Journal of the International Neuropsychological Society (2006), **12**, 688–696.

DOI: 10.1017/S1355617706060838

Detection of inadequate effort on the California Verbal Learning Test-Second edition: Forced choice recognition and critical item analysis

JAMES C. ROOT,[1] REUBEN N. ROBBINS,[2] LUKE CHANG,[3] AND WILFRED G. VAN GORP[4]

[1]Department of Psychiatry, Weill Medical College of Cornell University, Manhattan, New York
[2]Department of Psychology, Fordham University, Bronx, New York
[3]Department of Psychology, New School for Social Research, Manhattan, New York
[4]Department of Psychiatry, Columbia University, Manhattan, New York

(RECEIVED January 20, 2006; FINAL REVISION May 12, 2006; ACCEPTED May 15, 2006)

Abstract

The Forced Choice Recognition (FCR) and the Critical Item Analysis (CIA) indices of the California Verbal Learning Test-II (CVLT-II) have been identified by the CVLT-II test developers as potentially useful, brief screening indicators of effort in neuropsychological assessment. This retrospective study analyzes performance on these measures in three groups: (1) clinically referred individuals; (2) forensically referred individuals not suspected of inadequate effort; and (3) forensically referred individuals whose performance on freestanding tests of effort suggested inadequate effort. Performances on FCR were analyzed for their relation to actual memory impairment and with regard to concrete and abstract distractor endorsement. FCR and CIA performances were analyzed for agreement with formal tests of inadequate effort and their test characteristics. Incremental validity was assessed by hierarchical logistic regression with previously identified indices for detection of inadequate effort on the CVLT. Results indicate that (1) FCR and CIA performances are not related to decreased memory performance; (2) FCR and CIA indices exhibit higher specificity and lower sensitivity, with higher positive predictive value than negative predictive value; and (3) FCR and CIA indices exhibit modest incremental validity with previously identified indices. Implications for use of FCR and CIA indices in inadequate effort detection are discussed (*JINS*, 2006, *12*, 688–696.)

Keywords: Malingering, Neuropsychological tests, Memory, Forensic, Assessment, Neuropsychology

INTRODUCTION

Medicolegal settings are very often presented with the possibility of inadequate effort by evaluees and resulting invalid performance on neurocognitive measures. Although estimates of base rates for inadequate effort in various populations vary considerably, most estimates suggest that a significant portion of evaluees may be at risk. In clinical populations, in which inadequate effort as a result of secondary gain is presumably less prevalent, the base rate of cases in which inadequate effort is an issue has been estimated to be 2 to 7% (Schretlen, 1988) to approximately 8% (Mittenberg et al., 2002). Base rates for inadequate effort among litigating populations, however, have been estimated to range from 8.5 to 14% (Frederick et al., 1994), 18 to 33% (Binder, 1993), 40% (Larrabee, 2003), and as high as 64% (Heaton et al., 1978), with a survey of clinicians' estimates of inadequate effort ranging from 19 to 30% (Mittenberg et al., 2002)

To address the problem of inadequate effort, numerous freestanding tests have been developed specifically to detect inadequate effort: the Rey 15-Item Test (Arnett et al., 1995; Lee et al., 1992), the Test of Memory Malingering (TOMM; Tombaugh, 1996), the Validity Indicator Profile (VIP; Frederick, 1997), the Computerized Assessment of Response Bias (CARB: Conder et al., 1992), and the Word Memory Test (WMT; Green et al., 1996), among others. Whereas a subset of these instruments has proven to be valid and reliable in measuring inadequate effort, problems remain in adequate detection both due to test characteristics

Correspondence and reprint requests to: James C. Root, Ph.D., Department of Psychiatry, Weill Medical College of Cornell University, 1300 York Avenue, F1302, New York, NY 10021. E-mail: jcr2003@med.cornell.edu

(Vallabhajosula & van Gorp, 2001) as well as to increasing availability of information about the measures and methods used to assess effort available to evaluees (Bauer & McCaffrey, 2006; Ruiz et al., 2002).

One alternative method is the use of instruments not originally designed for detection of inadequate effort. The first form of the CVLT has been subjected to several analyses for the detection of unusual profiles within the original indices that may be suggestive of inadequate effort (Ashendorf et al., 2003; Coleman et al., 1998; Millis et al., 1995; Sweet et al., 2000). Millis et al. (1995) analyzed CVLT performance both for multivariate and univariate predictors of inadequate effort in a group of subjects with mild, moderate, and severe head injury. The authors found that multivariate analysis did not surpass univariate predictors in predicting inadequate effort and focused on Recognition Discriminability, and Recognition Hits in particular, as promising indicators of inadequate effort.

The second edition of the California Verbal Learning Test (CVLT-II; Delis et al., 2000) added two new components, Forced Choice Recognition (FCR) and Critical Item Analysis (CIA), to specifically aid in situations in which a brief screen of effort is needed. Forced Choice Recognition-Total (FCR-Total) is an optional trial that is administered 10 minutes after the end of the Yes/No Recognition trial. The examinee is read a pair of words and asked to choose the word in each pair from the original list, with distractors consisting of both concrete and abstract items. In the normative sample from the CVLT-II manual (Delis et al., 2000), 90% of individuals achieved perfect scores of 16 correct across all age groups on the FCR-Total, with no individuals obtaining less than 14 of 16 correct. Given such uniformly high performance, the authors suggest that the FCR-Total may be a useful subtest in the detection of inadequate effort (Delis et al., 2000; p. 54).

The CVLT-II also includes normative data for Critical Item Analysis (CIA), which analyzes performance contrasts between easier and more difficult trials. Two indices are generated: CIA-Recall compares recall for a given item on any recall trial with recognition of that same item in the FCR. CIA-Recognition compares yes/no recognition for a given item with recognition of that same item in the FCR. The logic of Critical Item Analysis holds that items remembered on more difficult trials should also be remembered on easier trials. Data on the CIA from the original normative sample confirms that items once remembered on more difficult free-recall trials are typically remembered on later, easier recognition trials in 90% or more of the normative sample.

These optional components to the CVLT-II may offer important information in evaluations in which a brief screen of adequate effort is needed. In describing the characteristics of the FCR and CIA indices, the developers of the CVLT-II suggested that these indices might be helpful as a brief screening tool in detecting blatantly inadequate effort but were less optimistic regarding the detection of subtle or sophisticated attempts (CVLT-II Manual; Delis et al., 2000). Some preliminary characteristics of the FCR are documented in two earlier studies. Using the original form of the CVLT (Delis et al., 1987), Connor et al. (1997) found that a cutoff of less than 14 of 16 correct successfully classified 95% of individuals exhibiting inadequate effort of the total group. In a recent study using the current CVLT-II FCR, Moore and Donders (2004) examined agreement between scores on the FCR and the TOMM. The authors used an *a priori* cutoff of less than 15 of 16 on the FCR and/or a score of less than 45 on the second trial of the TOMM in a cohort of individuals with varying levels of traumatic brain injury and found 89% agreement between the CVLT-II FCR and TOMM. Among the 20 participants who scored below the *a priori* TOMM and FCR cutoff points, the FCR identified 15 individuals with invalid performance, whereas the TOMM identified 11 individuals with invalid performance. Of these individuals, six failed criteria on both instruments. Although Moore and Donders provide important data in the utility of the FCR and the TOMM, and document relatively strong agreement between the two instruments, certain aspects of their study limit the interpretability of their results, including analysis of only one cutoff score at a single base rate, and no analysis of FCR-Abstract, FCR-Concrete, or CIA indices.

The current study is a retrospective analysis based on archival review of evaluees drawn from a medicolegal practice. This study has the following features as its main focus: (1) to examine performance on the FCR in individuals with documented decrease in memory performance; (2) to examine performance on the FCR and CIA in relation to our defined gold standard for inadequate effort; (3) to assess sensitivity, specificity, positive (PPV) and negative (NPV) predictive value of the FCR-Total, -Concrete, and -Abstract Trials, and CIA-Recall and -Recognition Trials at a range of cutoff scores and at three base rates; and (4) to examine incremental validity of FCR and CIA indices when used in addition to previously identified indices.

FCR and CIA performance is analyzed in three groups of individuals collected from a clinical and forensic sample based on archival file review: (1) individuals referred for clinical evaluation in whom no obvious potential for inadequate effort or secondary gain was present (Clinical); (2) individuals referred for forensic evaluation who did not evidence inadequate effort as indicated by formal tests (Forensic-Adequate Effort, Forensic-AE); and (3) individuals referred for forensic evaluation whose performance on formal tests of effort indicated inadequate effort (Forensic-Inadequate Effort, Forensic-IE).

Given the relatively low demand characteristics of the FCR and CIA, we hypothesized that clinically referred individuals with decreased memory performance would perform similarly to the normative sample presented in the original standardization of the CVLT-II, and furthermore, that memory difficulties would not be significantly related to performance on these indices. With regard to the test characteristics of the FCR and CIA indices, given the relatively low demand characteristics of these indices and their

original development as brief screening measures, we hypothesized that the FCR and CIA indices would have higher specificity than sensitivity in detecting individuals with inadequate effort, (i.e., that individuals with adequate effort would be correctly classified as such, whereas those with inadequate effort would be able to avoid detection and be misclassified as exhibiting adequate effort). With regard to PPV and NPV, we hypothesized that each index would have higher PPV than NPV (i.e., that test positive scores would be more accurate indicators of inadequate effort than test negative scores would be of adequate effort).

We used two free-standing measures of effort: the TOMM (Tombaugh, 1996) and the VIP (Frederick, 1997) for determination of inadequate effort. We chose these two measures for their psychometric properties, their accuracy in identification of inadequate effort, and their acceptance by practitioners and professionals in forensic contexts (Vallabhajosula & van Gorp, 2001). In using the TOMM and VIP, it is noted that sensitivity and specificity of the CVLT-II FCR and CIA indices cannot exceed those of the TOMM and the VIP as these measures define our inadequate effort subgroup. As such, values, interpretation, and discussion of the CVLT-II FCR and CIA indices is qualified and limited by the gold standard measures to which they are compared. With regard to sensitivity and specificity values for the TOMM, previous studies indicate sensitivity rates ranging between 77% and 100% in simulation studies of malingering, with corresponding specificity rates all at 100% (Rees et al., 1998; Tombaugh, 1996, 1997). With regard to sensitivity and specificity of the VIP, the original validation study indicates sensitivity rates of the Verbal subtest at 67%, the Nonverbal subtest at 74%, either subtest at 78%, and both subtests at 63%; corresponding specificity rates were 83% for the VIP Verbal subtest, 86% for the Nonverbal subtest, 78% for either subtest, and 93% for both subtests (Frederick, 1997).

With regard to proposed cutoffs, we examine the performance of the CVLT-II indices in light of the Daubert standard. Because measures of effort are most often used in legal contexts, it is important to attend to court standards regarding admissibility. In the *Daubert v. Merrell Dow Pharmaceuticals* (1993) ruling, the Supreme Court held that, in evaluating expert testimony, FRE must be relied upon in assessing relevance and validity of such testimony. Opinion differs on what constitutes an acceptable level of accuracy in detection of inadequate effort, and, by extension, validity, in the legal context. Because no definitive level of accuracy has been established to meet the requirements of legal validity, we provide error rates at a range of cutoff scores to aid in the determination of acceptable levels of accuracy.

METHOD

Participants

Data included in this manuscript were obtained in compliance with regulations of the respective institutions through Institutional Review Board review and approval. Retrospective file review included 77 evaluees. Of the 77, 25 (33%) were clinical evaluees and 52 (68%) were forensically referred evaluees. Final analysis of the forensic group yielded 27 (52%) evaluees in the Forensic-AE group, and 25 (48%) individuals in the Forensic-IE group, as determined by the TOMM or the VIP. Of 77 records, we have probable diagnoses on 52% and believe these persons to be representative of the clinical sample as a whole. Diagnoses per group are as follows: Clinical group: 42% Mild Cognitive Impairment (MCI) (5), 17% Attention Deficit Disorder (ADD) (2), 17% Learning Disability (LD) (2), 17% Mood Disorder (Major Depression, Bipolar Disorder) (2), 8% Traumatic Brain Injury (TBI) (1); Forensic-Adequate Effort group: 36% Mild Cognitive Impairment (MCI) (5), 36% Mood Disorder (Depression) (5), 7% Neurological Condition (Multiple Sclerosis) (1), 14% Traumatic Brain Injury (TBI) (2); Forensic-Inadequate Effort group: 57% Traumatic Brain Injury (TBI) (8), 14% Mild Cognitive Impairment (MCI; Chronic Fatigue Syndrome, Lyme disease) (2), 7% Mood Disorder (Major Depression) (1), 7% Posttraumatic Stress Disorder (PTSD) (1). Demographic characteristics and performances on the CVLT-II, TOMM, and VIP are presented in Table 1.

Procedure

All participants were administered comprehensive neuropsychological batteries, including the CVLT-II and the FCR. Individuals in the forensic group were administered the TOMM and/or the VIP to assess effort. Individuals in the Clinical group were not administered freestanding measures of effort as, in these cases, no issues of secondary gain or probable future litigation were evident. A subset of participants were classified as exhibiting inadequate effort by *a priori* criteria (invalid performance on either or both subtests of the VIP and/or invalid performance on Trial 2 or the Retention Trial of the TOMM). The optional Retention Trial of the TOMM was administered only in cases in which performance on Trial 2 of the TOMM was below the cutoff indicated by the test author, as indicated in the TOMM manual (Tombaugh, 1996).

FCR-Total performance was assessed at different levels of memory performance as determined by the Long Delay Free Recall Trial (LDFR) of the CVLT-II. Test characteristics were calculated at different cutoffs for the FCR-Total, FCR-Concrete, FCR-Abstract, CIA-Recall, and CIA-Recognition. Three base rates were used in determining PPV and NPV: 48%, 30%, 15%. Hierarchical, logistic regression was performed analyzing incremental validity of FCR and CIA indices together with Recognition Hits and Recognition Discriminability with all predictors entered as continuous variables (Royston et al., 2006).

RESULTS

Clinical and forensic groups did not differ with respect to demographic characteristics. To ensure that there were no

Table 1. Demographic and neuropsychological characteristics of participant groups

Characteristic	Clinical	Forensic-Adequate Effort	Forensic-Inadequate Effort
Sample size	25	27	25
Sex (% male)	72%	64%	52%
Age (range)	39 (18–77)	42 (20–59)	43 (22–59)
Education (range)	16 (10–22)	16 (10–22)	14 (3–20)
WAIS-III VIQ (*SD*)	110 (16)	107 (14)	95 (18)
WAIS-III-PIQ (*SD*)	101 (16)	102 (14)	87 (13)
WAIS-III-FSIQ (*SD*)	107 (15)	105 (13)	91 (16)
CVLT-II Trials 1–5-Raw (*SD*)	49 (12)	48 (12)	40 (10)
CVLT-II Short Free (Raw) (*SD*)	10 (4)	9 (4)	7 (3)
CVLT-II Long Free—(Raw) (*SD*)	10 (5)	9 (4)	6 (3)
TOMM Trial1(*SD*)	—	45 (4)	35 (9)
TOMM Trial2 (*SD*)	—	49 (1)	41 (9)
TOMM Retention (*SD*)	—	—	38 (10)
VIP Verbal (% valid)	—	100%	57%
VIP Nonverbal (% valid)	—	100%	8%

Note. WAIS-III = Wechsler Adult Intelligence Scale-Third Edition; FSIQ = Full-Scale IQ; VIQ = Verbal IQ; PIQ = Performance IQ; CVLT-II = California Verbal Learning Test-Second Edition; TOMM = Test of Memory Malingering; VIP = Validity Indicator Profile.

significant differences between forensic and clinical participants in overall intellectual functioning or relative levels of impairment, performance on the Wechsler Adult Intelligence Scale-III Full-Scale IQ (FSIQ) and CVLT-II of the Clinical and Forensic-AE groups were compared. Because performance of individuals in the Forensic-IE group cannot be assumed to be valid, data from these individuals were not included in this analysis. An independent samples *t*-test revealed no significant difference in FSIQ between Clinical and Forensic-AE groups for whom FSIQ values were available [$t(48) = .360$; $p = .720$]. Likewise, using the LDFR of the CVLT-II as an index of memory performance, no significant difference was found between the two groups in memory performance [$t(50) = .709$; $p = .482$]. Given these results, clinical and forensic groups in our sample are comparable with regard to general level of intellectual functioning and level of memory impairment.

Clinical and Forensic-AE Performance

We predicted that individuals in the Clinical group would perform similarly to individuals in the original normative sample on the FCR-Total, and, related to this, that decreased memory performance would not significantly influence performance on the FCR-Total. Decreased memory performance was determined by analysis of the LDFR of the CVLT-II and included those individuals performing at ≤ -1.5 standard deviations on this trial. Analysis of LDFR performance yielded 8 of 25 individuals in the Clinical group whose LDFR performance was ≤ -1.5 standard deviations below the mean. Review of LDFR performance in the Clinical group indicates that no individual with decreased memory performance (≤ -1.5 standard deviations on LDFR) received a score of less than 15 of 16 correct on the FCR index.

FCR-Total scores of individuals in the Forensic-AE group were less consistent but similar to the CVLT-II standardization; performance of the Forensic-AE group on the FCR index was generally 15 and above with the exception of two scores (12; 14). In the Forensic-AE group, 11 of 27 individuals received scores of ≤ -1.5 standard deviations on the LDFR. Results of our analysis confirm that individuals in the Forensic-AE group with decreased memory performance on the LDFR perform at levels consistent with the unimpaired sample presented in the CVLT-II normative sample (90% of individuals obtaining a 15 or above on FCR-Total).

To assess the relationship between LDFR and FCR performance in the Clinical and Forensic-AE groups, FCR and LDFR performance (−4.0z through 1.5z) were compared. No significant correlation was observed between FCR-Total performance and LDFR performance [$r(52) = .211$; $p = .133$], indicating that decreased memory performance and performance on the FCR-Total are not related.

Sensitivity and Specificity of the CVLT-II FCR and CIA Indices

We tested the prediction that FCR and CIA indices would be more specific than sensitive in the detection of inadequate effort. FCR and CIA performance was analyzed in Forensic-AE and Forensic IE groups with Clinical group performance omitted from analysis. Test characteristics were analyzed at the natural base rate for individuals with inadequate effort in our sample (.48) and at two manipulated base rates (.30 and .15), that is, common base rates estab-

Table 2. Forced Choice Recognition-Total (FCR-T) test characteristics at different cut scores (total correct) for Forensic-Adequate Effort and Forensic-Inadequate Effort

Cut score	SENS	SPEC	PPV .48/.30/.15	NPV .48/.30/.15
9	.04	1.00	1.00/1.00/1.00	.53/.71/.86
10	.08	1.00	1.00/1.00/1.00	.54/.72/.86
11	.12	1.00	1.00/1.00/1.00	.54/.73/.87
12	.20	.96	.83/.68/.47	.57/.74/.87
13	.36	.96	.90/.79/.61	.62/.78/.89
14	.44	.93	.85/.73/.53	.64/.79/.90
15	.60	.81	.75/.58/.36	.69/.83/.92

Note. Inadequate effort base rate = .48, .30, and .15. SENS = sensitivity; SPEC = specificity; PPV = positive predictive value; NPV = negative predictive value.

lished by the literature for medicolegal and clinical populations, respectively (Greve & Bianchini, 2004; Mittenberg et al., 2002). Maximum sensitivity and specificity were analyzed at each cutoff, excluding perfect scores. Excluding perfect scores, the FCR-Total was most sensitive at a cutoff of 15 and below, correctly identifying a maximum of 60% of individuals with inadequate effort (81% of individuals with adequate effort; Table 2). The FCR-Concrete index correctly identified a maximum of 56% of individuals with inadequate effort (89% of individuals with adequate effort) at a cutoff of 7 and below; the FCR-Abstract index correctly identified a maximum of 40% of individuals exhibiting inadequate effort (93% of individuals with adequate effort) at a cutoff of 7 and below (Tables 3 and 4). The CIA-Recall index correctly identified a maximum of 36% of individuals with inadequate effort (78% of individuals with adequate effort) at a cutoff of 1 or more errors; the CIA-Recognition index correctly identified a maximum of 32% of individuals with inadequate effort (81% of individuals with adequate effort) at a cutoff of 1 or more errors (Tables 5 and 6).

Table 3. Forced Choice Recognition-Concrete (FCR-C) test characteristics at different cut scores (total correct) for Forensic-Adequate Effort and Forensic-Inadequate Effort

Cut score	SENS	SPEC	PPV .48/.30/.15	NPV .48/.30/.15
3	.04	1.00	1.00/1.00/1.00	.53/.71/.86
4	.04	.96	.50/.30/.15	.52/.70/.85
5	.16	.96	.80/.63/.41	.55/.73/.87
6	.32	.93	.80/.66/.45	.60/.76/.89
7	.56	.89	.82/.69/.47	.69/.83/.92

Note. Inadequate effort base rate = .48, .30, and .15. SENS = sensitivity; SPEC = specificity; PPV = positive predictive value; NPV = negative predictive value.

Table 4. Forced Choice Recognition-Abstract (FCR-A) test characteristics at different cut scores (total correct) for Forensic-Adequate Effort and Forensic-Inadequate Effort

Cut score	SENS	SPEC	PPV .48/.30/.15	NPV .48/.30/.15
3	.00	1.00	*/*/*	.52/.70/.85
4	.04	1.00	1.00/1.00/1.00	.53/.71/.86
5	.04	1.00	1.00/1.00/1.00	.53/.71/.86
6	.24	1.00	1.00/1.00/1.00	.59/.75/.88
7	.40	.93	.83/.71/.50	.63/.78/.90

Note. Inadequate effort base rate = .48, .30, and .15. SENS = sensitivity; SPEC = specificity; PPV = positive predictive value; NPV = negative predictive value.

These results confirm our prediction that the CVLT-II indices would exhibit higher specificity and lower sensitivity. Overall, sensitivity rates of the FCR and CIA indices ranged from 32 to 60%. In contrast, specificity rates at the same cutoffs as those cited for sensitivity ranged from 81 to 93%. These characteristics indicate that these indices will fail to identify between 40% and 68% of cases of inadequate effort even when using a cutoff of highest sensitivity, while a minority (7 to 19%) of individuals with adequate effort will be falsely identified as exhibiting inadequate effort.

PPV and NPV of the FCR and CIA Indices

Given evidence of widely varying base rates of inadequate effort depending on clinical and forensic context (Binder, 1993; Frederick et al., 1994; Trueblood & Schmidt, 1993), PPV and NPV were analyzed at the natural base rate of our sample (.48), as well as at two manipulated base rates (.30, and .15) as recommended by Rosenfeld et al. (2000). Tables 2, 3, 4, 5, and 6, report PPV and NPV for each CVLT-II index at each of the three base rates for inadequate effort. The recorded PPV and NPV values suggest that the

Table 5. CIA-Recall characteristics at different cut scores (total missed) for Forensic-Adequate Effort and Forensic-Inadequate Effort

Cut score	SENS	SPEC	PPV .48/.30/.15	NPV .48/.30/.15
6	.04	1.00	1.00/1.00/1.00	.53/.71/.85
5	.04	1.00	1.00/1.00/1.00	.53/.71/.85
4	.12	1.00	1.00/1.00/1.00	.55/.73/.87
3	.16	.96	.80/.63/.41	.55/.73/.87
2	.24	.96	.86/.72/.51	.58/.75/.88
1	.36	.78	.60/.41/.22	.57/.74/.87

Note. Inadequate effort base rate = .48, .30, and .15. SENS = sensitivity; SPEC = specificity; PPV = positive predictive value; NPV = negative predictive value.

Table 6. CIA-Recognition characteristics at different cut scores (total missed) for Forensic-Adequate Effort and Forensic-Inadequate Effort

Cut score	SENS	SPEC	PPV .48/.30/.15	NPV .48/.30/.15
6	.00	1.00	*/*/*	.52/.70/.85
5	.00	1.00	*/*/*	.52/.70/.85
4	.04	1.00	1.00/1.00/1.00	.53/.71/.86
3	.08	1.00	1.00/1.00/1.00	.54/.72/.86
2	.16	1.00	1.00/1.00/1.00	.56/.74/.87
1	.32	.81	.62/.42/.23	.56/.74/.87

Note. Inadequate effort base rate = .48, .30, and .15. SENS = sensitivity; SPEC = specificity; PPV = positive predictive value; NPV = negative predictive value.

CVLT-II indices achieve reasonable accuracy in identification of inadequate effort and are less accurate in the determination of adequate effort. This finding is particularly the case in the FC-Abstract and CIA-Recognition indices at a base rate of .48, where a cutoff of more than one error yields a test positive accuracy of 100% but a test negative accuracy of only 59% and 56%, respectively.

Proposed Cutoff Scores

At a base rate of .48, as established in the current study for the base rate of inadequate effort, FCR-Total yields a PPV of .85 and an NPV of .64 at a cutoff of 14 and below, misclassifying 56% of individuals with inadequate effort as adequate and 7% of individuals with adequate effort as inadequate. FCR-Concrete yields a PPV of .80 and an NPV of .60 at a cutoff of 6 and below, misclassifying 68% of individuals with inadequate effort as adequate and 7% of individuals with adequate effort as inadequate. FCR-Abstract yields a PPV of 1.00 and an NPV of .59 at a cutoff of 6 and below, misclassifying 76% of individuals with inadequate effort as adequate and correctly classifying all individuals with adequate effort. For CIA-Recall, a cutoff of 2 or more errors yields a PPV of .86 and an NPV of .58, misclassifying 76% of individuals with inadequate effort as adequate and 4% of individuals with adequate effort as inadequate. For CIA-Recognition, a cutoff of 2 or more errors yields a PPV of 1.00 and an NPV of .56, misclassifying 84% of individuals with inadequate effort as adequate and correctly classifying all individuals with adequate effort. As is the case when reducing the base rate of the target condition, PPV decreases while NPV increases; PPV and NPV for lower base rates of .30 and .15 are reported in Tables 2, 3, 4, 5, and 6.

Incremental Validity of FCR and CIA Indices With CVLT Variables

To determine the incremental validity of the FCR and CIA indices, a hierarchical, logistic regression was performed analyzing the FCR and CIA indices together with Recognition Hits (RH) and Recognition Discriminability (RD), two indices previously examined for their utility in detecting inadequate effort on the CVLT (Millis et al., 1995). Results from 10 hierarchical, logistic regressions (5 forward, 5 reversed) are presented in Table 7. Effect sizes of individual indices and of changes with the addition of predictors in individual models are estimated by w' (Cohen, 1988).

Results of this analysis suggest that the CIA-Recall index and all FCR indices were statistically significant in predicting group membership individually, with individual index effect sizes (w') of these predictors ranging from a minimum of .28 (CIA-Recall) to a maximum of .48 (FCR-Total). In forward models, in which FCR and CIA indices were the second step in the regression model, estimated effect size of change (w'_{chg}) indicated that all indices exhibited small to moderate effect sizes when analyzed jointly in the regression, with a minimum of .15 (RH, RD/CIA-Recognition) to a maximum of .36 (RH, RD/FCR-Abstract). With the addition of either FCR or CIA indices in forward models, clinical utility, as indicated by change in Overall Correct Classification, ranged from .00 (RH, RD/CIA-Recognition) to .10 (RH, RD/FCR-Total). Change in PPV ranged from .00 (RH, RD/FCR-Concrete; RH, RD/CIA-Recall; RH, RD/CIA-Recognition) to a maximum of .11 (RH, RD/FCR-Total), whereas change in NPV ranged from 0 (RH, RD/CIA-Recognition) to .09 (RH, RD/FCR-Total).

DISCUSSION

The FCR-Total, FCR-Concrete, FCR-Abstract, CIA-Recall, and CIA-Recognition indices have been suggested as measures that may be useful in assessing inadequate effort in clinically and forensically referred patients. Although initial results from the original CVLT-II standardization are encouraging toward this end, performance of clinical and forensic groups was not reported. In the current study, our main area of focus included the performance of clinically referred individuals on the FCR-Total index, the effect of decreased memory performance on the FCR-Total performance, test characteristics of the FCR and CIA indices, and incremental validity of these indices with previously identified indices.

That decreased memory performance does not correlate with performance on the FCR suggests that the learning and memory demand of this task is minimal and that the FCR and CIA indices may be used to assess effort in individuals exhibiting learning and memory difficulties in the absence of frank dementia. Clinically referred individuals, in whom inadequate effort is not suspected, consistently performed at near perfect levels on FCR-Total, and, therefore, by definition, within the range of adequate effort on FCR-Concrete, FCR-Abstract, and CIA-Recall, and CIA-Recognition. Given that forensically referred individuals with adequate effort did obtain relatively lower scores on the FCR-Total may suggest that the range of responding in forensic individuals is somewhat wider, or alternatively, that

Table 7. Logistic regression analyses predicting group membership of the Forensic-Adequate Effort *versus* Forensic-Inadequate Effort groups

Model	$\chi^2(df)$	w'	$\chi^2_{chg}(df)$	w'_{chg}	OCC	SENS	SPEC	PPV	NPV
Forward models									
Model 1: RH, RD; FCR-T									
Block: RH, RD	13.99 (2)**	.52	—	—	.69	.60	.78	.71	.68
Step: FCR-Total	20.39 (3)**	.63	6.41 (1) *	.35	.79	.72	.85	.82	.77
Model 2: RH, RD; FCR-C									
Block: RH, RD	13.99 (2)**	.52	—	—	.69	.60	.78	.71	.68
Step: FCR-Concrete	17.14 (3)**	.57	3.15 (1)	.25	.71	.68	.74	.71	.71
Model 3: RH, RD; FCR-A									
Block: RH, RD	13.99 (2)**	.52	—	—	.69	.60	.78	.71	.68
Step: FCR-Abstract	20.81 (3)**	.63	6.82 (1) **	.36	.75	.64	.85	.80	.72
Model 4: RH, RD; CIA-Recall									
Block: RH, RD	13.99 (2)**	.52	—	—	.69	.60	.78	.71	.68
Step: CIA-Recall	15.58 (3)**	.55	1.60 (1)	.18	.71	.68	.74	.71	.71
Model 5: RH, RD; CIA-Recog									
Block: RH, RD	13.99 (2)**	.52	—	—	.69	.60	.78	.71	.68
Step: CIA-Recognition	15.08 (3)**	.54	1.10 (1)	.15	.69	.60	.78	.71	.68
Reversed models									
Model 1: FCR-T; RH, RD									
Step: FCR-Total	11.86 (1)**	.48	—	—	.71	.60	.81	.75	.69
Block: RH, RD	20.39 (3)**	.63	8.54 (2)*	.41	.79	.72	.85	.82	.77
Model 2: FCR-C; RH, RD									
Step: FCR-Concrete	8.00 (1)**	.39	—	—	.73	.56	.89	.82	.69
Block: RH, RD	17.14 (3)**	.57	9.14 (2)*	.42	.71	.68	.74	.71	.71
Model 3: FCR-A; RH, RD									
Step: FCR-Abstract	10.60 (1)**	.45	—	—	.67	.40	.93	.83	.63
Block: RH, RD	20.81 (3)**	.63	10.21 (2)**	.44	.75	.64	.85	.80	.72
Model 4: CIA-Recall; RH, RD									
Step: CIA-Recall	3.99 (1)*	.28	—	—	.58	.36	.78	.60	.57
Block: RH, RD	15.58 (3)**	.55	11.59 (2)**	.47	.71	.68	.74	.71	.71
Model 5: CIA-Recog; RH, RD									
Step: CIA-Recognition	3.77 (1)	.27	—	—	.58	.32	.81	.62	.56
Block: RH, RD	15.08 (3)**	.54	11.32 (2)**	.47	.69	.60	.78	.71	.68

Note. Inadequate effort base rate = .48. RH = Recognition Hits; RD = Recognition Discriminability; w'_{chg} = the w' effect size computed for the χ^2_{chg}; OCC = Overall Correct Classification; SENS = sensitivity; SPEC = specificity; PPV = positive predictive value; NPV = negative predictive value.
*$p \leq .05$.
**$p \leq .01$.

our gold standard measures actually missed a subset of individuals with inadequate effort who obtained lower scores on the FCR-Total (scores of 12 and 14).

With regard to sensitivity and specificity of the FCR and CIA measures, our analysis suggests relatively lower sensitivity in detection of inadequate effort and generally higher specificity as is expected, such that a larger proportion of individuals with inadequate effort are missed at a wide range of cutoffs, whereas only a minority of individuals with adequate effort were misclassified as putting forth inadequate effort. Thus, the indices under review appear to err on the side of caution in identification of inadequate effort, with this interpretation further supported by PPV and NPV values discussed below.

Following on sensitivity and specificity values, PPV and NPV values demonstrate considerably varying levels of accuracy, depending on whether the test result is positive or negative. The predictive values of positive and negative test results suggests that positive test results are more definitive and indicative of the presence of inadequate effort than negative test results are of adequate effort. The transparently low task-demands of the several indices under consideration may lead to general conservatism in identifying inadequate effort, leading to a higher level of accuracy in the case of positive identifications, with relatively less accuracy in ruling out inadequate effort. Although all FCR and CIA indices are conservative in identification of inadequate effort, they do exhibit a range of conservatism, from no false positive identifications of inadequate effort and several false negatives, to relatively higher false positive rates and lower false negatives. On the conservative end of the spectrum, this finding is most evident in what might be considered the least demanding and transparent indices, FCR-Abstract and CIA-Recognition, in which errors are rela-

tively rare. In both indices, individuals with adequate effort never made more than one error, yielding a PPV of 1.00 at our proposed cutoff of two or more errors with a base rate of .48 inadequate effort. This level of conservatism leads to an imbalance of predictive value between positive and negative identifications that will be important to attend to when reviewing the incremental validity of these measures, discussed below; whereas 100% of individuals scoring below the cutoff for FCR-Abstract and CIA-Recognition were positive for inadequate effort, a significant portion were not detected (76% and 84%, respectively). Less transparently easy tasks, such as FCR-Total, FCR-Concrete, and CIA-Recall, exhibit a similar but less extreme pattern, correctly identifying more instances of inadequate effort than either FCR-Abstract or CIA-Recognition, but with a corresponding increase in false positive identifications.

Analysis of incremental validity indicates that the FCR-Total and FCR-Abstract indices significantly added to the predictive capacity of previously identified indices of inadequate effort (RH, RD) in forward models. Estimated effect sizes of change in predictive capacity in FCR-Total and FCR-Abstract indices were moderate, increasing overall correct classification between .06 and .10, respectively. As such, FCR-Total and FCR-Abstract, when used in conjunction with RH and RD indices, exhibit moderate incremental validity over the use of RH and RD in isolation. In contrast, FCR-Concrete, CIA-Recall, and CIA-Recognition did not significantly add to predictive capacity of RH and RD, as is illustrated by change in overall correct classification values of .02, .02, and .00, respectively. One caveat to the interpretation of overall correct classification is that this value is based on the total proportion of correct classifications (positive and negative) and may obscure the tendency of an individual index to err on either predominantly positive or negative identifications. As noted above, the FCR and CIA indices are relatively conservative in determination of inadequate effort and, as a result, yield higher PPV than NPV. In the present sample, the FCR-Total, -Concrete, and -Abstract indices also yield higher PPV in isolation than the combined RH and RD indices, with differences in PPV ranging from .04 (FCR-Total) to .12 (FCR-Abstract) over RH and RD indices alone. As a result, the utility of the FCR indices when used together with RH and RD indices will vary in relation to positive and negative findings.

With regard to disagreement between our gold standard measures and the FCR and CIA indices under review, one issue in need of further clarification is the possibility of missed identifications of inadequate effort by the TOMM and/or VIP. Several possibilities exist for possible false negatives by the gold standard measures. Chief among these are the levels of sensitivity and specificity documented for these measures. As discussed previously, any comparison between a gold standard measure and a candidate measure will be limited by the ultimate sensitivity and specificity values of the gold standard. Although both the TOMM and VIP have acceptable sensitivity and specificity values, a subset of individuals exhibiting inadequate effort are missed by both measures in previous studies (Frederick, 1997; Rees et al., 1998; Tombaugh, 1996, 1997). Other measures, such as the Word Memory Test (Green et al., 1996), may offer increased agreement with the FCR and CIA, due to reported higher sensitivity rates (Gervais et al., 2004). Apart from issues such as sensitivity of the gold standard measures, comparability of tasks between the TOMM, VIP, FCR, and CIA may also be an issue. The TOMM ostensibly measures memory for visually presented material, whereas the FCR and CIA indices both involve recognition of verbally presented material; the VIP, in contrast, does not involve any memory task, with the Verbal subtest consisting of a semantic matching task and the Nonverbal subtest consisting of a matrix completion task. It is possible that a subset of individuals in the current study were highly selective in putting forth inadequate effort on only those measures involving verbal memory. As such, further studies analyzing the FCR and CIA indices for their agreement with primarily verbal memory tasks with documented higher sensitivity may be suggested.

The FCR and CIA indices, developed within the CVLT-II as brief screens of effort, exhibit strong predictive value in positive findings of inadequate effort. Given the relatively brief format of the FCR (2 to 3 minutes administration time), it will serve as a useful measure of effort in situations in which a brief screen of effort is needed. Whereas a negative finding should not be relied upon as evidence of adequate effort, results of the current study suggest that a positive finding, in the absence of frank dementia, will be strongly suggestive of inadequate effort and would indicate the need for further testing of level of effort in the neuropsychological assessment.

ACKNOWLEDGMENTS

The information in this manuscript and the manuscript itself are new and original and have never been published either electronically or in print. No financial or other relationships that may create a conflict of interest exist. The authors thank Scott Millis, Ph.D., for his consultation on statistical methodology in the manuscript.

REFERENCES

Arnett, P.A., Hammeke, T.A., & Schwartz, L. (1995). Quantitative and qualitative performance on Rey's 15-Item Test in neurological patients and dissimulators. *Clinical Neuropsychologist*, *9*, 17–26.

Ashendorf, L., O'Bryant, S.E., & McCaffrey, R.J. (2003). Specificity of malingering detection strategies in older adults using the CVLT and WCST. *Clinical Neuropsychology*, *17*, 255–262.

Bauer, L. & McCaffrey, R.J. (2006). Coverage of the Test of Memory Malingering, Victoria Symptom Validity Test, and Word Memory Test on the Internet: Is test security threatened? *Archives of Clinical Neuropsychology*, *21*, 121–126.

Binder, L.M. (1993). Assessment of malingering after mild head trauma with the Portland Digit Recognition Test. *Journal of Clinical and Experimental Neuropsychology*, *15*, 170–182.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Coleman, R.D., Rapport, L.J., Millis, S.R., Ricker, J.H., & Farchione, T.J. (1998). Effects of coaching on detection of malingering on the California Verbal Learning Test. *Journal of Clinical Experimental Neuropsychology*, *20*, 201–210.

Conder, R., Allen, L., & Cox, D. (1992). *Computerized assessment of response bias test manual*. Durham, NC: Cognisyst.

Connor, D.J., Drake, A.I., Bondi, M.W., & Delis, D.C. (1997). *Detection of feigned cognitive impairments in patients with a history of mild to severe closed head injury*. Paper presented at the American Academy of Neurology, Boston.

Delis, D.C., Kramer, J.H., Kaplan, E., & Ober, B.A. (1987). *California Verbal Learning Test*. San Antonio, TX: The Psychological Corporation.

Delis, D.C., Kramer, J.H., Kaplan, E., & Ober, B.A. (2000). *California Verbal Learning Test-2nd ed*. San Antonio, TX: The Psychological Corporation.

Frederick, R.I. (1997). *Validity indicator profile manual*. Minnetonka, MN: NCS Assessments.

Frederick, R.I., Sarfaty, S.D., Johnston, J.D., & Powel, J. (1994). Validation of a detector of response bias on a forced-choice test of nonverbal ability. *Neuropsychology*, *8*, 118–125.

Gervais, R.O., Rohling, M.L., Green, P., & Ford, W. (2004). A comparison of WMT, CARB, and TOMM failure rates in non-head injury disability claimants. *Archives of Clinical Neuropsychology*, *19*, 475–487.

Green, P., Allen, L.M., & Astner, K. (1996). *The Word Memory Test: A user's guide to the oral and computer administered forms, US version 1.1*. Durham, NC: Cognisyst.

Greve, K.W. & Bianchini, K.J. (2004). Setting empirical cut-offs on psychometric indicators of negative response bias: A methodological commentary with recommendations. *Archives of Clinical Neuropsychology*, *19*, 533–541.

Heaton, R.K., Smith, H.H., Lehman, R.A., & Vogt, A.T. (1978). Prospects for faking believable deficits on neuropsychological testing. *Journal of Consulting and Clinical Psychology*, *46*, 892–900.

Larrabee, G.J. (2003). Detection of malingering using atypical performance patterns on standard neuropsychological tests. *Clinical Neuropsychology*, *17*, 410–425.

Lee, G.P., Loring, D.W., & Martin, R.C. (1992). Rey's 15-item visual memory test for the detection of malingering: Normative observations on patients with neurological disorders. *Psychological Assessment*, *4*, 43–46.

Millis, S.R., Putnam, S.H., Adams, K.M., & Ricker, J.H. (1995). The California Verbal Learning Test in the detection of incomplete effort in neuropsychological evaluation. *Psychological Assessment*, *7*, 463–471.

Mittenberg, W., Patton, C., Canyock, E.M., & Condit, D.C. (2002). Base rates of malingering and symptom exaggeration. *Journal of Clinical and Experimental Neuropsychology*, *24*, 1094–1102.

Moore, B.A. & Donders, J. (2004). Predictors of invalid neuropsychological test performance after traumatic brain injury. *Brain Injury*, *18*, 975–984.

Rees, L.M., Tombaugh, T.N., Gansler, D.A., & Moczynski, N.P. (1998). Five validation experiments of the Test of Memory Malingering (TOMM). *Psychological Assessment*, *10*, 10–20.

Rosenfeld, B., Sands, S.A., & van Gorp, W.G. (2000). Have we forgotten the base rate problem? Methodological issues in the detection of distortion. *Archives Clinical Neuropsychology*, *15*, 349–359.

Royston, P., Altman, D.G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine*, *25*, 127–141.

Ruiz, M.A., Drake, E.B., Glass, A., Marcotte, D., & van Gorp, W.G. (2002). Trying to beat the system: Misuse of the Internet to assist in avoiding the detection of psychological symptom dissimulation. *Professional Psychology: Research and Practice*, *33*, 294–299.

Schretlen, D.J. (1988). The use of psychological tests to identify malingered symptoms of mental disorder. *Clinical Psychology Review*, *85*, 451–476.

Sweet, J.J., Wolfe, P., Sattlberger, E., Numan, B., Rosenfeld, J.P., Clingerman, S., & Nies, K.J. (2000). Further investigation of traumatic brain injury versus insufficient effort with the California Verbal Learning Test. *Archives of Clinical Neuropsychology*, *15*, 105–113.

Tombaugh, T.N. (1996). *Test of Memory Malingering*. Toronto, ON: Multi Health Systems.

Tombaugh, T.N. (1997). The Test of Memory Malingering (TOMM): Normative data from cognitively intact and cognitively impaired individuals. *Psychological Assessment*, *9*, 260–268.

Trueblood, W. & Schmidt, M. (1993). Malingering and other validity considerations in the neuropsychological evaluation of mild head injury. *Journal of Clinical Experimental Neuropsychology*, *15*(4), 578–590.

Vallabhajosula, B. & van Gorp, W.G. (2001). Post-Daubert admissibility of scientific evidence on malingering of cognitive deficits. *Journal of the American Academy of Psychiatry and the Law*, *29*, 207–215.