Romanian Journal of Applied Psychology
2021, Vol. 23, No. 2, 32-39

Official journal of West University of Timisoara &
The Euroregional Center for Applied Psychology
DOI: 10.24913/rjap.23.2.02

Effectiveness of immediate vs. delayed recall in detecting invalid performance in coached and uncoached simulators: Results of two experimental studies

Iulia Crișan[1]*, Florin Alin Sava[1] & Laurențiu Paul Maricuțoiu[1]

[1] *West University of Timisoara, Romania*

Received 08.05.2021; Received revised 28.10.2021; Accepted 31.10.2021
Available online 30.12.2021

Objective: Two experimental studies were conducted to compare the ability of immediate and delayed recall indicators to discriminate between performances of simulators and full-effort clinical and non-clinical participants. *Methods:* Three groups of simulators (uncoached, symptom-coached, and test-coached), one group of community controls, and one group of cognitively impaired patients were assessed with four experimental memory tests, in which the immediate and delayed recall tasks were separated by three other tasks. *Results:* Across both studies, delayed recall demonstrated higher accuracy than immediate recall in classifying simulated performances as invalid, as compared to performances of bona fide clinical participants. ROC curve results showed sensitivities below 50% for both indicators at specificities of ≥ 90%. Computing performance curves across recall trials revealed descending trends for all three simulator groups indicating a suppressed learning effect as a marker of noncredible performances. Among types of coaching, test-coaching proved to decrease differences between simulators and patients. *Discussion:* The effectiveness of such indicators in clinical evaluations and their vulnerability to information about test-taking strategies are discussed.

Keywords: invalid performance, validity indicators, recall, simulation design, coaching

Address of correspondence: Iulia Crișan, West University of Timisoara, Department of Psychology. Address: 4 Vasile Parvan Bvd., 300223 Timisoara, Romania. E-mail: iulia.crisan@e-uvt.ro

ORCID: Iulia Crișan - https://orcid.org/0000-0003-1302-9921
Florin Alin Sava - https://orcid.org/0000-0001-8898-1306
Laurențiu Paul Maricuțoiu - https://orcid.org/0000-0001-8358-3130

Introduction

Decades of research show that standard performance validity tests (PVTs) assessing short-time memory (e.g., recall, recognition) are especially effective in detecting invalid performances (Bigler, 2014; Larrabee, 2003). A perspective on detecting noncredible performance is offered by looking into failures on specific tasks (e.g., recognition vs. recall), which would confirm a diagnosis of invalid performance. For instance, failure in forced-choice tasks is known to indicate invalid performance even in impaired samples (Bigler, 2014; Larrabee, 2003) and worse performance in recognition than in recall indicates noncredible responding (Greiffenstein et al., 1996; Suhr & Gunstad, 2000). Also, forced-choice testing has been deemed the best method to discriminate invalid responses from genuine performances by numerous empirical studies (Denning, 2012; Gunner et al., 2012; Inman & Berry, 2002; Strauss et al., 2002), reviews (Leighton et al., 2014), and meta-analyses (Crișan et al., 2021; Sollman & Berry, 2011). Despite their effectiveness, forced-choice measures are not infallible, as some tests are vulnerable to coaching (i.e., coached simulators producing above-chance performances on recognition tests), and online information is available about them (Rüsseler et al., 2008; Strauss et al., 2002). In addition, there are a few studies that support the idea that indicators based on recall might also be effective in detecting noncredible performance, yet not so identifiable by feigners (Strauss et al., 2002; Suhr & Gunstad, 2000). Hence, we designed two experimental studies to test the effectiveness of recall tasks in discriminating simulated from genuine cognitive impairment.

Types of recall: Immediate vs. delayed recall

Results of a recent meta-analysis on types of detection strategies moderated by types of stimuli and coaching revealed instruments relying on recognition to be most accurate in classifying invalid performances across experimental and criterion-group designs (Crișan et al., 2021). Concerning recall, both stand-alone and embedded indicators were found to generate modest effects in simulation designs compared with known-groups studies (Cohen's d = .66 vs. d= 1.05 for stand-alone indices, d = .65 vs. d = 1.21 for embedded indices). Another interesting finding for embedded measures based on recall was that test-coaching reduced differences between simulators and clinical participants, as opposed to methods relying on other strategies, in which case symptom-coaching was superior in reducing differences between means.

In their review on methodological characteristics of PVTs, Leighton and colleagues (2014) argued that, from the multitude of methodological moderators that require more investigation (e.g., types of stimuli, number of learning trials, trials characteristics), the influence of learning trials on performance in tasks eliciting *recall* has received far less attention than recognition. The authors also noted a need for more research concerning delayed recall performances in noncredible groups.

Regarding *learning characteristics* of PVT items, although there is evidence suggesting that one single exposure of test material (i.e., learning trial) may be enough to ensure recognition (Gunner et al., 2012; Denning, 2012), other findings prove that multiple encoding in *recall* sessions further facilitates recognition and long-term retention in cognitively impaired samples – or what is known as *the test-effect* (Leighton et al., 2014). The suppression of this effect would therefore be an indicator of invalid performance. When multiple retention trials are used, the learning effect may be displayed as a *performance curve*, showing ascending trends in the case of full-effort participants and descending trends for noncredible respondents (Bender & Rogers, 2004; Suhr & Gunstad, 2000; Rose, Hall & Szalda-Petree, 1998; Wogar et al., 1998).

Types of coaching: symptom-coaching vs. test-coaching

Currently, it is a well-known fact that coaching has a moderating effect on simulated performance (Bender & Rogers, 2004; Gorny & Merten, 2006; Suhr & Gunstad, 2000; Brennan et al., 2009). Still, the influence of types of coaching on the classification accuracy of assessment measures remains an ongoing issue in research, yielding some controversies. On the one hand, numerous empirical studies have found either test-coaching alone (i.e., instructing participants about detection strategies used by tests; DiCarlo et al., 2000; Bender & Rogers, 2004; Powell et al., 2004; Weinborn et al., 2012) or a mix of test-coaching and symptom-coaching (Rose, Hall & Szalda-Petree, 1998; Rüsseler et al.; 2008; Lau et al., 2017) to be more effective than symptom-coaching alone (i.e., supplying information about symptoms of the condition to be feigned) in reducing differences between scores of simulators and bona fide patients. On the other hand, two meta-analyses on validity indicators revealed symptom-coaching to be superior to test-coaching in reducing differences between groups (Crişan et al., 2021; Sollman & Berry, 2011). Therefore, more research on the differences between types of coaching is needed. In addition, as most studies used PVTs with good face validity and high classification accuracies (e.g., standard forced-choice tests), we propose investigating other experimental indicators' ability to classify performances moderated by coaching.

To conclude, we set the following research objectives for the present studies:

(1) To investigate the accuracy of immediate vs. delayed recall indicators in detecting noncredible performance in simulators compared with clinical patients and community controls. In this regard, we hypothesized that delayed recall would be superior to immediate recall.

(2) To compare performances of uncoached, symptom-coached, and test-coached simulators with performances of full-effort patients on the delayed recall task. We hypothesized that test-coached participants would show scores closer to clinical patients than the other two groups.

Study 1

Methods

Participants

The general sample was composed of 190 participants. Experimental participants were 90 psychology undergraduates (27 males and 63 females) who volunteered to take part in the study and received course credits for their involvement. Participants were randomized into three groups of simulators (uncoached N = 23, symptom-coached N = 22, and test-coached N = 22) and one full-effort group (N = 23). The undergraduate full-effort group was aggregated with 30 community volunteers (15 males and 15 females), recruited from the acquaintances of the researchers, to form the non-clinical control group, with no reported history of mental illness or cognitive dysfunction and no current involvement in lawsuits. Clinical patients were 70 neurological outpatients (38 males and 32 females) with cognitive impairment of heterogenous etiologies: traumatic brain injury (TBI) (N = 10); cerebrovascular accident (CVA) (N = 25); dementia of various etiologies (N = 35). Patients were included in the study if they had intact perceptual functions and reading and writing abilities. Two female CVA patients had to be excluded because of severe dysgraphia, leaving 23 patients in this group. No patient was involved in litigation at the time of the assessment, and none expressed interest regarding external benefits. All clinical participants were treated at an outpatient clinic specialized in cognitive and motor dysfunctions and were assessed with the Mini-Mental State Examination (MMSE; Folstein et al., 1999). The minimum score for inclusion was 15. Scores in our sample ranged between 18 and 29, with an average of 25.26 ± 1.87. The only significant difference between the three clinical groups was related to the patients' age (F = 15.470, p = .001), with TBI patients being younger than dementia and CVA patients.

In the analysis, groups of simulators, full-effort non-clinical, and genuine clinical controls were aggregated into three groups that showed significant differences in age, gender, and education between them (see table below).

Procedure

The experimental procedure was described in full in a different article (see Crişan et al., 2021). After being recruited and randomized into groups, experimental participants received via email the simulation instructions, adapted from previous studies (Brennan & Gouvier, 2006; Rüsseler et al., 2008; see table 1 of the appendix). The participants in the full-effort group were asked to react to tasks putting in their best effort. All participants read and signed an informed consent form with information about the study and indications of not disclosing experimental instructions. An extra incentive was provided: They were told that an unspecified sum of money would be awarded to the most credible simulated performance or the best performance in the case of the full-effort group. After collecting the data, the equivalent of $25 was given to one random participant from each group.

Participants in each experimental group were assessed individually by a licensed clinical psychologist who was unaware of the feigning conditions. After completing the test, all simulators had to complete a post-test questionnaire with manipulation checks and items referring to malingered performance and employed strategies. One male participant from the uncoached group was found uncompliant with experimental instructions and had to be excluded from the study, leaving a total sample of 89 experimental participants (22 in each simulator group and 23 in the full-effort group).

All control and clinical participants were assessed by a licensed psychologist and were asked to put their best effort into their test performance. They were not monetarily rewarded.

Assessment instruments

All participants were individually assessed using a battery with five memory tasks, of which the present paper concerns only indicators used to assess immediate and delayed recall performance. The analysis of the other indicators was presented in a different paper (Crişan et al., 2021).

First, 12 pictures of common objects were shown to the participant whose task was to name and memorize each picture. After being presented with all 12 objects, the participant was asked to recall all the memorized objects in any order. For the second trial, the procedure was repeated with identical instructions. A mean of correctly recalled items across both trials was computed as an immediate recall indicator.

The next tasks consisted of two forced-choice trials where participants had to choose each of the 12 memorized items from pairs with similar foils, and a word completion task where participants had to complete word stems first by including the 12 items, then by excluding them. Next, the BVRT - Benton Visual Retention Test, set A, was used as a distractor (i.e., a task with different stimuli, inserted before the delayed recall phase to divert the participant's attention from the original set of 12 items and provide the timeframe for the delay).

Finally, the participant had to recall the 12 items presented in the first recall phase. The total of correctly recalled items represented the delayed recall indicator.

Measures for internal consistency were computed for each type of stand-alone indicator: Cronbach's alpha was .894 for the immediate and delayed recall tasks, .913 for recognition tasks, and .893 for the process dissociation indicator.

Table 1. Demographic characteristics of individual and aggregated groups

Group	N	Mean Age ± SD (Range)	Gender		Mean Education ± SD
			m	f	
SIMULATORS	66	22.65 ± 6.56 (19-56)	17 (26%)	49 (74%)	13.21 ± 0.69
1 (NC)	22	20.73 ± 1.07 (19-43)	8 (36%)	14 (64%)	13.18 ± .58
2 (SC)	22	23.77 ± 8.62 (19-56)	6 (27%)	16 (73%)	13.23 ±.75
3 (TC)	22	23.45 ± 7.22 (19-48)	3 (14%)	19 (86%)	13.23 ± .75
CONTROLS	53	34.92 ± 15.32 (18-71)	24 (45%)	29 (55%)	13.43 ± 1.43
4 (FE)	23	23.61 ±7.27 (18-46)	9 (39%)	14 (61%)	13.00 ± .01
5 (CV)	30	43.60 ±14.19 (21-71)	15 (50%)	15 (50%)	13.77 ± 1.85
PATIENTS	68	63.66 ±13.98 (36-88)	38 (56%)	30 (44%)	11.46 ± 2.52
6 (Dem)	35	69.46 ± 10.51 (57-88)	17 (49%)	18 (51%)	11.51 ± 2.82
7 (CVA)	23	62.39 ± 11.69 (36-79)	13 (57%)	10 (43%)	11.39 ±2.38
8 (TBI)	10	46.30 ± 15.33 (27-72)	8 (80%)	2 (20%)	11.40 ± 1.83
TOTAL	187	41.04 ± 21.65	79 (42%)	108 (58%)	12.64 ± 1.96
Differences between individual groups		F = 94.55, p = .001	Kruskal-Wallis Test p = .009		F = 7.26, p = .001
Differences between aggregated groups		F = 193.350, p = .001	Kruskal-Wallis Test p = .002		F = 24.48., p = .001

Abbreviations: NC: uncoached simulators; SC: symptom-coached simulators; TC: test-coached simulators; FE: full-effort experimental participants; CV: community volunteers: Dem: Dementia Patients; CVA: Cerebrovascular Accident patients; TBI: Traumatic brain-injured patients

Table 2. Overall results on indicators of immediate and delayed recall

Variables (m, SD)	Simulators				Controls			Patients				Cohen's for SIM PTS	Cohen's for C vs. PTS
	NC	SC	TC	SIM	FE	CV	C	Dem	CVA	TBI	PTS		
N	22	22	22	66	23	30	53	35	23	10	68		
Mean	6.18	7.13	6.95	6.75	9.82	9.66	9.73	6.83	6.82	7.45	6.92	.072	1.527
Recall	*2.58*	*2.15*	*2.02*	*2.27*	*1.63*	*1.34*	*1.46*	*2.23*	*2.14*	*2.04*	*2.16*		
Delayed	5.32	6.32	6.41	6.02	10.96	10.83	10.89	7.54	8.26	8.30	7.90	.635	1.239
Recall	*3.24*	*2.75*	*2.55*	*2.86*	*1.43*	*1.62*	*1.52*	*3.47*	*2.58*	*2.54*	*3.05*		

Abbreviations: NC: Uncoached simulators; SC: Symptom-coached simulators; TC: Test-coached simulators; SIM: Aggregated simulator sample; FE: Full-effort experimental participants; CV: Community volunteers; C: Aggregated control sample; Dem: Dementia patients, CVA: Cerebrovascular accident patients; TBI: Traumatic brain-injured patients; PTS: Aggregated patient sample

Table 3. ROC curve analysis of immediate and delayed recall indicators in simulators vs. controls and simulators vs. patients

Validity Indicators	AUC	Std. Error	Sig.	95% CI Lower Bound	Upper Bound	Effect size (Cohen's *d*)	Cutoff	Sn	Sp
SIM vs. C									
Immediate Recall	.876	.031	.000	.815	.937	1.561	≤ 8.00	.667	.906
Delayed Recall	.936	.021	.000	.894	.977	2.126	≤ 8.00	.667	.962
SIM vs. PTS									
Immediate Recall	.508	.050	.871	.410	.606	.072	≤ 4.00	.121	.926
Delayed Recall	.690	.046	.000	.601	.780	.635	≤ 4.00	.197	.912

Abbreviations: SIM vs. C: simulators vs. community controls; SIM vs. PTS: simulators vs. clinical patients

Results

Data analysis

Overall data were analyzed using IBM SPSS Statistics 25. Descriptive statistics for the individual and aggregated groups on the recall indicators are displayed in Table 2.

Manipulation checks – differences between groups

As our participants came from three different populations, one-way ANCOVAs were conducted between scores of the three aggregated groups on the immediate and delayed recall indicators whilst adjusting for age, gender, and education, at a 99% confidence interval. We found significant differences between the three groups on the delayed recall indicator [F (2, 181) = 53.448, p = .001, Eta² = .371], and on the immediate recall indicator [F (2, 181) = 38.547, p = .001, Eta² = .299]. LSD post hoc tests showed significant differences (p = .001) between simulators and controls and between controls and patients on both indicators, but significant differences between controls and patients were only found in delayed recall performance (p = .004). Comparing the estimated marginal means further showed that the simulator group produced significantly lower scores on this indicator than the control group and the neurological patient group. The immediate recall indicator failed to produce significant differences between simulators and patients at a 99% confidence interval. The observed statistical power of 1.000 for both indicators would allow us to conclude that the study was well designed for the examination of our hypotheses.

Immediate vs. delayed recall

Next, we wanted to determine the ability of immediate vs. delayed recall indicators to discriminate between scores of simulators and patients at cutoffs with specificities of ≥ 90%. Results are shown above (Table 3).

Results showed marked differences between the two contrast categories concerning the classification accuracy of the two indicators. Both immediate and delayed recall demonstrated high to excellent abilities to discriminate between simulators and non-clinical controls, classifying noncredible performances with 66.7% sensitivities at cutoffs ≤ 8.00. However, in the simulator vs. patient contrast, only the delayed recall indicator demonstrated a significant AUC value and a moderate effect size, thus confirming our first hypothesis. The indicator for immediate recall failed to significantly discriminate between simulators and patients (as previously indicated by the results of the ANCOVA). Still, both indicators' failure to produce acceptable sensitivities in simulators vs. patients limits their accuracy in this type of contrast.

Differences between simulators: Performance curves of recall

To assess the influence of learning trials on performance in recall and to see whether the suppression of the learning effect was characteristic of simulators' performances, we conducted independent samples t-test comparisons between the means of correctly recalled items from the first two immediate recall trials and the correctly recalled items from the delayed recall phase. At this stage, we took each experimental condition of feigning separately for comparison. Full-effort controls and neurological patients were again considered as aggregated groups. Results are shown in table 4.

Table 4. Comparison of recall performances – simulators vs. controls and simulators vs. patients

Comparison	Variable	Levene's Test for equality of variances		t-test for equality of means						Cohen's *d*
		F	Sig.	t	Sig.	Mean Difference	Std. Error Difference	95% CI		
								Lower	Upper	
NC vs. C	Immediate RC	12.884	.001	-6.060	.001**	-3.554	.470	-4.492	-2.615	1.692
	Delayed RC	27.005	.001	-7.708	.001**	-5.569	.549	-6.663	-4.474	2.201
SC vs. C	Immediate RC	3.531	.064	-6.059	.001**	-2.599	.429	-3.454	-1.744	1.409
	Delayed RC	10.350	.002	-7.337	.001**	-4.569	.497	-5.559	-3.579	2.056
TC vs. C	Immediate RC	1.852	.178	-6.671	.001**	-2.781	.416	-3.612	-1.950	1.575
	Delayed RC	12.018	.001	-7.665	.001**	-4.478	.477	-5.429	-3.526	2.134
NC vs. Pts	Immediate RC	1.743	.190	-1.326	.188	-.737	.556	-1.842	.368	.309
	Delayed RC	.700	.405	-3.426	.001**	-2.579	.760	-4.089	-1.068	.819
SC vs. Pts	Immediate RC	.010	.922	.410	.683	.217	.529	-.834	1.269	.100
	Delayed RC	.213	.645	-2.188	.031*	-1.579	.732	-3.033	.125	.544
TC vs. Pts	Immediate RC	.280	.598	.068	.946	.035	.521	-1.001	1.072	.017
	Delayed RC	.296	.588	-2.092	.039*	-1.488	.722	-2.922	-.054	.530

** Significant at p < .01; * Significant at p < .05

Abbreviations: NC vs. C: uncoached simulators vs. community controls; SC vs. C: symptom-coached simulators vs. community controls; TC vs. C: test-coached simulators vs. community controls; NC vs. PTS: uncoached simulators vs. clinical patients; SC vs. PTS: symptom-coached simulators vs. clinical patients; TC vs. PTS: test-coached simulators vs. clinical patients; RC: recall

We obtained significant differences between performances of each feigning group contrasted with full-effort controls, as demonstrated by very large effect sizes (Cohen's *d* > 1.5) generated by both indicators of recall. There were no significant differences between simulators and patients in terms of immediate recall performance. However, significant differences (p < .05) were found in delayed recall performance, with moderate effect sizes irrespective of the feigning condition. The smallest effect for delayed recall was observed between test-coached simulators and patients (Cohen's d = .53) and the largest difference was obtained for uncoached simulators vs. patients (Cohen's d = .819). We then computed individual differences between group means, reflecting the change in recall performance across the entire test battery.

Observing the individual differences between group means, we noted that full-effort participants, either healthy or impaired, demonstrated an increase in cognitive performance across recall trials. In other words, the positive difference between the score of the delayed recall phase and the first two recall trials showed that full-effort respondents retained significantly more words across tasks, irrespective of the distraction provided by the BVRT, thus proving a learning effect. Interestingly, this increment in performance seemed to be maintained in neurological patients despite their implicit cognitive impairment (0.99), closely matching the improvement in performance of healthy controls (1.16). On the other hand, all groups of simulators demonstrated a negative difference between recall scores, attesting a suppressed learning effect across test trials, thereby supporting a diagnosis of invalid performance. Of note, the highest decrease in performance was scored by the uninstructed (-0.86) and symptom-coached groups (-0.81), while test-coached participants showed a moderate difference (-0.54). These results supported our second hypothesis, indicating that test-coaching decreased the differences between delayed recall performances of simulators and clinical patients. Comparing the performance curves that accounted for the presence or absence of learning effects across groups showed descending trends for all three simulator groups and ascending trends for both full-effort groups, regardless of their clinical status.

Discussion

Results of the first study confirmed both of our hypotheses: delayed recall was found more effective than immediate recall in discriminating between simulators and patients, although both of these indicators' classification accuracies failed to reach acceptable sensitivities at ≥ 90% specificities. Comparing performances across recall trials in simulators and full-effort participants showed the suppression of the learning effect in all three simulator groups, displayed as a descending performance curve which is a marker for invalid performance (Bender & Rogers, 2004; Wogar et al., 1998). Consistent with our second hypothesis, test-coached simulators demonstrated the smallest decrement in recall performance which supports the impact of test-coaching on simulated performance in the sense of increasing its credibility. A second study was designed to verify our findings.

Study 2. A replication study

Methods

Participants

A total sample of 108 participants was used. Experimental participants were 48 psychology undergraduates (17 males and 31 females), randomly allocated to three groups of simulators of 16 participants each (uncoached, symptom-coached, and test-coached). All volunteered to take part in the study and received course credits for their participation. Control participants (12 males and 18 females) included ten undergraduate participants and 20 community volunteers recruited from the social networks of the researchers, with no reported history of mental conditions or cognitive impairment, and no present involvement in any type of litigation.

The clinical group was composed of 30 outpatients (15 males and 15 females), with psychiatric diagnoses like major depressive disorder (MDD) (N = 2), panic disorder (N = 2), chronic alcoholism (N = 1), and delusional disorder (N = 1); and neurological diagnoses, such as polyneuropathy (N = 2), traumatic brain injury (TBI) (N = 2), cerebrovascular accident (CVA) (N = 7), Alzheimer's (N = 5) and Parkinson's dementia (N = 8). All patients had intact perceptual functions and reading and writing abilities. All were psychiatrically and physically treated at a rehabilitation clinic specialized in cognitive impairment. They were tested as part of standard neuropsychological assessment, which included the Mini-Mental State Examination (MMSE), with the minimum score for inclusion of ≥ 15. Scores ranged between 21 and 29, with an average of 25.80 ± 2.64. Like in the first study, the only difference was related to the patients' age (F = 13.680, p = .001), with dementia and CVA patients being older than the other participants.

The demographic characteristics of the individual and aggregated groups are displayed in the table below. Significant differences in age and education were observed between the three aggregated groups, simulators being younger than the other two groups, and patients having a lower average of education.

Procedure

The same protocol as in the first study was used, but the assessment procedure was adapted for online testing (see Assessment instruments below).

Assessment instruments

The same memory tasks as in the first study were administered, but instead of the BVRT, we used a distractor task consisting of recalling and recognizing two rows of digits, because it was considered more suitable for online assessment and its duration was similar to the BVRT in the first study (i.e., approximately 5 minutes). Similar to the BVRT, participants' scores on these tasks were not included in the analysis. Measures for internal consistency were again computed for each type of stand-alone indicator and were found to closely match our first study: Cronbach's alpha was .890 for the immediate and delayed recall tasks, .897 for recognition tasks, and .886 for the three process dissociation indicators.

Table 5. Demographic characteristics of individual and aggregated groups

Group	N	Mean age ± SD (Range)	Gender		Mean Education ± SD
			m	f	
SIMULATORS	48	25.25 ± 7.620 (19-50)	17 (35%)	31 (65%)	14.60 ± 2.303
NC	16	23.13 ± 6.820 (19-47)	6 (38%)	10 (63%)	14.375 ± 2.680
SC	16	27.06 ± 8.512 (19-50)	7 (44%)	9 (56%)	14.687 ± 2.625
TC	16	25.56 ± 7.384 (20-42)	4 (25%)	12 (75%)	14.750 ± 1.570
CONTROLS	30	36.43 ± 14.467 (18-65)	12 (40%)	18 (60%)	15.233 ± 2.775
PATIENTS	30	62.13 ± 13.351 (30-82)	15 (50%)	15 (50%)	12.400 ± 3.519
TOTAL	108	38.60 ± 19.190	44 (41%)	64 (59%)	14.166 ± 3.009
Differences between aggregated groups		F = 94.107, p = .001	Kruskal-Wallis Test p = .445		F = 8.641., p = .001

Abbreviations: NC: uncoached simulators; SC: symptom-coached simulators; TC: test-coached simulators

Table 6. Overall results on immediate and delayed recall indicators

Variables	Simulators				Controls	Patients	Cohen's *d* for SIM vs. PTS	Cohen's *d* for C vs. PTS
(m, SD)	NC	SC	TC	SIM	30	30		
N	161	16	16	48				
Mean Recall	6.187	6.250	7.093	6.510	10.333	7.300	.424	1.925
	2.122	*1.663*	*1.518*	*1.793*	*1.116*	*1.928*		
Delayed Recall	5.190	4.940	6.630	5.580	11.370	8.330	.979	1.527
	2.994	*2.886*	*2.872*	*2.952*	*.928*	*2.657*		

Abbreviations: NC: Uncoached simulators; SC: Symptom-coached simulators; TC: Test-coached simulators; SIM: Aggregated simulator sample; FE: Full-effort experimental participants; CV: Community volunteers; C: Aggregated control sample; Dem: Dementia patients, CVA: Cerebrovascular accident patients; TBI: Traumatic brain-injured patients; PTS: Aggregated patient sample

Table 7. ROC curve analysis of immediate and delayed recall indicators in simulators vs. controls and simulators vs. patients
Abbreviations: SIM vs. C: simulators vs. community controls; SIM vs. PTS: simulators vs. clinical patients

Validity Indicators	AUC	Std. Error	Sig.	95% CI		Effect size (Cohen's *d*)	Cutoff	Sn	Sp
				Lower Bound	Upper Bound				
SIM vs. C									
Immediate Recall	.972	.014	.000	.944	1.000	2.559	≤ 8.00	.833	.967
Delayed Recall	.966	.016	.000	.938	.999	2.646	≤ 8.00	.792	1.000
SIM vs. PTS									
Immediate Recall	.623	.066	.070	.493	.752	.424	≤ 4.00	.063	.933
Delayed Recall	.759	.056	.000	.650	.868	.979	≤ 4.00	.458	.933

Results

Data analysis

Overall data were analyzed using IBM SPSS Statistics 25. Descriptive statistics for the individual and aggregated groups on the recall indicators are displayed in Table 5.

Manipulation checks – differences between groups

To test differences between participants, one-way ANCOVAs were conducted between scores of the three aggregated groups on the performance validity indicators whilst adjusting for age, gender, and education, at a 99% confidence interval. Results yielded significant differences for both the immediate recall indicators [F (2,102) = 46.904, p = .001, Eta^2 = .479] and the delayed recall indicator [F (2, 102) = 48.517, p = .001, Eta^2 = .488]. LSD post-hoc tests largely confirmed our initial findings: both indicators yielded significant differences between simulators and controls (p = .001) and between controls and patients (p = .001), but while immediate recall failed to discriminate between simulators and patients at acceptable probabilities (p = .039), the delayed recall indicator produced significant differences between these groups (p = .001). In addition to our first study, no significant differences were observed between delayed recall performances of community vs. clinical controls (p = .068). Comparing the estimated marginal means further showed that the simulators produced significantly lower scores on these indicators than controls and neurological patients. Moreover, Again, the observed statistical power was 1.000 for both indicators.

Immediate vs. delayed recall

Results of ROC curve analyses matched our previous study, confirming our first hypothesis: while both indicators demonstrated high discrimination ability between simulators and community controls and very large effect sizes between group means, only the delayed recall indicator generated an AUC value in the fair range that discriminated simulators from patients. Still, in this contrast, both indicators failed to produce acceptable sensitivities at specificities of ≥ 90%

Differences between simulators: Performance curves of recall

Independent samples t-test comparisons between scores of immediate recall and delayed recall across types of contrasts yielded no significant differences between simulators and patients in immediate recall performance. Congruent with initial findings, moderate to large effects for the delayed recall indicator showed noncredible performance in the case of all three simulator groups as compared to genuine patients.

Next, differences between performances in immediate vs. delayed recall were computed across contrasts between the three simulator groups vs. controls and patients, to generate curves of recall performance. The results matched our previous findings: while both full-effort groups (controls and patients) showed an increase in recall performance across the test (of approximately 1 point) and ascending curves, all three groups of simulators displayed a decrease in performance (indicating a suppressed learning effect) with descending curves. In contrast to the first study, symptom-coached feigners produced the largest difference (-1.31), and consistent with initial findings, the smallest difference was observed for test-coached simulators (-0.46). Therefore, across both studies, participants receiving coaching about test-taking strategies demonstrated performances that were closer to genuine patients, and an effect of test-coaching on making performance more credible could be inferred.

Table 8. Comparison of recall performances – simulators vs. controls and simulators vs. patients – study 2

Comparison	Variable	Levene's Test for equality of variances		t-test for equality of means						Cohen's *d*
		F	Sig.	t	Sig.	Mean Diff	Std. Error Diff	95% CI Lower	Upper	
NC vs. C	Immediate RC	7.010	.011	-7.323	.001**	4.145	.566	2.963	5.328	2.454
	Delayed RC	22.155	.001	-8.052	.001**	6.179	.767	4.557	7.802	2.788
SC vs. C	Immediate RC	5.162	.028	-8.817	.001**	4.083	.463	3.123	5.042	2.883
	Delayed RC	23.538	.001	-8.675	.001**	6.429	.741	4.863	7.995	2.999
TC vs. C	Immediate RC	1.612	.211	-8.251	.001**	3.239	.392	2.448	4.030	2.431
	Delayed RC	29.109	.001	-6.427	.001**	4.742	.738	3.183	6.300	2.220
NC vs. Pts	Immediate RC	.159	.692	-1.803	.078	1.112	.616	-.130	2.355	.550
	Delayed RC	.323	.573	-3.660	.001**	3.146	.859	1.414	4.878	1.109
SC vs. Pts	Immediate RC	.080	.778	-1.841	.072	1.050	.570	-.099	2.199	.583
	Delayed RC	.239	.627	-4.008	.001**	3.396	.847	1.688	5.103	1.222
TC vs. Pts	Immediate RC	.764	.387	-.370	.713	0.206	.556	-.916	1.328	.119
	Delayed RC	.397	.532	-2.020	.050*	1.708	.846	.004	3.413	.614

** Significant at p < .01; * Significant at p < .05

Abbreviations: NC vs. C: uncoached simulators vs. community controls; SC vs. C: symptom-coached simulators vs. community controls; TC vs. C: test-coached simulators vs. community controls; NC vs. PTS: uncoached simulators vs. clinical patients; SC vs. PTS: symptom-coached simulators vs. clinical patients; TC vs. PTS: test-coached simulators vs. clinical patients

General Discussion

The present studies compared the effectiveness of immediate vs. delayed recall indicators in discriminating between performances of simulators vs. full-effort clinical and non-clinical comparison groups. We used groups of uncoached, symptom-coached, and test-coached simulators to explore the impact of coaching on recall performance. We hypothesized that (1) delayed recall would be superior to immediate recall at detecting invalid performance in the general simulator sample and (2) test-coached simulators would display smaller differences in performance than the other two groups, as compared to genuine patients.

The results of both studies confirmed our first hypothesis. Both indicators yielded significant differences between simulators and non-clinical controls, whilst controlling for age, gender, and education, but delayed recall was more accurate than immediate recall in distinguishing between simulators and patients. As results on this type of contrast are more salient to clinical settings (Vickery et al., 2001), using indicators based on delayed rather than immediate recall would be more appropriate for discriminating noncredible from impaired performance in the assessment of clinical participants (Crişan et al., 2021). Still, ROC curve analyses showed marked differences in the classification accuracy of both indicators across types of contrasts. While in simulators vs. controls, both immediate and delayed recall showed high to excellent AUC values of similar ranges, their accuracy decreased in simulators vs. patients, and sensitivities for both indicators failed to reach the "Larrabee limit" at cutoffs of ≤ 4 (i.e., ≥ 50% sensitivity at ≥ 90% specificity). In both studies, modest AUC values that were statistically not significant for the immediate recall indicator showed that it was less reliable than delayed recall in distinguishing invalid from genuinely dysfunctional responses (see Suhr & Gunstad, 2000; Strauss et al., 2002). However, its limited effectiveness in clinical comparisons imposes caution on interpreting failures in delayed recall as a single indicator of invalid performance, recommending the association with other types of indicators for a more rigorous assessment.

Of note, although cutoffs set at ≤ 8 for both indicators discriminated between simulators and non-clinical controls with sensitivities between 66% and 83%, reaching up to 100% specificity in the second study, they had to be significantly lowered to achieve acceptable specificities (≥ 90%) in the simulator vs. patient contrast. This finding was expectable given the presence of cognitive impairment in the clinical group. At this point, we stress the importance of setting differential cut scores for indicators according to the level of impairment of the full-effort comparison groups. Therefore, besides including a community control group, the presence of a clinical group performing with full-effort is mandatory, as the scores of these patients set a threshold for real impairment below which invalid performance might be suspected (Bender & Rogers, 2004; DiCarlo et al. 2000; Kanser et al., 2018). Therefore, in comparisons with impaired populations, cutoffs must be lowered to keep false positives to a minimum (Green et al., 2011; Merten et al., 2007). Unfortunately, this was achieved at the expense of low sensitivities for both recall indicators. Of note, a cut score of ≤ 4 for the delayed recall indicator yielded the highest sensitivity in the second study (45.8%) pointing to limited effectiveness in simulator vs. patients, which proved nonetheless superior to immediate recall.

We analyzed differences between immediate recall and delayed recall scores across contrasts, to verify the *test-effect* (i.e. how multiple learning trials and recall sessions of test material influenced retention throughout the test). These differences were graphically displayed as performance curves. Results across both studies showed a decrement in recall performance for all simulator groups that varied across studies, while both patients and normal controls revealed a performance increment of approximately 1 p. These results suggested that in full-effort groups, regardless of the presence of cognitive impairment, a learning effect occurred, despite confrontation with a distractor with different stimuli. On the other hand, in groups of feigners, the performance decrement suggested the intentional withholding of memorized items thus indicating noncredible responding. Our findings thus offer input on how the difference between immediate and delayed recall in memory tasks might be a useful indicator in assessment, addressing the lack of evidence noted by Leighton et al. (2014). Our studies also provide a new method of computing performance curve indicators, therefore contributing to knowledge in this field (Rose, Hall & Szalda-Petree, 1998; Wogar et al., 1998; Suhr & Gunstad, 2000; Bender & Rogers, 2004).

All three groups of simulators were discriminated from both full-effort groups by demonstrating descending curves of recall performance, regardless of the coaching type. The fact that test-coached simulators showed a smaller decrement in performance than the other two groups suggested the moderating influence of test-coaching on performance in the sense of bringing simulated test presentations closer to bona fide impairment. In this regard, our findings appear consistent with studies that support the superiority of test-coaching over other types of coaching (Bender & Rogers 2004; Powell et al., 2004; Rüsseler et al., 2008; Weinborn et al., 2012). Our results also indicate the vulnerability of recall measures to test-coaching, and therefore their association with more robust measures (e.g. forced-choice) would be more suitable in the assessment of noncredible performance.

Limitations

Several limitations may be attributed to our studies. Firstly, the employed indicators are experimental and need to be further tested in research before issuing any final statements about their classification accuracies. Secondly, the small sample sizes in both studies and the fact that simulators were recruited among psychology restrict the generalizability of results to other populations. Thirdly and most importantly, the absence of a standard PVT to determine whether the clinical and non-clinical comparison groups performed with full-effort is another limitation. Without it, it could not be stated with certainty that the presentations of clinical patients and community controls were genuine. Therefore, the replication of findings with a criterion-PVT to account for response validity appears mandatory. Future studies should address these limitations by including simulator samples recruited from various populations and replicating experimental results in comparisons of clinical criterion groups.

Conclusion

The present paper compared the accuracy of two experimental validity indicators based on recall memory in detecting noncredible performances of simulators compared with full-effort controls and clinical patients. Results of two experiments highlighted delayed recall as superior to immediate recall in distinguishing simulators from patients, yet low sensitivities pointed to this indicator's limited classification ability in clinical assessment. Comparing immediate vs. delayed recall performances of simulating vs. full-effort participants showed a suppression of the learning effect in all three simulator groups, with test-coached participants exhibiting the smallest difference from scores of genuine patients. Observing the absence of a learning effect or a descending performance curve in examinees may provide some information about an examinee's response validity, however, caution is recommended when interpreting such results in cases where bona fide impairment is present.

Availability of data and material
Data are available upon request.

References

Bender, S.D., Rogers, R. (2004). Detection of neurocognitive feigning: Development of a multi-strategy assessment. *Archives of Clinical Neuropsychology, 19*(1), 49-60. https://doi.org/10.1016/S0887-6177(02)00165-8

Bigler, E. D. (2014). Effort, symptom validity testing, performance validity testing, and traumatic brain injury. *Brain Injury, 28*(13-14), 1-16. https://doi.org/10.3109/02699052.2014.947627

Brennan, A.M., Meyer, S., David, E., Pella, R., Hill, B.D., & Gouvier, W.D. (2009). The vulnerability to coaching across measures of effort. *The Clinical Neuropsychologist, 23*(1), 314-328. https://doi.org/10.1080/13854040802054151

Crișan, I., Maricuțoiu, L.P., & Sava, F.A. (2021). Strategies to detect invalid performance in cognitive testing: An updated and extended meta-analysis. *Current Psychology.* https://doi.org/10.1007/s12144-021-01659-x

Crișan, I., Sava, F.A., Maricuțoiu, L.P., Ciumăgeanu, M.D., Axinia, O., Gîrniceanu, L., & Ciotlăuș, L. (2021). Evaluation of various detection strategies in the assessment of noncredible memory performance: Results of two experimental studies. *Assessment*. Advance online publication. https://doi.org/10731911211040105.

Denning, J.H. (2012). The efficiency and accuracy of the Test of Memory Malingering Trial 1, errors on the first 10 items of the Test of Memory Malingering, and five embedded measures in predicting invalid test performance. *Archives of Clinical Neuropsychology, 27*(3), 417-432. https://doi.org/10.1093/arclin/acs044

DiCarlo, M.A., Gfeller, J.D., Oliveri, M.V. (2000). Effects of coaching on detecting feigned cognitive impairment with the Category Test. *Archives of Clinical Neuropsychology, 15*(5), 399-413. https://doi.org/10.1016/S0887-6177(98)90389-4

Folstein, M.F, Folstein S.E, & McHugh P.R. (1999). Mini-Mental State Examination (MMSE) In Burns, A., Lawlor, B., Craig, S. (Eds.), *Assessment scales in old age psychiatry* (pp. 34-35). Martin Dunitz. (Original work published 1975)

Gorny, I. & Merten, T. (2006). Symptom information – warning – coaching. How do they affect successful feigning in neuropsychological assessment*? Journal of Forensic Neuropsychology,* *4*(4), 71-97. https://doi.org/10.1300/J151v04n04_05

Green, P., Montijo, J., & Brockhaus, R. (2011). High specificity of the Word Memory Test and Medical Symptom Validity Test in groups with severe verbal memory impairment. *Applied Neuropsychology*, *18*(2), 86-94. https://doi.org/10.1080/09084282.2010.523389

Greiffenstein, M.F., Baker, W.J., & Gola, T. (1996). Comparison of multiple methods for Rey's malingered amnesia measures. *Archives of Clinical Neuropsychology, 11*(4), 283-293. https://doi.org/ 10.1016/0887-6177(95)00038-0

Gunner, J.H., Miele, A.S., Lynch, J.K., & McCaffrey, R.J. (2012). The Albany Consistency Index for the Test of Memory Malingering. *Archives of Clinical Neuropsychology, 27*(1), 1-9. https://doi.org/10.1093/arclin/acr089

Inman, T. H., & Berry, D. T. R. (2002). Cross-validation of indicators of malingering: A comparison of nine neuropsychological tests, four tests of malingering, and behavioral observations. *Archives of Clinical Neuropsychology*, *17*, 1–23. https://doi.org/10.1016/S0887-6177(00)00073-1

Kanser, R.J., Rapport, L.J., Bashem, J.R., & Hanks, R.A. (2018). Detecting malingering in traumatic brain injury: combining response time with performance validity test accuracy. *The Clinical Neuropsychologist*, 33(1), 1-18. https://doi.org/10.1080/13854046.2018.1440006

Larrabee, G.J. (2003). Detection of malingering using atypical performance patterns on standard neuropsychological tests. *The Clinical Neuropsychologist, 17*(3), 410-425. https://doi.org/10.1076/clin.17.3.410.18089

Lau, L., Basso, M.R., Estevis, E., Miller, A., Whiteside, D.M., Combs, D., & Arentsen, T.J. (2017). Detecting coached neuropsychological dysfunction: a simulation experiment regarding mild traumatic brain injury. *The Clinical Neuropsychologist*, *31*(8), 1-20. https://doi.org/10.1080/13854046.2017.1318954

Leighton, A., Weinborn, M., & Maybery, M. (2014). Bridging the gap between neurocognitive processing theory and performance validity assessment among the cognitively impaired: a review and methodological approach. *Journal of the International Neuropsychological Society, 20*(1), 873-886. https://doi.org/10.1017/S135561771400085X

Merten, T., Bossink, L., & Schmand, B. (2007). On the limits of effort testing: Symptom validity tests and severity of neurocognitive symptoms in nonlitigant patients. *Journal of Clinical & Experimental Neuropsychology*, *29*(3), 308–318. https://doi.org/10.1080/13803390600693607

Powell, M.R., Gfeller, J.D., Hendriks, B.L, & Sharland, M. (2004). Detecting symptom- and test-coached simulators with the Test of Memory Malingering. *Archives of Clinical Neuropsychology*, *19*(5), 693-702. https://doi.org/10.1016/j.acn.2004.04.001

Rees, L., Tombaugh, T., Gansler, D., Moczynski, N. (1998). Five validation experiments of the Test of Memory Malingering. *Psychological Assessment, 10*(1), 10-20. https://doi.org/10.1037/1040-3590.10.1.10

Rose, F. E., Hall, S., & Szalda-Petree, A. D. (1998). A comparison of four tests of malingering and the effects of coaching. *Archives of Clinical Neuropsychology, 13*(4), 349–363. https://doi.org/10.1016/S0887-6177(97)00025-5

Rüsseler, J., Brett, A., Klaue, U., Seiler, M., Münte, T. (2008). The effect of coaching on the simulated malingering of memory impairment. *BMC Neurology, 37*(8), 1-14. https://doi.org/10.1186/1471-2377-8-37

Sollman, M. J., & Berry, D. T. (2011). Detection of inadequate effort on neuropsychological testing: a meta-analytic update and extension. *Archives of Clinical Neuropsychology*, *26*(8), 774-789. https://doi.org/10.1093/arclin/acr066

Strauss, E., Slick, D. J., Levy-Bencheton, J., Hunter, M., MacDonald, S. W. S., & Hultsch, D. F. (2002). Intraindividual variability as an indicator of malingering in head injury. *Archives of Clinical Neuropsychology, 17*, 423–444. https://doi.org/10.1093/arclin/17.5.423

Suhr, J.A., & Gunstad, J. (2000). The effects of coaching on the sensitivity and specificity of malingering measures. *Archives of Clinical Neuropsychology, 15*(5), 415-424. https://doi.org/10.1093/arclin/15.5.415

Vickery, C.D., Berry, D.T., Inman, T.H., Harris, M.J., & Orey, S.A. (2001). Detection of inadequate effort on neuropsychological testing: A meta-analytic review of selected procedures. *Archives of Clinical Neuropsychology*, *16*(1), 45-73. https://doi.org/10.1093/arclin/16.1.45

Weinborn, M., Woods, S.P., Nulsen, C., & Leighton, A. (2012). The effects of coaching on the Verbal and Nonverbal Medical Symptom Validity Tests. *The Clinical Neuropsychologist*, *26*(5), 832-849. https://doi.org/10.1080/13854046.2012.686630

Wogar, M.A., van den Broek, M.D., Bradshaw, C.M., Szabadi, E. (1998). A new performance-curve method for the detection of simulated cognitive impairment. *The British Journal of Clinical Psychology, 37*(3), 327-339. https://doi.org/10.1111/j.2044-8260.1998.tb01389.x

APPENDIX

Table 1. Simulation instructions

Condition	Instructions
No coaching (NC)	General instructions: "Imagine it's the last week of the final semester and you need to hand in a paper which has a 50% weight in the final evaluation. Without it, you cannot participate in the final examination. You realize you have no possibility of finishing the paper on time, but you know that the university may allow exceptions in cases of accidents followed by residual impairment. You decide to feign a traumatic brain injury (TBI): you declare that you have been involved in a car accident, you have been hospitalized, and told that you had suffered a minor TBI. Gradually you started feeling better but, given the current exam circumstances, you decide to exaggerate your symptoms so that you still appear impaired. A psychologist will hand you a few tests to assess your cognitive state and to check the authenticity of your symptoms. You can avoid the exam situation only if you convince the evaluator that your symptoms are real. You have one week to prepare – you can choose any source of information. Remember: you have to appear truly dysfunctional and convincing in your presentation so that the evaluator does not realize that you are feigning."
Symptom-coaching (SC)	General instructions + "Some of common TBI symptoms include sensorial problems (blurred vision, ringing in the ears, bad taste, changes in smell, sensitivity to light and sound), physical/somatic problems (temporary loss of consciousness, persistent headache, nausea, fatigue, loss of balance), cognitive problems (difficulties in speech and reading, confusion and disorientation, difficulty remembering new information, memory and attention difficulties), affective problems (emotional dysregulations, irritability, feelings of sadness or anxiety)."
Test-coaching (TC)	General instructions + "Some tests are designed to detect feigning, so mind the following aspects: do not fail more than half of the items. Try to fail more difficult tasks rather than easier ones. Tasks that appear easy are usually easy and can be accomplished by patients with genuine cognitive impairment. The performance of cognitively impaired persons is constant, so try to react equally correct/incorrect to all assessment tasks."