median of the three scores for each participant.

Results

A total of 12 servicemembers (7 men and 5 women) participated in this study. The mean time to perform the CQDT was 19.6 min; 7 of 12 participants completed the task in <20 min and 11 of 12 in <23 min. The maximum test duration was 31.9 min. The average number of transits was 10.5. Table 1 provides the IRR results. Rule breaks and frequency of rule breaks were not reliable, with ICCs of .66 and .64, respectively. Task completion, transits, and total time were highly reliable, with ICCs of .94, .98, and .98, respectively.

Discussion

Occupational therapists are charged with developing and implementing measurement strategies that characterize the extent to which impairments impede daily life performance (Baum, Perlmutter, & Dunn, 2005). Doing so is difficult when impairments such as executive dysfunction are potentially difficult to detect, as in servicemembers with concussion. Performance-based assessments that involve multitasking have demonstrated the potential to discriminate between

Table 1. Preliminary Interrater Reliability Results for the Charge of Quarters Duty Task (N = 12)

Item	Reliability (ICC)	95% CI	Mean (SD)	Range
Task completion	.94	[.86, .99]	27.6 (5.6)	13–33
Rule breaks	.66 ^a	[.39, .88]		
Frequency of rule breaks	.64 ^b	[.32, .90]		
Transits	.98	[.96, .99]	10.5 (4.0)	5–18
Total time (min)	.98	[.96, .99]	19.6 (4.8)	13.2–31.9

Note. CI = confidence interval; ICC = intraclass correlation coefficient; SD = standard deviation. The mean, standard deviation, and range are reported only for metrics that achieved an ICC of .90.

^aFour of 12 triplets did not agree. ^bSix of 12 triplets did not agree.

healthy control participants and people with executive dysfunction (Alderman et al., 2003; Baum et al., 2008; Morrison et al., 2013; Wolf, Morrison, & Matheson, 2008) and may be an alternative to traditional measures of cognitive domains, which often fail to detect existing deficiencies in complex task performance (Tranel, Hathaway-Nepple, & Anderson, 2007). Although such tests do not appear to be subject to the ceiling effects of more structured measures of performance (Hall et al., 1996; Scott et al., 2011), they are typically complex to administer and score (Morrison et al., 2013). More multitasking tests that are specific to various clinical populations and life situations are needed. IRR specific to servicemembers with concussion and discriminant validity remain untested for the CQDT, but the preliminary evaluation of IRR in healthy participants suggests progress in the development of a multitask assessment of executive functioning for servicemembers with concussion.

The current evaluation of preliminary IRR highlights easily scored metrics for multitasking assessment and those requiring further refinement by the research team. IRR for task completion improved from previous versions of multitasking assessments because the score sheet was redesigned to include operational definitions and list performance tasks chronologically. These elements helped clarify scoring criteria and reduce rater disagreements regarding task performance.

Unfortunately, behavioral aspects of rule breaks and frequency of rule breaks were not as well specified, accounting for continued but solvable problems with IRR. Rater disagreements in how to score vocalizations directed at the examiners (e.g., asking the examiner questions) and the number of PVC parts brought to the assembly area largely explained the unacceptable ICCs for rule

breaks and frequency of rule breaks. Operational definitions were not clear enough to account for the unpredictable nature of human performance in these areas. Additionally, the restricted range resulting from only four rules may have had a negative impact on the ICC values. With a restricted range, one missed observation in rule breaks can affect the ICC value to a greater degree than with a greater number of rules. In preparation for future data collection, operational definitions have been revised and piloted to improve IRR for rule breaks.

Limitations and Future Directions

The CQDT is in relative infancy in terms of test development. Thus far, clinical feasibility and IRR for the CQDT have been evaluated in only a small number of healthy participants. Results of future data collection will determine IRR and clinical feasibility of the CQDT in a clinical population and, most important, will ascertain whether it discriminates between healthy control participants and servicemembers with concussion. If so, further research will need to be conducted to determine whether the CQDT predicts successful return to duty. Finally, the team is exploring the development of a civilian version of the CQDT that could be used as a stand-alone assessment of executive dysfunction.

Implications for Occupational Therapy Practice and Research

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

- Performance-based assessments of multitasking may enable occupational therapy practitioners to identify executive function deficits after concussion.

- Because of the complexity of scoring a multitask assessment, operational definitions for scoring are best developed on the basis of observed variations in task performance and differences in interpretation of that performance by multiple evaluators.
- The lessons learned in the development of the CQDT may benefit occupational therapy practitioners interested in developing performance-based assessments of executive dysfunction tailored to populations and practice settings other than the military.

Conclusion

There remains a need for reliable, valid, and clinically feasible assessments that can be used to identify executive dysfunction. Performance-based assessments that incorporate multitask methods and accurately simulate job demands may prove useful for occupational therapy practitioners in determining return-to-activity timelines in various populations. ▲

Acknowledgments

This ongoing work was funded by the U.S. Army Medical Research and Materiel Command. We thank the soldiers who provided valuable feedback to improve the face validity of the Charge of Quarters Duty Task and the soldiers who participated in this study. The opinions or assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the Army or the U.S. Department of Defense.

References

- Alderman, N., Burgess, P. W., Knight, C., & Henman, C. (2003). Ecological validity of a simplified version of the Multiple Errands Shopping Test. *Journal of the International Neuropsychological Society*, 9, 31–44. <http://dx.doi.org/10.1017/S1355617703910046>
- Baum, C. M., Connor, L. T., Morrison, T., Hahn, M., Dromerick, A. W., & Edwards, D. F. (2008). Reliability, validity, and clinical utility of the Executive Function Performance Test: A measure of executive function in a sample of people with stroke. *American Journal of Occupational Therapy*,

- 62, 446–455. <http://dx.doi.org/10.5014/ajot.62.4.446>
- Baum, C. M., Perlmutter, M., & Dunn, W. (2005). Establishing the integrity of measurement data. In M. Law, C. Baum, & W. Dawn (Eds.), *Measuring occupational performance* (pp. 49–64). Thorofare, NJ: Slack.
- Bonett, D. G. (2002). Sample size requirements for estimating intraclass correlations with desired precision. *Statistics in Medicine*, 21, 1331–1335. <http://dx.doi.org/10.1002/sim.1108>
- Brenner, L. A., Terrio, H., Homaifar, B. Y., Gutierrez, P. M., Staves, P. J., Harwood, J. E., . . . Warden, D. (2010). Neuropsychological test performance in soldiers with blast-related mild TBI. *Neuropsychology*, 24, 160–167. <http://dx.doi.org/10.1037/a0017966>
- Burgess, P. W. (2000). Strategy application disorder: The role of the frontal lobes in human multitasking. *Psychological Research*, 63, 279–288. <http://dx.doi.org/10.1007/s004269900006>
- Carroll, L. J., Cassidy, J. D., Peloso, P. M., Borg, J., von Holst, H., Holm, L., . . . Pépin, M.; WHO Collaborating Centre Task Force on Mild Traumatic Brain Injury. (2004). Prognosis for mild traumatic brain injury: Results of the WHO Collaborating Centre Task Force on Mild Traumatic Brain Injury. *Journal of Rehabilitation Medicine*, 43(Suppl.), 84–105. <http://dx.doi.org/10.1080/16501960410023859>
- Cuberos-Urbano, G., Caracuel, A., Vilar-López, R., Valls-Serrano, C., Bateman, A., & Verdejo-García, A. (2013). Ecological validity of the Multiple Errands Test using predictive models of dysexecutive problems in everyday life. *Journal of Clinical and Experimental Neuropsychology*, 35, 329–336. <http://dx.doi.org/10.1080/13803395.2013.776011>
- Dawson, D. R., Anderson, N. D., Burgess, P., Cooper, E., Krpan, K. M., & Stuss, D. T. (2009). Further development of the Multiple Errands Test: Standardized scoring, reliability, and ecological validity for the Baycrest version. *Archives of Physical Medicine and Rehabilitation*, 90(Suppl.), S41–S51. <http://dx.doi.org/10.1016/j.apmr.2009.07.012>
- Frisch, S., Förstl, S., Legler, A., Schöpe, S., & Goebel, H. (2012). The interleaving of actions in everyday life multitasking demands. *Journal of Neuropsychology*, 6, 257–269. <http://dx.doi.org/10.1111/j.1748-6653.2012.02026.x>
- Hall, K. M., Mann, N., High, W. M., Wright, J., Kreutzer, J. S., & Wood, D. (1996). Functional measures after traumatic brain injury: Ceiling effects of FIM, FIM+FAM DRS and CIQ. *Journal of Head Trauma Rehabilitation*, 11, 27–39. <http://dx.doi.org/10.1097/00001199-199610000-00004>
- Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1, 77–89. <http://dx.doi.org/10.1080/19312450709336664>
- Helmick, K., Baugh, L., Lattimore, T., & Goldman, S. (2012). Traumatic brain injury: Next steps, research needed, and priority focus areas. *Military Medicine*, 177(Suppl.), 86–92. <http://dx.doi.org/10.7205/MILMED-D-12-00174>
- McCrea, M., Iverson, G. L., McAllister, T. W., Hammeke, T. A., Powell, M. R., Barr, W. B., & Kelly, J. P. (2009). An integrated review of recovery after mild traumatic brain injury (MTBI): Implications for clinical management. *Clinical Neuropsychologist*, 23, 1368–1390. <http://dx.doi.org/10.1080/13854040903074652>
- Morrison, M. T., Giles, G. M., Ryan, J. D., Baum, C. M., Dromerick, A. W., Polatajko, H. J., & Edwards, D. F. (2013). Multiple Errands Test–Revised (MET–R): A performance-based measure of executive function in people with mild cerebrovascular accident. *American Journal of Occupational Therapy*, 67, 460–468. <http://dx.doi.org/10.5014/ajot.2013.007880>
- Radomski, M. V., & Morrison, M. T. (2014). Assessing abilities and capacities: Cognition. In *Occupational therapy for physical dysfunction* (7th ed., pp. 121–143). Baltimore: Lippincott Williams & Wilkins.
- Radomski, M. V., Weightman, M. M., Davidson, L. F., Finkelstein, M., Goldman, S., McCulloch, K., . . . Stern, E. B. (2013). Development of a measure to inform return-to-duty decision making after mild traumatic brain injury. *Military Medicine*, 178, 246–253. <http://dx.doi.org/10.7205/MILMED-D-12-00144>
- Scott, J. C., Woods, S. P., Vigil, O., Heaton, R. K., Schweinsburg, B. C., Ellis, R. J., . . . Marcotte, T. D.; San Diego HIV Neurobehavioral Research Center Group. (2011). A neuropsychological investigation of multitasking in HIV infection: Implications for everyday functioning. *Neuropsychology*, 25, 511–519. <http://dx.doi.org/10.1037/a0022491>
- Shallice, T., & Burgess, P. W. (1991). Deficits in strategy application following frontal lobe damage in man. *Brain*, 114, 727–741. <http://dx.doi.org/10.1093/brain/114.2.727>
- Tranel, D., Hathaway-Nepple, J., & Anderson, S. W. (2007). Impaired behavior on real-world tasks following damage to the ventromedial prefrontal cortex. *Journal of Clinical and Experimental Neuropsychology*, 29, 319–332. <http://dx.doi.org/10.1080/13803390600701376>
- Vagnozzi, R., Signoretti, S., Tavazzi, B., Floris, R., Ludovici, A., Marziali, S., . . . Lazzarino, G. (2008). Temporal window of metabolic brain vulnerability to concussion: A pilot ¹H-magnetic resonance spectroscopic study in concussed athletes—Part III. *Neurosurgery*, 62, 1286–1295, discussion 1295–1296. <http://dx.doi.org/10.1227/01.neu.0000333300.34189.74>
- Wolf, T. J., Morrison, T., & Matheson, L. (2008). Initial development of a work-related assessment of dysexecutive syndrome: The Complex Task Performance Assessment. *Work*, 31, 221–228.