Rubric scoring of a clinical test of executive functioning

Executive functions (EF) are complex abilities that allow one to successfully complete independent, deliberate, and novel goal-directed activities (Lezak, Howieson, & Loring, 2004). EF tests require solving problems with minimal direction from the examiner (Baddeley, 1992; Shallice & Burgess, 1991). Because EF skills tend to show up globally, Lezak et al. (2004) suggested clinicians will learn more about one's EF abilities by observing how he or she goes about solving a problem than from a test score. If this is the case, a "rubric" score that took into consideration "how" a problem was solved may provide the clinician with better information for treatment planning than a test score, as long as it did not greatly affect test sensitivity or specificity. Sensitivity and specificity are important factors in determining the usefulness of EF tests. Sensitivity refers to the probability of identifying abnormal functioning in an impaired individual or "hit rate" of a test, whereas specificity reflects the probability of correctly identifying healthy individual with the test (Cartoni & Lincoln, 2005; Kiel & Kaszniak, 2002).

Rubrics are useful scoring tools that divide tasks into component elements and provide a description of levels of performance for each element (Goodrich, 2005; Hanna & Smith, 1998). Rubrics have been widely used to assess student performance (Andrade, 2000; Falchikov, 1986; Goodrich, 1997), but have not been used to score EF tests. The aim of this study was to examine sensitivity and specificity for a clinical test of EF, the Rapid Assessment of Problem Solving test (RAPS) when scored with a rubric that allowed the examiner to describe the quality of performance using a standard that was developed from a large body of normative research. This differs from using the three traditional test scores from the RAPS that require time intensive calculations.

Method

Rapid assessment of problem solving test (RAPS)

The RAPS is a clinical test of problem solving based on Mosher and Hornsby's 20 Questions (20Q) task (1966). Materials for the RAPS include nine arrays of 32 pictures from known semantic categories, half in color and half black and white. Each array has three categories of 4, two categories of 6, and one category of 8 pictures from the same category. Three 20Q problems are solved on the RAPS. For each problem, the client is instructed to ask yes/no questions to identify an unknown target picture "with as few questions as possible." Affected pictures are crossed out after each question asked, and a problem is solved when the number of pictures has been reduced to 2 or 3. Three scores (See Table 1), number of questions (MQ), percent constraint questions (%C), and mean question-asking efficiency (MEQ) are used to score the RAPS.

Rubric and cut off scores

The rubric created for the RAPS was based on five types of questions used by normal subjects in solving problems on the test and the impact (positive and negative) these questions have had in solving problems (See Table 2). Rubric elements, shown in Table 3, were (a) planning, (b) strategy choice, (c) strategy execution, (d) attention to category size, (e) appropriate use of narrowing questions, and (f)

number of questions. To score an element, the examiner asks a specific question, and assigns a score of 2, 1, or zero. Element scores are summed for the three problems for a total rubric score. Tests from 350 normal subjects who had participated in research with the RAPS were scored with the rubric and with the MQ, %C, and MQE scores described in Table 1. Means and standard deviations were calculated and cut off scores established at 1 SD below the mean, or in the case of the MQ score, 1 SD above the mean (See Table 4). Inter- and intra-examiner agreement for scoring the elements of the rubric was also assessed.

Sensitivity and Specificity

To determine sensitivity and specificity for the RAPS for rubric scoring and the MQ, %C, and MQE scores, tests from neurologically compromised subjects with Alzheimer disease (AD; n = 18), aphasia (APH; n = 16, traumatic brain injury (TBI; n = 29), blast injuries (BI; n = 23) and severe mental illness (SMI; n = 30) and equal numbers of healthy controls matched for age, gender, and education (See Table 5) were also scored with the rubric and the MQ, % C, and MQE scores. Individual scores from the neurologically compromised and healthy controls were then examined in relationship to the cut off scores to compare test sensitivity and specificity for the different scores.

Results

Reliability

Percentages of intra- and inter-scorer agreements were calculated for 1692 and 1662 randomly selected rubric element scores respectively. Point-to-point agreement was attained for 95.2% and 94.3% of the intra-and inter-scorer comparisons respectively. Discrepant scores were reviewed by the two examiners and a joint decision was reached about whether the element should be scored 2, 1, or zero.

Sensitivity

Table 6 shows the number and percentage of subjects in each neurologically compromised group (AD, APH, TBI, BI and SMI) and the total number and percentage of all neurologically compromised subjects with scores outside the cut off scores. These data clearly reflect that the sensitivity of the RAPS is enhanced with rubric scoring. Table 6 shows that the rubric successfully identified many more neurologically compromised subjects in each group and more total subjects than any of the non-composite test scores.

Specificity

Table 7 shows the number and percentage of healthy controls and total number and percentage of all healthy controls with rubric and test scores inside the cut off scores. When the healthy controls are considered as a single group, specificity for RAPS is slightly higher (more healthy controls are found not

to have a problem solving deficit) for the MQ (91%), %C (91%), and MQE (93%) scores than for the rubric (78%).

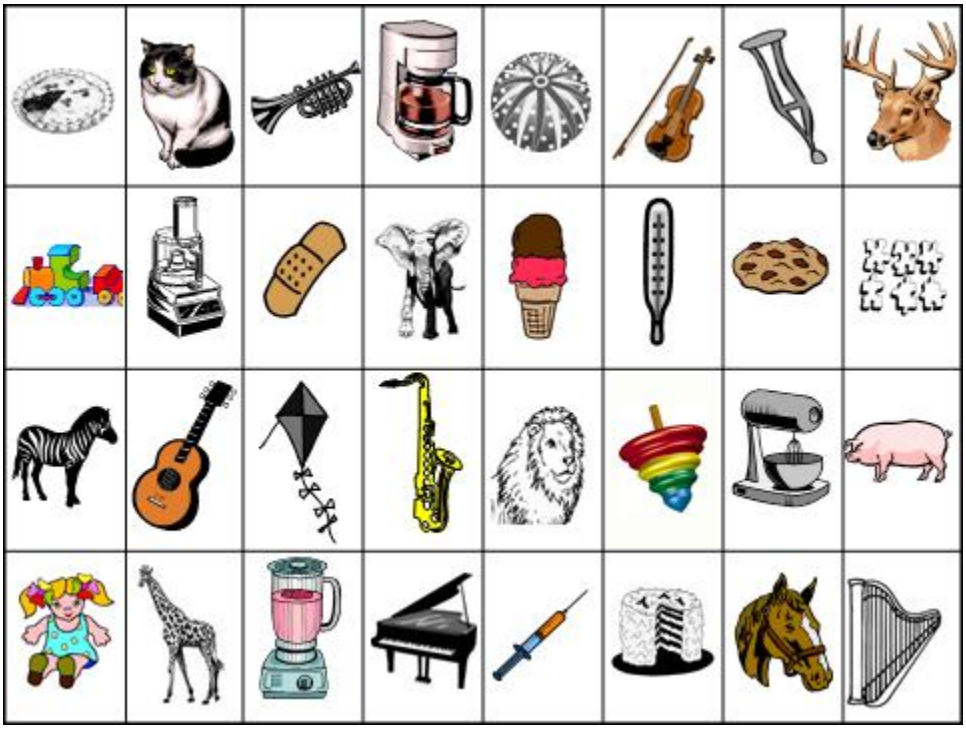
Discussion

Results support using a rubric to score the RAPS. Far more neurologically compromised subjects (90%) were identified with problem solving deficits with the rubric than with any of the three calculated scores. In contrast, slightly fewer healthy controls (78%) were identified with the rubric as not having problem solving deficits compared to the other scores. Lowered specificity, however, is not a contraindication to rubric scoring inasmuch as all healthy controls would not perform optimally on an EF measure but to reflect a range of performance levels (Kafer & Hunter, 1997).

In addition to increasing sensitivity, rubric scoring provided a more accurate representation of the problem solving abilities of some neurologically compromised subjects. For example, some subjects obtained MQ or %C scores inside the cut off scores by asking 1 or 2 “lucky questions.” Rubric scores, however, tended to negate this and reflect these subjects did not exercise the fluid reasoning necessary to solve problems on the RAPS.

While further research is needed to support rubric scoring of EF tests, similar scoring tools have been used successfully to quantify complex communication behaviors (Ambrose & Yairi, 1999). In discussing this paper, the authors will argue that rubric scoring is an attractive compromise between non-objective clinical ratings and traditional objective scores that might provide a mechanism for clinicians to make use of what they learn from observation in planning treatment.

TABLE 1. Example of a problem from the Rapid Assessment of Problem Solving Test (RAPS) and the numerical scores used for the test.



Mean number of questions (MQ). A problem is considered solved after the client's questions have reduced the number of pictures to 2 or 3 or the target picture is identified with a guess. The MQ score is obtained by computed the mean number of questions asked across three problems.

Percentage of constraint questions (%C). Two kinds of yes/no questions are acceptable on the RAPS: constraints (questions that eliminate more than one picture from consideration regardless of whether they are answered yes or no, e.g., "Is it an animal?) and guesses (questions that eliminate one picture with a "no" and solve the problem with a "yes," e.g., "is it the guitar?). The %C is obtained by dividing the total number of questions (constraints and guesses) by the number of constraint questions.

Mean question asking efficiency (MQE). Efficiency scores are calculated for the first 4 questions of each problem by dividing the number of pictures targeted or eliminated (whichever is less) by the number of pictures remaining when the question is asked and multiplying the result by 2. An efficiency score can range from 0-to-100. Scores are averaged across the three problems.

TABLE 2. Definitions of the types of questions asked by normal subjects on the Rapid Assessment of Problem Solving test (RAPS).

Type of Question	Definition
Category-limited constraint questions	Questions that target all pictures in one semantic category. Ordinarily this is 4, 6, or 8 pictures as these are the sizes of categories represented. However, it is possible for a category-limited question to target fewer pictures depending on prior questions asked.
Narrowing questions	Constraint questions asked after the target picture's category is known; these further reduce "narrow" the number of pictures and are congruent with the goal of solving the problem with as few questions as possible.
Novel constraint questions	Questions that target 9 or more pictures and/or have efficiency scores above 50%
Inefficient constraint questions	Questions with efficiency scores less than 50% that are not category-limited, narrowing, or novel questions
Guesses	Questions that target one picture in the array, solve the problem with a "yes" answer, and contribute little to the problem solving effort with a "no" answer.

Note: Research with the RAPS has shown that all questions asked by normal subjects in solving problems on the RAPS can be classified as one of the above mentioned types of questions. Citations documenting this research are not allowable under the CAC rules governing provision of information related to identification.

Table 3. Rubric elements and scoring.

Element	Scoring
1. Planning	How many pictures are targeted by the first question? 9-16 = 2 4-8 = 1 0-3 = 0
2. *Strategy choice*	What type of question is asked first? Novel = 2 Category limited = 1 Other constraint or guess = 0
3. Strategy execution	What type of question is asked second? Novel = 2 Category-limited = 1 Other constraint or guess = 0
4. *Awareness of category size*	What is the size of the picture category targeted by the *first* category-limited question? Largest category = 2 Next largest category = 1 Smallest = 0
5. Use of narrowing questions	What type of question is asked when a narrowing question is called for? Efficient narrowing question = 2 Any narrowing question = 1 Guess = 0
6. N*umber of questions*	How many questions are needed to solve the problem? 4 or less = 2 5 = 1 6 or more = 0 Score is zero if guessing is done on question 1-3

Note: Rationale for inclusion of all elements in the rubric is available, but this cannot be provided here because of CAC submission rules regarding author identification.

TABLE 4. Mean, standard deviation (SD) and cut off scores for the RAPS for 350 normal subjects.

Score	Mean	SD	Cut off Score
Mean Number of Questions	4.2	0.8	5.0
Percent Constraint Questions	88.2	12.4	75.8
Mean Question Asking Efficiency	62.0	11.0	51.0
Total Rubric Score	21.5	4.1	17.4

TABLE 5. Mean and standard deviations for age and education level in years, and gender distribution for neurologically compromised subjects with Alzheimer's disease (AD), aphasia (APH), traumatic brain injury (TBI), blast injuries (BI), and severe mental illness (SMI) and healthy controls (HC)

Group	Age Mean (SD)	Education Mean (SD)	M:F
AD (n=18)	80 (8)	13.5 (2.5)	10:8
HC (n=18)	77 (5)	14.2 (2.4)	8:10
APH (n=16)	53 (17)	13.8 (3.3)	12:4
HC (n=16)	51 (13)	14.6 (3.1)	12:4
TBI (n=29)	34 (17)	13.8 (3.4)	16:13
HC (n=29)	38 (16)	14 (2.4)	18:11
BI (n=23)	31 (7)	13.1 (1.9)	22:0
HC (n=23)	32 (8)	13.6 (2.4)	22:0
SMI (n=30)	38 (13)	11.6 (1.5)	14:16
HC (n=30)	37 (11)	14.4 (2.4)	15:15

TABLE 6. Number and percentage of subjects with Alzheimer disease (AD; n = 18), aphasia (APH; n = 16), traumatic brain injury (TBI; n = 29), blast injuries (BI; n = 23), and severe mental illness (SMI; n = 30) and number and percentage of all neurologically compromised subjects falling outside the cut off scores for normal subjects on the Rubric, MQ, %C, and MQE scores.

Group	Rubric		MQ		%C		MQE	
	#	%	#	%	#	%	#	%
AD	17	94	14	78	13	72	13	72
APH	12	75	6	38	9	56	10	62
TBI	25	86	16	55	15	52	17	57
BI	20	87	9	39	8	35	8	35
SMI	27	90	19	63	25	83	24	80
All	104	90	64	55	70	60	72	62

TABLE 7. Number and percentage of healthy controls matched to neurologically compromised subjects and number and percentage all healthy controls with scores falling within the cut off scores for normal subjects on the Rubric, MQ, %C, and MQE scores.

Group	Rubric		MQ		%C		MQE	
	#	%	#	%	#	%	#	%
HC-AD	16	89	16	89	15	83	17	94
HC-APH	16	100	15	94	16	100	15	94
HC-TBI	20	69	27	73	27	93	29	97
HC-BI	20	87	21	91	20	87	21	91
HC-SMI	19	63	27	90	27	90	26	87
All	91	78	106	91	105	91	108	93

References

Ambrose, N. G. & Yairi, E. (1999). Normal disfluency data for early childhood stuttering. *Journal of Speech, Language, and Hearing Research, 42,* 885-909.

Andrade, H. G. (2000). Using rubrics to promote thinking and learning. *Educational Leadership,* 57, *13-18.*

Baddeley, A. D. (1992). Working memory. *Science, 255,* 559.,

Cartoni, A., & Lincoln, N. B. (2005). The sensitivity and specificity of the Middlesex Elderly Assessment of Mental State (MEAMS) for detecting cognitive impairment after stroke. *Neuropsychological Rehabilitation, 15, I55-67.*

Falchikov, N. (198). Produce comparison and process benefits of collaborative peer group and self-assessments. *Assessment and Evaluation in Higher Education, 11,* 146-166.

Goodrich, H. A. (2005). Teaching with rubrics. The good, bad, and ugly. *College Teaching,* 3, 27-30.

Goodrich, H. A. (1997). Understanding rubrics. *Educational Leadership, 54,* 14-17.

Hanna, M. A., & Smith, J. (1998). Using rubrics for documentation of clinical work supervision. *Counselor Education and Supervision.*

Kafer, K. I., & Hunter, M. (1997). On testing the face validity of planning/problem-solving tasks in a normal population. *Journal of International Neurological Society, 3,* 108-119.

Kiel, K., & Kaszniak, A. (2002). Examining executive functioning in individuals with brain injury. A review. A*phasiology, 16,* 305-336.

Lezak, M., Howieson, D., & Loring, D. (2004). *Neurological assessment* (4th ed.). New York: Oxford University Press.

Mosher, F. A., & Hornsby, J. R. (1966). On asking questions. In: Bruner, J. S. et al. (Eds.). *Studies in cognitive growth.* (pp. 86-102). New York: Wiley.

Shallice, T. & Burgess, P. W. (1991). Deficits in strategy application following frontal lobe damage in man. *Brain, 114,* 727-741.