International Journal of Business and Social Science Research

ISSN: 2709-2143 **Vol: 7, Issue: 4** **April/2026**
DOI: https://doi.org/10.47742/ijbssr.v7n4p1

Check for Updates

Original Article | Open Access | Peer Reviewed

IARP

A Comparison of Traditional and Non-traditional True-False Measures in a Business Task

Dr. Richard Perlow
Dean
School of Business
MacEwan University
Email: Perlowr@macewan.ca
ORCID ID https://orcid.org/0000-0001-9727-0637
Canada

Lori S. Kopp
Associate Professor
Dhillon School of Business
University of Lethbridge
Email: Lori.Kopp@uleth.ca
ORCID ID https://orcid.org/0000-0002-9371-5843
Canada

ARTICLE INFO

Article History:

Received	: 16 March 2026
Revised	: 8 April 2026
Accepted	: 16 April 2026
Publication	: 30 April 2026
DOI	: 10.47742/ijbssr.v7n4p1

ABSTRACT

Beliefs regarding the usefulness of true-false tests are mixed. Many of these opinions stem from research comparing true-false test performance to that on traditional paper-and-pencil tests. However, little is known about how true-false test scores relate to performance measures requiring knowledge application, or whether different scoring algorithms vary in their ability to predict such performance. To address these gaps, we examined the relationships between traditional and modified true-false scoring methods and outcomes on a business simulation designed to assess complex knowledge application. Our results showed that posttest true-false scores were associated with simulation performance, with the gap between high and low scorers widening over time. Scoring formats that incorporated confidence ratings demonstrated higher reliability and predictive power, but were not substantially more correlated with performance than traditional methods. These findings suggest that true-false tests can serve as effective measures of performance on complex tasks.

KEYWORDS: Business simulation, measurement, scoring formulas, evaluation methods, confidence testing

Introduction

A variety of techniques exist to assess knowledge and the acquisition of complex skills. These assessments generally fall into two categories: objectively scored and subjectively scored measures. Research suggests that objectively scored tests offer several advantages over subjectively scored measures like essay tests (Anastasi, 1988; Thorndike, 1982). Objective measures tend to have superior psychometric properties and can cover a broader range of material than subjective formats, such as essay exams. One common type of objective test uses true-false items.

There are mixed opinions on the usefulness of true-false measures. Some believe the format produces poor measures (Hancock, Theids, Sax, & Michael, 1993; Sax, 1989; Storey, 1966). Others claim the research literature doesn't support the poor reputation of this method (Downing, 1992; Ebel, 1970; Frisbie, 1992). A third group believes that it is premature to draw definitive conclusions on true-false tests (e.g., Gose and Escudero, 1986). As can be seen, most of the literature regarding the debate is dated, and we could not find more recent research addressing the issue. This is unfortunate, given the mixed opinions on its usefulness, as it might lead some to abandon the use of true-false tests in situations where they might be useful, or lead people to use the measures when they should not. There are also gaps in the literature that need to be addressed.

One gap in the literature pertains to the criterion often used to assess the usefulness of true-false measures. Nearly all research has examined the extent to which true-false scores relate to paper-and-pencil test performance. While addressing that issue is important, an alternative research stream is to investigate true-false test relations in tasks that require people to apply acquired knowledge when solving problems, such as those that occur during discovery learning. Discovery learning is a pedagogical technique in which people acquire knowledge through experience and use it to solve problems. Perhaps true-false tests can be a way to assess the accuracy of propositions people develop during learning. Unfortunately, we are aware of no research that has examined the relationship between true-false measures and performance on complex non-paper-and-pencil tasks. We contribute to the testing literature by addressing this fundamental issue.

We also extend the literature by noting that, to our knowledge, no research has assessed whether true-false measures are sensitive enough to detect the application of knowledge acquired during the performance of a complex task over time. People rely on subject-matter knowledge and beliefs to frame a problem and develop problem-solving strategies during task performance. They then use

International Journal of Business and Social Science Research

ISSN: 2709-2143 Vol: 7, Issue: 4 April/2026
DOI: https://doi.org/10.47742/ijbssr.v7n4p1

Check for Updates

Original Article | Open Access | Peer Reviewed IARP

the outcomes of actions based on those strategies to evaluate their basic beliefs, and that evaluation contributes to the development of new knowledge and, hopefully, a better understanding of the task. This sequence of applying knowledge to strategy development and execution, observing the outcome, and using that information to refine or abandon the strategy is fundamental to solving problems (Carver & Scheier, 2001; Newell & Simon, 1972). Documentation of sensitivity to detect learning provides evidence of the usefulness of true-false measures in assessing the efficacy of training programs that deploy, in whole or in part, discovery learning methods and experiential activities.

Rationale

The use of true-false tests is appropriate for assessing certain aspects of problem-solving performance. Given cognitive limitations and/or environmental constraints such as time, people may satisfice by making a decision that is just good enough rather than determining the best possible solution (Simon, 1955, 1956). One example occurs when a supervisor makes a yes-or-no decision about whether a subordinate possesses the knowledge, skills, and/or ability to perform a task. Perhaps yes/no perceptions that lead to subsequent decisions are amenable to assessment with true/false measures.

Supervisors also hold fundamental beliefs. One example is the belief about how to motivate people (e.g., "The best way to motivate people is to tell them to do their best. Setting difficult goals improves performance."). Given that people take mental shortcuts when making decisions (Kahneman & Tversky, 1979; 1984), that people sometimes make either-or decisions, and that people hold certain fundamental beliefs, perhaps the true-false format is a viable way to gather information about decisions and the propositions people hold. In sum, there are occasions when people make absolute judgments, and true-false questions may be useful for assessing those judgments.

H1. True-false test scores are related to performance on a dynamic, complex task.

H2. True-false tests can detect learning that occurs during problem-solving.

Scoring algorithms

One long-standing criticism of true-false tests is that guessing artificially inflates participants' scores (Greene, 1929). This is particularly true for those who have no or little idea of the correct answer to a question (i.e., blind guessing). Guessing is more likely to benefit less knowledgeable respondents because, by guessing on more test items than individuals with superior subject-matter knowledge, the former group capitalizes on the benefits of guessing to a greater degree than more knowledgeable people (Grosse & Wright, 1985). On the other hand, the effects of guessing may be overstated. The odds of guessing correctly on many items are low (Ebel, 1970). Guessing also occurs infrequently (Ebel, 1968; Campbell, 2015). While research has examined methods to understand guessing behavior better (Campbell, 2015; Chiu & Camilli, 2013; Dutke & Barenberg, 2015; Espinosa & Gardeazabal, 2010; Kang, Pashler, Cepeda, Rohrer, Carpenter, & Mozer, 2011; Lau, Lau, Hong, & Usop, 2011; Siddiqui, Bhavsar, Bhavsar & Bose 2016), one gap in the literature is the degree to which scoring algorithms that account for guessing impact test relations with the performance in tasks that people perform over time. This is important in that the test-performance relation assessed only once neither captures changes in learning, or lack thereof, that occur over time nor enables understanding of the temporal relation between test scores and performance. In the present research, we extend the scoring algorithm literature by assessing the relationship among a combination of confidence ratings and scoring algorithms in a business simulation that simulates a dynamic task performed 15 times.

Confidence ratings

One possible way to modify true-false tests to improve their predictive power is to incorporate confidence ratings for each true-false item. Respondents read a question, indicate the accuracy of the statement, and indicate the degree to which they believe their response is correct. Research suggests that this type of measure can be useful for assessing knowledge, as effective learning increases the availability of knowledge and learners' confidence in correctly recalling and applying it (Dutke & Barenberg, 2015). Based on that reasoning, we propose the following hypothesis.

H3: The relation of performance on a dynamic complex task with true-false test scores incorporating confidence ratings will be greater than the relation of performance on true-false test scores that do not incorporate confidence ratings.

Method

Participants[1]

213 undergraduate students enrolled in management and psychology classes received extra credit for participating in this study. Computer malfunctions and missing survey data reduced the sample to 189 participants.

Performance Measure

Participants completed the Furniture Factory computer program, which simulates the role of a special-order manager (Wood & Bandura, 1989a, 1989b). In this computer program, participants assume the position of a special-order manager. The participants' task was to assign three Furniture Factory employees in this computer program to jobs and to motivate them to complete their tasks quickly.[2] Poor employee placement in jobs that didn't match their skill set or poor motivational decisions resulted in the employees taking longer to complete their assigned tasks. The appendix contains both the description of the three employees and the three jobs. As shown in the appendix, the three fictitious employees varied in their skill sets, and the jobs differed in the skills required for successful performance.

The computer program required study participants to make four decisions that affected the time it took for the three employees, as a group, to complete their work. The first decision involved placing employees in jobs that matched their skill sets. Incorrectly assigning an employee to a job resulted in the employee taking longer to complete the job. Participants also decided on the performance goal to give each employee (i.e., a goal equal to the estimated time to do the job, a time goal that was easier than the estimated time to do the job, a goal that was harder than the time estimated to do the job, a "do your best" vague goal, or no goal). The third decision involved the kind of feedback, if any, to give each employee in this simulation after learning how they performed after each trial (i.e., no feedback, the time it took to do the task, the reasons for the performance attained, or both feedback on the time it took to do the task and the reasons for the performance). The fourth decision involved the

[1] Human subjects approval was obtained for this study.
[2] The computer program instructions contained information on the skill sets of nine employees and descriptions of nine jobs but allows the end user to limit the number of employees and jobs. We chose to limit the number of employees and jobs to three.

International Journal of Business and Social Science Research

ISSN: 2709-2143 Vol: 7, Issue: 4 April/2026

DOI: https://doi.org/10.47742/ijbssr.v7n4p1

Check for Updates

Original Article | Open Access | Peer Reviewed IARP

consequence, if any, of giving each worker (i.e., no reward, a verbal compliment, and a posted written acknowledgement). In sum, our study's participants made placement, goal-setting, feedback, and consequence decisions.

The simulation provides information on the effectiveness of the participant's four decisions: a) the hours each of the three employees took to perform their job versus the time it should have been taken to do the job had the participant made correct decisions, b) the total time for all three simulation employees to complete assigned tasks in comparison to the estimated time to do all three jobs, and c) the difference between the actual and estimated time for all employees as a group expressed as a percent. Participants perform well when they correctly apply the principles governing employee performance and learn the correct match between employee characteristics and task demands. Wood, Bandura, & Bailey (1990) present a complete description of the simulation. Wood and Bailey (1985) discuss the program's logical and mathematical foundations.[3]

Better participant job placement and motivation decisions resulted in the computer program's employees completing their jobs in fewer hours. Thus, the faster the employees completed their work, the better the study's participants' performance. As in golf, lower scores (i.e., fewer hours the three employees took to complete their work in this simulation) reflect better performance. Conversely, either assigning the employees to jobs that did not match their skill set and/or deploying poor motivation strategies, the longer it took the employees to complete their jobs. Poorer placement and motivation decisions are reflected in a greater number of hours that the employees in the simulation needed to complete their tasks. Like golf, higher scores reflect poorer participant performance. In our analysis, this means we expect higher true-false test scores to be negatively related to performance on the simulation.

Although participant performance is based on the number of hours it takes employees in the simulation to perform their jobs, it is important to note that this simulation is not a speed test. Participants are not scored on either how quickly they can make decisions or how quickly they can perform the simulation. Rather, the placement and motivation decisions participants make when completing the jobs affect the number of hours it takes the three workers to complete their tasks, as determined by the software's algorithm. The time participants took to make their four decisions had no bearing on performance.

In summary, we operationalized performance as the total time it took the three employees, as a group, to complete the jobs. The better decisions participants made regarding job placement and motivation, the fewer hours it took the employees to complete their job. We did not evaluate how fast participants made their four decisions. We only assessed the quality of the decisions based on the three employees' time to complete their jobs. We assessed performance 15 times. The reliability of the 15-trial criterion measure was excellent ($\alpha = .97$)

Measures

Participants completed two true-false measures during the investigation. The pretest contained two sections. One section comprised a 15-item measure assessing knowledge of fundamental motivation principles (e.g., giving employees specific, difficult goals improves their performance; the only way to get substandard employees to improve their performance is through rewards). After reading the instructions, but before performing the simulation, participants completed a 13-item measure. This measure assessed knowledge of six employee characteristics (e.g., Bert is a skilled metal worker; Dave is a careful worker). These items assessed the characteristics of employees who were not part of the simulation used in the present investigation, thereby avoiding cueing participants. The number of correct true-false items on each measure was either close to or exceeded the number of correct true items because research shows a bias to respond to items as being true when in doubt of the answer to true-false questions (Cronbach,1942). The posttest included items assessing the characteristics of the three employees used in the simulation and questions assessing knowledge of fundamental motivational principles from the pretest.

In addition to answering each true-false item, respondents indicated the confidence they had in their response on a 5-point scale (1 = Not Confident; 5 = Very Confident). Respondents did not receive information on how the confidence ratings affected their scores.

True-false Scoring Schemes

We compared three scoring schemes for true-false tests in this study. The first scoring method is the traditional method (**TTF**). The score is the number of items answered correctly.

The second method assessed participant confidence and correctness for each item on a 0-9 scale (**CTF**). Participants answering an item correctly received 9 points if they had also circled "5" on the confidence rating associated with that question. They received 8 points if they answered correctly and circled "4" on the confidence rating for the item. They received 7 points if they answered correctly and had circled "4" on the confidence rating, and so on. The reverse was true when respondents answered incorrectly. Incorrect responses coupled with greater confidence were associated with lower scores than incorrect answers associated with less confidence. Participants received a score of 4 if they answered an item incorrectly and circled "1" on the confidence rating for that question. They received a score of 3 if they answered incorrectly and circled "2" on the item's confidence rating. Participants received a score of 2 if they answered incorrectly and circled "3" on the confidence rating. Participants received a zero on an item they answered incorrectly and circled "5" on the confidence rating for that question.

The third method was the right-only method (**CTF-RO**). Participants who answered correctly received 5 points if they also circled "5" on the confidence rating for that question. They received 4 points if they answered correctly and circled "4" on the confidence rating for the item, etc. People received a score of zero for items answered incorrectly, regardless of their response confidence. Cronbach alphas were low for all scoring schemes (ranging from .42 (traditional) to .53 (right only) on the pretest and from .49 (traditional) to .59 (right only) on the posttest).

Procedure

Participants completed the first section of the pretest, then read the simulation instructions from a computer monitor. Instructions included a description of the task, of nine employees, and nine tasks. Participants read the instructions at their own pace. They

[3] The task is dynamic. Participant's motivational decisions affect the simulated employees' performance on later trials. Moreover, an incorrect job placement and/or incorrect goal assignment decision on any one trial reduces employee performance improvement stemming from correct motivational decisions made in previous trials.

International Journal of Business and Social Science Research

ISSN: 2709-2143 Vol: 7, Issue: 4 April/2026
DOI: https://doi.org/10.47742/ijbssr.v7n4p1

Check for Updates

Original Article | Open Access | Peer Reviewed IARP

completed the second pretest component immediately after reading the instructions. All participants then received a hard copy of all employee and job information for reference during the simulation. Participants performed the self-paced simulation 15 times. They completed the posttest true-false measure after the last simulation trial.

Design

This study used a mixed-effects design. Performance was the within-subjects' variable. True-false test score was the between-subjects' variable.

Results

Table 1 presents descriptive statistics and intercorrelations among the three measures and the average performance criterion. Data reveal that, unlike the pretest, the posttest was related to performance.

Table 1: Descriptive statistics and intercorrelations

Variable	M	SD	1	2	3	4	5	6	7
1. Mean across 15 trials	85.84	12.27	.97						
2. Pretest Traditional Scoring	15.96	3.07	-.03	.42					
3. Pretest Confidence Scoring	137.29	18.34	-.05	.93	.50				
4. Pretest Right Only Scoring	115.32	24.86	-.02	.94	.96	.53			
5. Posttest Traditional Scoring	14.66	2.88	-.28	.25	.30	.27	.49		
6. Posttest Confidence Scoring	126.44	18.41	-.31	.26	.32	.31	.94	.54	
7. Posttest Right Only Scoring	111.28	24.35	-.30	.25	.32	.33	.94	.97	.59

Note. n = 186. List wise deletion. $p \leq .01$ for correlation values greater than ± .25. Numbers in the diagonal are coefficient alpha values (n=189).

We used Lee and Preacher's (2013) software to calculate the differences between the dependent correlations of each scoring method with performance and present the results in Table 2. Results indicate that the relation of the posttest with performance was stronger than the relation of performance with the pretest.

The differences in the correlation coefficients of performance with each of the three pretest scoring schemes were not meaningful. We draw the same conclusion after examining the correlation coefficients of performance with each of the three posttest scoring schemes. Given that there were no differences across the scoring schemes, we report only the results of our analyses using the traditional method for scoring true-false questions.

Table 2: Difference between pre and posttest correlations with performance

Algorithm	Pretest	Posttest	r difference	Z score
Traditional	-.03	-.28	.25	2.84**
Confidence	-.05	-.31	.26	3.12**
Right Only	-.02	-.30	.28	3.34**

n = 186; ** p < .01 (two-tailed)

Table 3 contains the means of performance for each trial. Inspection of the means suggests two things. First, our means are increasing over time. Recall that the desired performance is fewer hours to complete the jobs, indicating that our participants performed worse over time. One viable explanation for the declining performance pertains to strategy. Ideally, participants should change one parameter at a time (e.g., the nature of the goals) before performing the next trial to determine the effects of the change on the simulated workers' behaviour. Changing multiple parameters (e.g., goals and rewards) makes it difficult to determine the effects of each alteration. Perhaps our participants deployed this and/or other ineffective problem-solving strategies when performing the simulation. We also notice a non-linear trend in the data that we will assess in our analyses.

Table 3: Means and standard deviations across trials

Performance	M	SD
Trial 1	79.6	7.9
Trial 2	78.7	9.1
Trial 3	81.9	11.9
Trial 4	83.8	12.1
Trial 5	84.4	13.6
Trial 6	85.6	13.7
Trial 7	87.0	15.0
Trial 8	87.2	15.6
Trial 9	87.9	16.0
Trial 10	88.0	16.8
Trial 11	88.5	16.6
Trial 12	89.0	16.8
Trial 13	89.3	16.8
Trial 14	89.6	17.6
Trial 15	89.3	17.4

Note. n=189

International Journal of Business and Social Science Research

ISSN: 2709-2143 Vol: 7, Issue: 4 April/2026

DOI: https://doi.org/10.47742/ijbssr.v7n4p1

Check for Updates

Original Article | Open Access | Peer Reviewed IARP

We used the nlme mixed-effects modeling package (Pinheiro, Bates, DebRoy. & Sakar (2020) in R (R Core Team, 2020) when evaluating our hypotheses. Using mixed-effects modeling has an advantage over ANOVA for assessing repeated measures. The latter method assumes that errors are uncorrelated. Our data do not meet that assumption because performance is nested within people, and decisions participants made on earlier trials affect performance on later trials. Failing to account for correlated errors in repeated measures designs can produce biased parameter estimates. Mixed-effects modeling can include correlated error terms in the model equations after developing the appropriate model, prior to hypothesis testing.

Model development

Our first analysis assesses the unconditioned means model to address whether participant performance intercepts vary. The results of that analysis reveal that our intercept (i.e., the grand mean) was 85.99 hours with a standard deviation of 12.01 (95% CI 10.81-13.34). The confidence interval of the standard deviation does not contain zero, which indicates there is significant variance in the intercept to be accounted for. In addition, we created and compared two models: one model where we fixed the participants' intercepts to equal each other, and the other that allowed the intercepts to vary. Results revealed that the random intercept model fit the data better than the fixed intercept model (Likelihood ratio*2 = 2196.09, p<.0001). Given these findings, we will analyze our hypotheses using a random intercepts model.

Our next analysis assesses the unconditioned slope model to determine whether our participants' performance over time follows the same slope or varies among participants. We created and compared two random-intercept models, in which we fixed the slope in one model and allowed it to vary in the other. Results showed that the random slopes model fit the data better than the fixed slope model (Likelihood ratio*2 = 748.03, p < .0001). Our hypothesis testing will include models with both random intercepts and slopes.

We examined whether models including autocorrelations in the error term and another including autocorrelations and increasing variances over time would fit the data better than models without autocorrelations and/or variance differences in performance. Results showed that the model incorporating autocorrelations fit the data better than one that did not (Likelihood ratio*2 = 98.98, p< .0001). Modeling increasing variance in performance yielded an over fitted model that failed to produce parameter estimates.

Finally, we created models specifying non-linear trends in performance over time. The model incorporating a quadratic term fits the data (t = -6.76, df 2644, p<.01); the one incorporating a cubic term did not (t = .21, df 2643, p=.84).

In sum, we conducted several analyses to determine the existence of and account for possible statistical biases that could affect our hypothesis tests. The hypothesized models that we assess below incorporated random intercepts, random slopes, autocorrelations, and a quadratic function.

Hypothesis tests

We conducted our hypothesis tests using each of the three posttest scoring schemes: traditional true-false, confidence interval true-false, and right only true-false. Results from the three mixed-effects modeling analyses revealed there was little difference among the scoring methods. Given this finding, we report below the results using the traditional true-false scoring scheme, as it is the least complex.

In the first step of our analysis, we entered the linear and quadratic trends and the posttest score into the model. Table 4 contains the results. In support of the first hypothesis specifying that true-false test scores were related to performance on a complex task, the posttest was statistically significant (γ= -.51, 95% CI = -.89 -.14, t = -2.70, p = .008).

Table 4: Performance and posttest score relations over time

Source	Estimâtes	CI	t
	Step 1		
Intercept	93.51	87.80 to 99.22	32.10**
Time (Linear)	170.48	133.47 to 207.50	9.03**
Time (Quadratic)	-56.21	-72.50 to -39.92	-6.77**
Test score	-.51	-.89 to -.14	-2.70**
	Step 2		
Intercept	102.96	94.21 to 111.70	23.09**
Time (Linear)	432.68	242.63 to 622.72	4.47**
Time (Quadratic)	-175.01	259.78 to -90.24	-4.05**
Test score	-1.16	-1.74 to -.14	-3.88**
Time (Linear)× Test Score	-17.86	-30.58 to -5.15	-2.76**
Time (Quadratic) × Test Score	8.09	2.42 to 13.76	-2.80**

Note. $^{*}p \leq .05$, $^{**}p \leq .01$.

In the second step, we entered the linear trend x posttest and the quadratic trend x posttest interaction terms in the model. Results reveal that the quadratic trend x posttest interaction contributed unique variance in performance (γ = 8.09, 95% CI = 2.42-13.76, t = 2.80, p=.006). We plotted the performance-fitted values for participants who were at least ±1 standard deviation from the true-false posttest mean. Figure 1 contains that graph. As can be seen in that figure, the curvilinear trend was both lower and less steep for the group that scored better on the true-false posttest than those with better posttest scores. Recalling that higher scores are associated with poorer performance on the simulation, the data suggest that the posttest was sensitive enough to detect the different rates of learning that occurred during performance of this complex task. That result is consistent with the second hypothesis.

International Journal of Business and Social Science Research

ISSN: 2709-2143 Vol: 7, Issue: 4 April/2026

DOI: https://doi.org/10.47742/ijbssr.v7n4p1

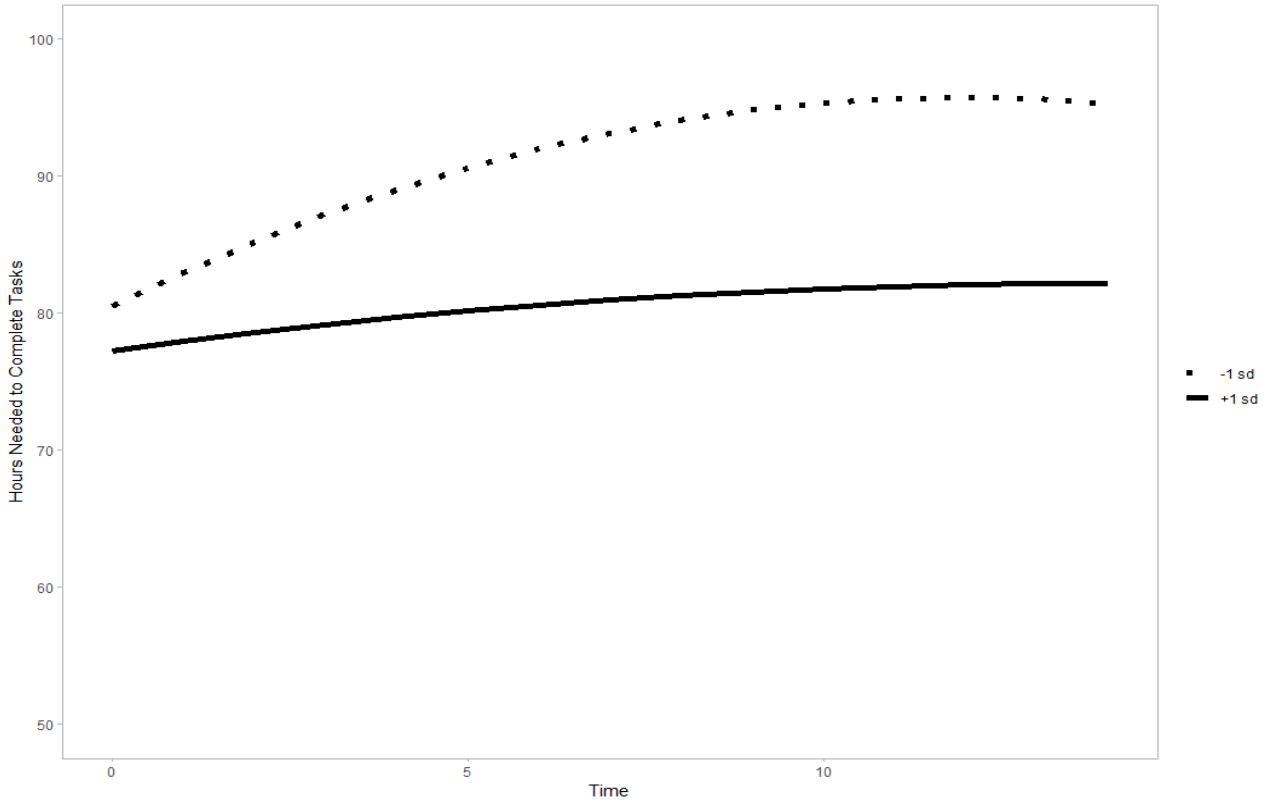
Check for Updates

Original Article | Open Access | Peer Reviewed

IARP

As previously mentioned, the results of the analyses of the two other scoring schemes were consistent with the traditional scoring method but did not appear markedly more predictive. Thus, we did not find support for the third hypothesis.

Figure 1: Traditional True-false Test x Time Interaction



Note. Fewer hours to complete tasks reflects better performance

Discussion

Results show that true-false measures assessing people's beliefs are useful. Participants who obtained higher posttest scores performed better in a business simulation than participants with lower posttest scores. The performance gap between these two groups grew over time. Data also reveal that differences in prior knowledge of the concepts measured in the true-false test did not explain the performance differences between the high and low-posttest score groups. However, we found no difference in the usefulness of different true-false scoring algorithms across groups.

It may seem paradoxical when we state that true-false tests are useful predictors of performance, given the pretest-performance relations we found when comparing groups with different pretest scores. Prior knowledge of motivational principles and employees' capabilities, which participants acquired after reading the simulation's instructions, was not related to later simulation performance. This is not surprising because participants neither received formal instruction on techniques designed to enhance performance nor did they have the opportunity to study the material describing employee characteristics intensively. What is meaningful is that the scores people received on the true-false posttest measure reflected knowledge acquired during performance of the computer simulation. Beliefs held by some participants changed during the simulation, and the posttest was sensitive enough to detect that change and distinguish between participants whose beliefs shifted and those who did not.

A frequent criticism of true-false tests is that respondents have a 50-50 chance of correctly responding to a test item when they guess blindly, thereby inflating their scores and reducing the validity of inferences drawn from the scale's scores. Scholars have developed scoring algorithms based on confidence ratings to reduce bias. However, our research suggests that scores based on confidence ratings did not markedly increase the relationship with performance compared to the traditional true-false scoring scheme in a dynamic environment. This is an important practical finding for assessors because the time and effort required to incorporate confidence ratings into a true-false measure, and the time test-takers require to complete the measure, did not translate into markedly better predictive power.

Although we believed that confidence ratings would improve test-performance relations over scoring algorithms that do not incorporate such ratings, the data did not support our hypothesis. Perhaps a decrease in confidence in an incorrect answer did not imply or immediately lead to the adoption of a correct strategy. Rather, disregarding an incorrect strategy may lead to increased confidence in adopting another incorrect strategy. Thus, any attempt to quantify degrees of not knowing, as is done with point deductions for wrong answers, may not lead to improved performance prediction. Being somewhat confident in an incorrect strategy leads to no better performance than being very confident in an alternative strategy that is also incorrect. Thus, our data is consistent with the notion that there should be no difference in test usefulness when the degree of wrongness is assessed.

The same explanation may not apply to correct responses. With correct responses and confidence in them, attention and effort can be directed at solving other aspects of the problem. We expect that greater confidence in correct strategies and their use leads to feedback confirming their correct use. In turn, confidence in that strategy increases, and that confidence, along with subsequent responses, leads to testing and verification of strategies aimed at other aspects of the problem. This might, in part, account for the differences in problem-solving strategies and behaviors between

International Journal of Business and Social Science Research

ISSN: 2709-2143 Vol: 7, Issue: 4 April/2026
DOI: https://doi.org/10.47742/ijbssr.v7n4p1

Check for Updates

Original Article | Open Access | Peer Reviewed IARP

experts and novices identified in models of expertise (Chi, Feltovich, & Glaser, 1981; Anderson, 1987).

Contributions

Using a business simulation, we extend the true-false test literature. Most true-false research is dated, and much of it correlated the true-false measure with other academic paper-and-pencil tests. While that stream of research is informative, it does not address the extent to which true-false test scores are useful for assessing knowledge acquired through experience, as we do, particularly in real time. This is important because people develop and test propositions during dynamic problem-solving. After observing the outcomes of decisions informed by their underlying beliefs, people modify, discard, or create new tactics to solve the problem at hand. By assessing performance in a dynamic environment over time, we documented the usefulness of true-false tests in a way that correlates true-false tests with other academic knowledge tests not done in a one-time administration. To our knowledge, incorporating a performance-based criterion into a repeated-measures design to analyze the usefulness of true-false tests has not been done before.

The difference between the pre and post-test scores suggests that true-false tests can detect learning from a simulation that requires people to discover principles of motivation, feedback, and consequences through the feedback it provides. To our knowledge, this has not been previously documented. The finding is important because our results show that instructors deploying the discovery learning method as a pedagogical technique by itself or coupled with other teaching methods can rely on true-false measures to assess propositions people possess as well as pedagogical efficacy.

Limitations and future research

There are limitations that constrain the generalizability of our findings. One limitation is that we did not assess cultural differences among the participants. This is important because cultural differences may exist in item response sets as documented in survey research (e.g., Dolnicar & Grün, 2007). Avoiding generalizations of group data to all members of any one culture and using country data as a proxy for inferences about cultures notwithstanding (Brewer & Venaik, 2014; Tsui, Nifadkar, & Ou, 2007; Venaik & Brewer, 2016) future research should investigate specific response variables to have a better understanding of how people respond to true-false testing and how the propensity to respond to items impact the validity of inferences based on test scores.

We did not assess gender effects. Research documents that women are more risk-averse than men when there is a reward for not answering test questions (Iriberri & Rey-Biel, 2020) and for omitting questions when there is a penalty for providing a wrong answer (Balart, Ezquerra, and Hernandez-Arenaz, 2020; Baldiga, 2014; Espinosa & Gardeazabal, 2020). Research also shows that pressure moderated gender effects on test performance, with women performing better than men under low-pressure conditions but worse when the stakes for successful performance were high (Montolio & Taberner, 2021). Although participants were unaware of how their confidence ratings affect scoring, one avenue for future research is to examine gender effects on confidence judgments and what role the judgments might have in predicting performance using a dynamic, complex activity such as the one we deployed in our study.

Research has documented the usefulness of multiple true-false tests (Brassil & Couch, 2019; Couch, Hubbard, Brassil, 2018; Frisbie, 1992; Frisbie & Sweeney, 1982). In this format, respondents are presented with a question and several options in a multiple-choice format. Their task is to indicate the veracity of each option. Future research could assess the extent to which our results, using a non-paper-and-pencil performance measure as a criterion, generalize to this multiple-choice format.

The relations obtained in this study are modest, in part, due to the scales' internal consistency. Reliability constrains relations among variables (Hunter and Schmidt, 1990). One possible reason for the low internal consistency values we found here includes how people learned the principles underlying successful performance. Participants' acquisition of human resource allocation and performance management principles was through self-discovery. Given that people don't learn as well with self-discovery pedagogy as with direct instruction and that enhanced discovery methods yielded better learning than comparison methods (Alfieri, Brooks, Aldrich, & Tenenbaum, 2011), future research can test the results found here using other methods.

Our performance measure utilized an attainment criterion. We did not assess how quickly participants made decisions or completed the simulation's 15 trials. We based performance on the number of hours the simulation employees spent completing their assigned tasks, which were determined by the decisions the study's participants made. Reaction time, or how quickly study participants can complete an activity, is a performance criterion for some tasks. Additional research examining true-false measures' relations with a variety of performance criteria would enable scholars and practitioners to better assess the usefulness of true-false test scores in a variety of situations.

References

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based learning instruction enhance learning. *Journal of Educational Psychology*, *103*(1), 1-18. https://doi.org/10.1037/a0021017

Anderson, J. R. (1987). Skill acquisition: Compilation of weak-method problem solutions. *Psychological Review*, 94, 192 -210. https://psycnet.apa.org/doi/10.1037/0033-295X.94.2.192

Anastasi, A. (1988). *Psychological testing (6th ed.)*. Upper Saddle River, NJ. Prentice-Hall.

Balart, P., Ezquerra, L. & Hernandez-Arenaz, I. (March 18, 2020). Framing effects on risk-raking behavior: Evidence from a field experiment. Available at SSRN: https://ssrn.com/abstract=3556710 or http://dx.doi.org/10.2139/ssrn.3556710

International Journal of Business and Social Science Research

ISSN: 2709-2143 **Vol: 7, Issue: 4** **April/2026**
DOI: https://doi.org/10.47742/ijbssr.v7n4p1

Check for Updates

Original Article | Open Access | Peer Reviewed IARP

Baldiga, K. (2014). Gender differences in willingness to guess. *Management Science*, *60*(2), 434-448. http://dx.doi.org/10.1287/mnsc.2013.1776

Brassil, C. E., & Couch, B. A. (2019). Multiple-true-false questions reveal more thoroughly the complexity of student thinking than multiple-choice questions: a Bayesian item response model comparison. *International Journal of STEM Education*, *6*, 16, 1-17. https://doi.org/10.1186/s40594-019-0169-0

Brewer, P., & Venaik, S. (2014). The ecological fallacy in national culture research. *Organizational Studies*, *35*(7), 1063-1086. https://doi.org/10.1177%2F0170840613517602

Campbell, M. L. (2015). Multiple-choice exams and guessing: Results from a one-year study of general chemistry tests designed to discourage guessing. *Journal of Chemical Education*, *92*, 1194-1200. https://doi.org/10.1021/ed500465q

Carver, C. S., & Scheier, M. F. (2001). *On the self-regulation of behavior*. Cambridge, United Kingdom: Cambridge University Press.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*, 121-152. https://doi.org/10.1207/s15516709cog0502_2

Couch, B. A., Hubbard, J. K., & Brassil, C. E. (2018). Multiple–true–false questions reveal the limits of the multiple–choice format for detecting students with incomplete understandings. *BioScience*, *68*(6), 455–463. https://doi.org/10.1093/biosci/biy037

Cronbach, L. J. (1942). Studies of acquiescence as a factor in the true-false test. *Journal of Educational Psychology*, *33(6)*, 401-415. https://psycnet.apa.org/doi/10.1037/h0054677

Dolnicar, S. & Grün, B. (2007). Cross-cultural differences in survey response patterns. *International Marketing Review*, *24 (2)*, 127-143. https://doi.org/10.1108/02651330710741785

Downing, S. M. (1992). True-false, alternate-choice, and multiple-choice items. *Educational Measurement-Issues and Practice*, *11(3)*, 27-30. https://doi.org/10.1111/j.1745-3992.1992.tb00248.x

Dutke, S., & Barenberg, J. (2015). Easy and informative: Using confidence-weighted true-false items for knowledge tests in psychology courses. *Psychology Learning and Teaching*, *14(3)*, 250-259. https://doi.org/10.1177%2F1475725715605627

Ebel, R. L. (1968). Blind guessing on objective achievement tests. *Journal of Educational Measurement*, *5(4)*, 321-325. http://www.jstor.org/stable/1433785

Ebel, R. L. (1970). The case for true-false items. *School Review*, *78(3)*, 373-389. http://www.jstor.org/stable/1084159

Espinosa, M. P., & Gardeazabal, J. (2010). Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical Psychology*, *54*(5), 415–425. https://doi.org/10.1016/j.jmp.2010.06.001

Espinosa, M. & Gardeazabal, J. (2020). The Gender-bias Effect of Test Scoring and Framing: A Concern for Personnel Selection and College Admission. *The B.E. Journal of Economic Analysis & Policy*, *20*(3), 20190316. https://doi.org/10.1515/bejeap-2019-0316

Frisbie, D. A. (1992). The multiple true false format: A status review. *Educational Measurement-Issues and Practice*, *11(4)*, 21-26. https://doi.org/10.1111/j.1745-3992.1992.tb00259.x

Frisbie, D. A., & Sweeney, D. A. (1982). The relative merits of multiple true-false achievement tests. *Journal of Educational Measurement*, *19(1)*, 29-35. URL: http://www.jstor.org/stable/1434916.

Greene, E. B. (1929). Achievement and confidence on true-false tests of college students. *Journal of Abnormal and Social Psychology*, *23*, (4), 467-478. https://psycnet.apa.org/doi/10.1037/h0072335

Gose, M. D., & Escudero, R. M. (1996). Whether to use true-false items. *Educational Research Quarterly*, *20*, 37-47.

Grosse, M., & Wright, B. D. (1985). Validity and reliability of true-false tests. *Educational and Psychological Measurement, 45(1)*, 1-13. https://doi.org/10.1177%2F0013164485451001

Hancock, G. R., Thiede, K. W., Sax, G., & Michael, W. B. (1993). Reliability of comparably written two-option multiple-choice and true false test items. *Educational and Psychological Measurement, 53(3)*, 651-660. https://doi.org/10.1177%2F0013164493053003006

Hofstede, G. (1984). *Culture's consequences: International differences in work-related values*. Newbury Park, CA: Sage.

Hunter, J. E., & Schmidt, F. L. (1990). Methods of meta-analysis: Cumulating research findings across studies. Sage.

Iriberri, Nagore & Rey-Biel, Pedro. (2021). Brave boys and play-it-safe girls: Gender differences in willingness to guess in a large scale natural field experiment. *European Economic Review*, *131*(2), 103603. doi: 10.1016/j.euroecorev.2020.103603.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decisions under risk. *Econometrika*, *47*, 263–291. doi:10.2307/1914185

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, *39*, 341–350. https://psycnet.apa.org/doi/10.1037/0003-066X.39.4.341

Kang, J. K., Pashler, H., Cepeda, N. J., Rohrer, D., Carpenter, S. K., & Mozer, M. C. (2011). Does incorrect guessing impair fact learning? *Journal of Educational Psychology*, *103(1)*, 48-59. https://psycnet.apa.org/doi/10.1037/a0021977.

Lau, P. N. K., Lau, S. H., Hong, K. S., & Usop, H. (2011). Guessing, partial knowledge, and misconceptions in multiple-choice tests. *Educational Technology & Society*, *14 (4)*, 99–110. http://www.jstor.org/stable/jeductechsoci.14.4.99

International Journal of Business and Social Science Research

ISSN: 2709-2143 **Vol: 7, Issue: 4** **April/2026**
DOI: https://doi.org/10.47742/ijbssr.v7n4p1

Check for Updates

Original Article | Open Access | Peer Reviewed IARP

Lee, I. A., & Preacher, K. J. (2013, September). Calculation for the test of the difference between two dependent correlations with one variable in common [Software]. Retrieved from http://quantpsy.org/corrtest/corrtest2.htm on May 3, 2021.

Montolio, D., & Taberner, P. A. (2021). Gender differences under test pressure and their impact on academic performance: A quasi-experimental design. Journal of Economic Behavior and Organization *191*, 1065–1090. https://doi.org/10.1016/j.jebo.2021.09.021

Newell, A., & Simon, J. (1972). *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

Pinheiro J., Bates D., DebRoy S., & Sarkar D. (2020). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-144, https://CRAN.R-project.org/package=nlme.

R Core Team (2020). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria. URL https://www.R-project.org/.

Sax, G. (1989). *Principles of educational and psychological measurement and evaluation* (3rd ed.). Belmont, CA: Wadsworth.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, *59*, 99–118. https://doi.org/10.2307/1884852

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*, 129–138. https://psycnet.apa.org/doi/10.1037/h0042769

Storey, A. G. (1966). A review of evidence or the case against the true-false item. *Journal of Educational Research, 59(6),* 282-285. https://doi.org/10.1080/00220671.1966.10883357

Siddiqui, N.I., Bhavsar, V.H., Bhavsar, A.V. & Bose, S. Contemplation on marking scheme for Type X multiple choice questions, and an illustration of a practically applicable scheme. *Indian Journal of Pharmacology*, 48(2): 114–121. https://doi.org/10.4103/0253-7613.178836

Thorndike, R. L. (1982). *Applied Psychometrics*. Boston: Houghton:Mifflin.

Tsui, A., Nifadkar, S. S., & Ou, A. Y. (2007). Cross-national, cross-cultural organizational behavior research: Advances, gaps, and recommendations. *Journal of Management, 33*(3), 426-478. https://doi.org/10.1177%2F0149206307300818

Venaik, S., & Brewer, P. (2016). National culture dimensions: The perpetuation of cultural ignorance. *Management Learning*, *47*(5), 563-589. https://doi.org/10.1177%2F1350507616629356

Wood, R. E., & Bandura, A. (1989a). Social cognitive theory of organizational management. *Academy of Management Review*, *14*(3), 361-384. https://doi.org/10.5465/amr.1989.4279067

Wood, R. E., & Bandura, A. (1989b). Impact of conceptions of ability on self-regulatory mechanisms and complex decision making. *Journal of Personality and Social Psychology, 56(3),* 407-415. http://dx.doi.org/10.1037//0022-3514.56.3.407

Wood, R. E., Bandura, A., & Bailey, T. (1990). Mechanisms governing organizational performance in complex decision-making environments. *Organizational Behavior and Human Decision Processes, 46(2),* 181-201. https://doi.org/10.1016/0749-5978(90)90028-8

Wood, R. E. & Bailey, T. C. (1985). Some unanswered questions about goal effects: A recommended change in research methods. *Australian Journal of Management, 10*, 61-73. https:/ /doi.org/10.1177%2F031289628501000105