Evaluation of Learning Outcomes Through Multiple Choice Pre- and Post-Training Assessments

Thomas Samuel[1], Razia Azen[2] & Naira Campbell-Kyureghyan[1]

[1] Department of Industrial and Manufacturing Engineering, University of Wisconsin – Milwaukee, Milwaukee, USA

[2] Department of Educational Psychology, University of Wisconsin – Milwaukee, Milwaukee, USA

Correspondence: Naira Campbell-Kyureghyan, Department of Industrial and Manufacturing Engineering, University of Wisconsin – Milwaukee, Milwaukee, USA 53201. E-mail: campbeln@uwm.edu

Abstract

Training programs, in industry, are a common way to increase awareness and change the behavior of individuals. The most popular way to determine the effectiveness of the training on learning outcomes is to administer assessments with Multiple Choice Questions (MCQ) to the participants, despite the fact that in this type of assessment it is difficult to separate true learning from guessing. This study specifically aims to quantify the effect of the inclusion of the 'I don't know' (IDK) option on learning outcomes in a pre-/post-test assessment construct by introducing a 'Control Question' (CQ). The analysis was performed on training conducted for 1,474 participants. Results show a statistically significant reduction in the usage of the IDK option in the post-test assessment as compared to the pre-test assessment for all questions including the Control Question. This illustrates that participants are learning concepts taught in the training sessions but are also prone to guess more in the post-test assessment as compared to the pre-test assessment.

Keywords: training assessment, multiple choice question, I don't know, control question, adult learning, guessing behavior

1. Introduction

Training individuals is a common way for organizations to increase the knowledge of their workforce in specific competencies. Based on the Industry Report from 2000, US organizations with 100 or more employees budgeted to spend $54 billion in formal training (Arthur Jr., Bennett Jr., Edens, & Bell, 2003). These trends for formal training are also observed in Australia (Bahn & Barratt-Pugh, 2012) and have been reported to play an important role in how companies perceive that they can improve the safety of their employees and reduce incident rates. Overall in 2014 worldwide corporate spending on training was estimated at $170 billion (Bersin, 2014). As a significant amount of money is being dedicated annually around the globe to employee skill development and required changes in behavior, it is important to measure and verify the impact of the training. As a best practice for validating the benefits of training to the organizations, researchers agree on the importance of evaluating training effectiveness (Alliger & Janak, 1998). Although training programs are utilized worldwide (Calle, 2016), evaluation of the training methods is limited in non-Western countries (Ningdyah, 2018).

Of the many methods that can be used to measure the effectiveness of training, Kirkpatrick's model (Kirkpatrick, 1967) remains the one most frequently utilized by trainers (Arthur Jr. et al., 2003). The model consists of 4 evaluation levels as follows:

Level 1—Reaction: Assessed by asking the trainees how they liked and felt about the training

Level 2—Learning: Assessed by results of traditional tests of declarative knowledge

Level 3—Behavior: Assessed by on-the-job performance (i.e., work samples, work outputs and outcomes)

Level 4—Results: Assessed by organizational impact (i.e., productivity gains, customer satisfaction, cost savings).

Kontoghiorghes (2001) demonstrated that learning in a training setting, as measured by post-test assessments, is a good predictor of how people will apply their knowledge in their work environment. It was also shown that

there is a high correlation between the retention of the training material after training and follow up post-test scores. The author concluded that this can be considered a significant finding given that it statistically validates the importance of the training evaluation component, as has been advocated by many human resource development theorists. This finding also suggests that trainees will be more motivated to learn during training if they know that they are accountable for the training that they receive (Kontoghiorghes, 2001). Similarly, the methods used in the training are also important to help drive the required change in knowledge, attitude and behavior among the trainees. A meta-analysis of training related literature conducted by Burke and Baldwin (1999) concluded that any method that encourages engagement, dialog, and participation of the training participants was more effective than passive methods of training delivery like lectures, online training, and so on. A study by Campbell-Kyureghyan, Principe and Ahmed (2013) found that this method of participatory training, where participants can directly relate the learned material to their jobs, was shown to be effective at reducing work-related injuries. Campbell-Kyureghyan, Ahmed and Beschorner (2013) more importantly observed that dynamic work environments, where traditional approaches of workstation redesigns are not effective, are environments where there is an increased need for contextualized safety and ergonomic training to provide awareness, enhance knowledge, and change the attitude and behavior of the participants as it relates to job site safety.

Immediate post-training assessments of learning, Kirkpatrick's Level 2 assessment, are a common training practice. Knowledge is assessed by multiple choice test responses, answers to open-ended questions, listing of facts, and so forth. That is, trainees are asked to indicate, in one of several ways, how much they know about the topics trained. Alliger and Janak (1998) and Newble, Baxter and Elmslie (1979) indicate that traditional tests in the form of multiple choice questions are by far the most common to assess the knowledge gained

One of the frequent criticisms of Multiple Choice Question (MCQ) assessments is that they enable examinees to answer correctly by guessing. Many trainers and companies view any score gain from guessing as an incorrect representation of the participant's knowledge, which can negatively affect the participant's performance in a job environment. Also, multiple choice scores are generally perceived to be too high because scores from comparable short-answer or fill-in-the-blanks tests were found to be lower (Newble et al., 1979). Thus, it is important to have a grading procedure that accurately estimates the true score of the individual by accounting for guessing (Frary, 1998). Guessing can be interpreted by the illustration in Figure 1 and is defined here by the scenario where a participant does not know the answer yet answers the MCQ correctly. This is troublesome because guessing the correct answer artificially increases the score of the participant and is not an accurate measure of the participant's knowledge level of the subject. Hence, in any MCQ assessment, it is desirable to minimize the cases where the participant does not know the answer and yet answers correctly.

Figure 1. Outcomes of MCQ based answers based on the participant knowledge level

An extension of the post-test assessment model is defined by a pre-/post-test model (Level 2) that is essentially assessing participants twice. The pre-test assessment is administered before the training to gage the initial level of knowledge the participant has (baseline), and the post-test assessment is administered after the delivery of the training to gage the increase in knowledge due to training. Initial and final scores of the participants are tracked to determine change in assessment scores. Warr, Allan and Birdi (1999) observed that it is preferable to measure training outcomes in terms of changes from pre-test to post-test, rather than merely through post-test only scores, as this explains individual learning and an understanding of how different trainees have changed as a result of

their experiences. This is because there are often prior differences between trainees in the level of competence that they bring to the training. Although there is literature to illustrate methods to calculate score gains (Campbell, Stanley, & Gage, 1963; Herbig, 1976; Hendrix, Carter, & Hintze, 1978; Brogan & Kutner, 1980; van der Linden, 1981; Warr et al., 1999; Dimitrov & Rumrill, 2003; Arthur Jr. et al., 2003), there is a gap in the body of knowledge on using the pre-test/post-test method to predict correct guessing of answers on training assessments.

As a method to minimize guessing, a number of authors have suggested adding an 'I don't know' (IDK) option to the true-false answer choices in MCQ assessments (Sanderson, 1973; Newble et. al., 1979; Courtenay & Weidemann, 1985; Hammond, McIndoe, Sansome, & Spargo, 1998; van Mameren & van der Vleuten, 1999; Spears & Wilson, 2010). For example, van Mameren and van der Vleuten (1999) suggested the formula (total # correct answers)—(total # incorrect answers) for the score, with no penalty for IDK answers. Research conducted by Courtenay and Weidemann (1985) indicates that inclusion of the IDK option reduces the overall score of the respondents by 2.5% to 3.4% depending on the tests that were administered and decreases the percentage of questions that are answered incorrectly. Thus, the use of the IDK option is believed to compensate for guessing and increase the likelihood of a more accurate score.

A majority of the research on the IDK option has been conducted in the context of True or False (T/F) type questions (Sanderson, 1973; Newble et al., 1979; Courtenay & Weidemann, 1985; Hammond et al., 1998; van Mameren & van der Vleuten, 1999; Spears & Wilson, 2010). The work by Newble et al. (1979) included 19 multiple choice items in a post-test only assessment with an IDK option, but a gap in knowledge still exists on how the IDK option applies to MCQ with more than 2 options in a pre-/post-test assessment model. Therefore, the main goal of the research paper is to investigate and quantify the effect of the IDK option on guessing in an MCQ pre- and post-training assessment model.

The specific research questions (RQ) this study aims to answer are:

- RQ #1: How does the addition of the IDK option in the pre-test Level 2 MCQ assessment change the proportion of correct and incorrect answers?

o With the addition of the IDK option, we would expect the percentage of correct answers to stay the same and the percentage of incorrect answers to be reduced.

- RQ #2: How does the addition of the IDK option in the post-test Level 2 MCQ assessment change the proportion of correct and incorrect answers?

o With the addition of the IDK option, we would expect a reduction in the percentage of correct answers and a reduction in the percentage of incorrect answers.

- RQ #3: Does the addition of the IDK option truly reduce the amount of guessing in pre-test and post-test assessments?

o With the addition of the IDK option, we would expect participants to choose the IDK option instead of guessing on questions to which they do not know the answer.

- RQ #4: If the participant chooses IDK in the pre-test assessment, is there a difference in how that participant responds on the post-test assessment depending on the type of question (MCQ or a Control Question—CQ) —Details of the CQ are discussed in detail in the 'Methods' section below.

o For an MCQ, we would expect most of the participants to answer correctly in the post-test assessment if they answered IDK in the pre-test assessment.

o For a CQ, we would expect that most of the participants to answer IDK in the post-test assessment if they answered IDK in the pre-test assessment.

2. Method

A novel training method on workplace safety and ergonomics was developed for multiple sectors of the utility industry under a DOL Susan Harwood Training Grant by the team of researchers at the University of Wisconsin-Milwaukee. Training content was developed from a combination of onsite assessment observations, employee and management interviews, management concerns, ergonomic principles, nationwide injury and fatality records specific to the utility industry and known problematic operations and tasks. Table 1 illustrates the number of companies and participants that were trained in the three energy utility sectors.

Table 1. List of the number of companies and training participants in each industry

UTILITY SECTOR	# OF COMPANIES	# OF PARTICIPANTS	PARTICIPANT ROLE
Natural Gas	16	*Tier 1:* 500	Employee: 414
			Manager: 86
		Tier 2: 375	Employee: 375
Electric Transmission	15	*Tier 1:* 61	Employee: 54
			Manager: 7
		Tier 2: 359	Employee: 359
Power Generation	4	*Tier 1:* 22	Employee: 8
			Manager: 14
		Tier 2: 157	Employee: 157

To understand and re-define the ergonomic risks, particularly specific to small business utilities, onsite visits were conducted rather than relying solely on general ergonomic principles that are relevant to that utility. Data was gathered from managers/employee interviews and direct observation of all performed tasks using videotaping methods. Since the recruited utilities provide different services, utilize different tools, and are exposed to various ranges of risk-factors, the onsite visits identified the specific ergonomic risks and safety concerns of interest for each facility. The collected information was analyzed and combined with information acquired from nationwide injury and fatality statistics for the utility industry. The basic ergonomic risk factors and safety concerns present in utilities were identified from the observational data (Campbell-Kyureghyan et al., 2013).

The onsite training was split up into two distinct categories. Tier 1 training was conducted by the individuals who conducted the onsite visit and developed the training content. Tier 2 training was conducted by individuals who had participated in a train-the-trainer program conducted by the Tier 1 trainers. In each company both employees and managers were trained and their respective counts are detailed in Table 1. All employees received a base training of 4–5 hours. In addition, managers received an extra 2 hours of training specific to workplace risk assessment and program implementation. It is to be noted that Tier 1 trainers delivered first-hand training to both employees and managers, and Tier 2 trainers conducted primarily employee training.

2.1 Training Content

Newly developed content was based on research that specifically targeted the areas of safety and ergonomics of companies, utilities and contractors. All examples and applications in the training were based on the medium to high risk of injury utility-specific tasks that were observed and assessed with the applicable ergonomic methods and tools during onsite visits. Risk factors were classified into the following categories: physical factors such as lifting heavy loads, pushing/pulling, exposure to vibration, or awkward postures, and environmental factors such as exposure to heat or cold, noise, or slippery conditions. The training materials were organized in separate modules: slips/trips/falls, overexertion/repetitive injuries, noise, environment, PPE, and vehicle safety. The materials were developed with a diverse audience in mind, including some employees with less than a high school education or with English as a second language.

2.2 Training Assessments

Out of Kirkpatrick's 4 levels of assessments mentioned previously, only the first 2 levels are used in the current study. Due to a very diverse range of trainees with respect to prior competence on ergonomic concepts, years of experience, learning skills, etc., a pre-test and post-test model of training assessment was used.

The mode of training for all session was face to face with the number of participants ranging from 6–40. Both pre-test (baseline) and post-test assessments, using MCQ items, were administered to determine the knowledge of the delivered content that each individual acquired. Participants for all the training sessions were required to complete a 10–15-minute pre-training assessment (pre-test) as soon as they arrived for the training. Once the pre-test assessment was completed by all the participants, they were collected by a training team member for further analysis and the training session commenced. Upon completion of the training the same assessment items were administered to the participants post-test. Table 2 illustrates the number of multiple choice questions in the pre-test and post-test training assessments for each of the utility sectors based on the role of the participant.

Table 2. List of the number of assessment questions for managers and employees in each utility sector

UTILITY SECTORS	PARTICIPANT ROLE	# OF MCQs IN ASSESSMENT
Natural Gas	Employee	7
	Manager	7
Electric Transmission	Employee	9
	Manager	12
Power Generation	Employee	10
	Manager	13

Finally, the participants were given a Level 1 training reaction assessment that consisted of eight questions to determine the training quality, trainer quality, training material, training process, and the intent of the individuals to apply their new knowledge to their work environment.

2.3 Knowledge Testing

Control question (CQ) and IDK option: One question in both the pre- and post-test assessments was a question that was contextually similar to the content being trained in the session; however, that specific item was not covered in the training class. For example, the content of the training consisted of information applicable to most common risk factors present in every energy utility sector (natural gas and electric transmission and power generation) such as: slips/trips/falls, overexertion/repetitive injuries, noise, environment, PPE, and vehicle safety. For the assessment, the control question was NOT related to the content of the training, such as application of the NIOSH lifting equation in the case of employee training, and the selection of appropriate anthropometric measurements for office furniture design in the case of management training. In the CQ model, it is reasonable to assume that a correctly answered Control Question is not a consequence of the training, but rather can be explained by prior knowledge, or guessing.

During the pre-test and post-test assessments for the electric transmission and power generation utility sectors, participants were given an additional 'I don't know' option for each MCQ in addition to the CQ. Participants were instructed to choose the 'I don't know' options instead of guessing at the answers in both assessments. Table 3 summarizes the usage of the CQ and the 'I don't know' option in the various assessments for each utility sector.

Table 3. Usage of CQ and IDK option in MCQ assessments by utility sector

UTILITY SECTOR	TRAINEE TYPE	MCQ ASSESSMENT	
		CQ	IDK
Natural Gas	Tier 1 employee	x	
	Tier 1 Manager	x	
	Tier 2 employee	x	
Electric Transmission	Tier 1 employee		x
	Tier 1 Manager		x
	Tier 2 employee	x	x
Power Generation	Tier 1 employee	x	x
	Tier 1 Manager	x	x
	Tier 2 employee	x	x

2.4 Analysis

The data from the all pre- and post-test results (Level 2), as well as the feedback questionnaire (Level 1) were compiled for analysis, and the percentages of correct, incorrect and IDK usage were calculated for the MCQs and the CQs for all the utility sectors.

For research questions 1–3, we define 'P' as the proportion of correct answers out of the total number of questions answered. The first subscript (Y or N) indicates whether the IDK option was available and the second subscript (1 or 2) indicates whether the assessment was pre-test or post-test assessment, respectively. We define 'Q' as the proportion of incorrect answers out of the total, using the same subscripts. In cases (such as research question 3) where only control questions (CQs) were analyzed, this is indicated by a third subscript (C). So, for example, P_{Y2C} would indicate the proportion of CQs answered correctly (of the total number of CQs) on the post-test where there was an IDK option. We define 'I' as the proportion of IDK option chosen using the same subscripts. These definitions are summarized in Table 4.

Table 4. Summary of proportions used for the analysis

QUESTION TYPE	ASSESSMENT	IDK	PROPORTION CORRECT	PROPORTION INCORRECT	PROPORTION IDK
MCQs	Pre-Test	Yes	P_{Y1}	Q_{Y1}	I_{Y1}
		No	P_{N1}	Q_{N1}	
	Post-Test	Yes	P_{Y2}	Q_{Y2}	I_{Y2}
		No	P_{N2}	Q_{N2}	
CQs Only	Pre-Test	Yes	P_{Y1C}	Q_{Y1C}	I_{Y1C}
		No	P_{N1C}	Q_{N1C}	
	Post-Test	Yes	P_{Y2C}	Q_{Y2C}	I_{Y2C}
		No	P_{N2C}	Q_{N2C}	

Statistical analysis was performed using Minitab 17 (State College, PA, USA). Two-tailed two-proportion z-tests were conducted with a level of significance (α) of 0.05 for statistical analysis of all hypothesis that are detailed for each RQ below.

RQ #1: In order to quantify the impact of IDK addition to all MCQs on the pre-test assessment, the percentage of correct and incorrect answers were compared between two training groups, one of which did not have the IDK option in the pre-tests. Statistical analysis was performed for difference in percentage of correct (H_0: $P_{Y1} - P_{N1} = 0$) and incorrect (H_0: $Q_{Y1} - Q_{N1} = 0$) answers on the pre-tests with and without the IDK option.

RQ #2: Similar to research question 1, the effectiveness of IDK addition to all MSQs on the post-test was evaluated by comparing the percentage of correct and incorrect answers in the post-training assessment of two groups, one of which didn't have the IDK option. Statistical analysis was performed for two hypotheses: (H_0: $P_{Y2} - P_{N2} = 0$) and (H_0: $Q_{Y2} - Q_{N2} = 0$).

RQ #3: To understand if the addition of the IDK option truly reduces the amount of guessing in pre- and post-training assessments, the percentage of correct, incorrect and IDK answers for the CQ in the pre- and post-training tests were compared between two groups, one of which did not have the IDK option on their tests. Statistical analysis of difference between the percentage of correct (H_0: $P_{Y1C} - P_{N1C} = 0$) and incorrect (H_0: $Q_{Y1C} - Q_{N1C} = 0$) answers on the pre-tests for the CQ with and without the IDK option was conducted. Similar analysis was performed on the posts-tests between the percentage of correct (H_0: $P_{Y2C} - P_{N2C} = 0$) and incorrect (H_0: $Q_{Y2C} - Q_{N2C} = 0$) answers. Finally, statistical significance was tested for a difference in the percentage of IDK answers between the pre-test and the post-test for the CQ with and without the IDK option (H_0: $I_{Y1C} - I_{Y2C} = 0$).

RQ #4: To determine the difference in post-test response between MCQ and CQ if IDK was chosen during the pre-test we define P as a proportion out of the total pre-test questions answered IDK. The first subscript indicates whether the post-test answer (which was IDK on the pre-test) was correct (a), incorrect (b), or IDK (c). When only control questions (CQs) were analyzed, this is indicated by a second subscript (C). So, for example, if $P_{bC} = 0.3$, this would indicate that 30% of CQs answered IDK on the pre-test were changed to an incorrect answer on the post-test. These definitions are summarized in Table 5.

Table 5. Summary of proportions used for research question 4

QUESTION TYPE	POST-TEST ANSWER	PROPORTION CHANGED FROM PRE-TEST IDK
MCQs answered IDK	Correct	P_a
on pre-test	Incorrect	P_b
	IDK	P_c
CQs answered IDK	Correct	P_{aC}
on pre-test	Incorrect	P_{bC}
	IDK	P_{cC}

Then, based on this smaller data set, we examined each participant's response on the same question in the post-test assessment, and grouped them into 3 groups: 'Pre-IDK to post-Correct', 'Pre-IDK to post-Incorrect' and 'Pre-IDK to post-IDK'. Statistical analysis was conducted to test the difference in the percentage of IDK answers on the pre-tests that changed to correct (H_0: $P_a - P_{aC} = 0$), incorrect (H_0: $P_b - P_{bC} = 0$) or IDK (H_0: $P_c - P_{cC} = 0$) answers on the post-tests for all MCQs and CQ.

3. Results

The 1474 study participants well represented general demographics of the utility workforce in the US, with a majority (over 90%) males and none of the participants had an issue with literacy. More than half (54.3%) of participants reported having no prior ergonomic training, and most (71%) worked at the same company more than five years. The detailed demographics of the participants in the various training sessions are provided in Table 6.

Table 6. Demographic information of the training participants from each utility sector

	UTILITY SECTOR						Total (n)
	Natural Gas		Electric Transmission		Power Generation		
	Tier 1	Tier 2	Tier 1	Tier 2	Tier 1	Tier 2	
Number of Participants (n)	500	375	61	359	22	157	1474
Gender							
Male	94.9%	86.8%	100%	94.9%	90%	91.6%	1365
Female	5.10%	13.2%	0%	5.1%	10%	8.4%	99
Ethnicity							
African American	1.5%	0%	3.4%	0%	0%	0%	10
American Indian	0%	0%	3.5%	0%	0%	0%	2
White, Non-Hispanic	94.8%	95.5%	91.2%	93.5%	95%	96.1%	1395
Multi-ethnic Background	0%	0%	0%	3.1%	0%	0%	11
Other	3.7%	4.5%	0%	3.4%	5%	3.9%	55
Level of education							
HS Diploma / GED	42.5%	25.9%	20%	35.9%	10%	9.9%	468
Some college	27.2%	34.2%	43.6%	32.4%	20%	8.6%	425
2-Year degree	20%	36.9%	23.6%	18.7%	25%	59.3%	419
4-Year degree	3.7%	0%	9.1%	6.9%	40%	16%	83
Higher degree	2.8%	0%	0%	3.1%	0%	3.7%	31
Other	3.9%	3%	3.6%	3.1%	5%	2.5%	49
Prior Ergo Training							
No	58.2%	52.1%	52.7%	53.5%	55%	54.3%	808
Yes	41.8%	47.9%	47.3%	44.1%	45%	40.7%	650
Years with Company							
<1	3.4%	8.6%	10.5%	19.9%	0%	7.5%	139
1–5	19.6%	25.4%	0%	27%	20%	10%	310
6–10	13.1%	19.2%	15.8%	23.2%	10%	35%	288
11–15	12.3%	11.5%	10.5%	8.2%	15%	13.8%	165
16–20	11.2%	6.2%	12.3%	17.2%	15%	8.8%	166
20+	38.7%	28.9%	38.6%	0%	40%	22.5%	370

To understand the trends in answering the MCQs in the pre- and post-test assessments, Table 7 details the percentage and counts of the answers that had been answered correctly, incorrectly, and IDK (when applicable) and these percentages have been aligned with the previously defined variables

Table 7. Percentage of correct, incorrect and IDK answers in pre-test assessment

QUESTION TYPE	ASSESSMENT	IDK	PROPORTION CORRECT*	PROPORTION INCORRECT*	PROPORTION IDK*
MCQs	Pre-Test	Yes	P_{Y1} = 66% (n = 1661)	Q_{Y1} = 30% (n = 765)	I_{Y1} = 3% (n = 87)
		No	P_{N1} = 42% (n = 2111)	Q_{N1} = 58% (n = 2929)	
	Post-Test	Yes	P_{Y2} = 83% (n = 2094)	Q_{Y2} = 16% (n = 402)	I_{Y2} = 1% (n = 17)
		No	P_{N2} = 80% (n = 4031)	Q_{N2} = 20% (n = 1009)	
CQs Only	Pre-Test	Yes	P_{Y1C} = 14% (n = 68)	Q_{Y1C} = 24% (n = 116)	I_{Y1C} = 62% (n = 297)
		No	P_{N1C} = 12% (n = 103)	Q_{N1C} = 88% (n = 727)	
	Post-Test	Yes	P_{Y2C} = 40% (n = 190)	Q_{Y2C} = 27% (n = 128)	I_{Y2C} = 34% (n = 163)
		No	P_{N2C} = 24% (n = 203)	Q_{N2C} = 76% (n = 627)	

Note. *Where 'n' is the number of questions.

The results for RQ #1 indicate that there was a statistically significant difference ($z = 20.65$; $p < 0.05$) of 24% between the percentage of correct pre-test MCQ answers with ($P_{Y1} = 66\%$) and without ($P_{N1} = 42\%$) the IDK option. In addition, there was on average a 28% statistically significant difference ($z = -24.04$; $p < 0.05$) observed in the percentage of incorrect pre-test MCQ answers with ($Q_{Y1} = 30\%$) and without ($Q_{N1} = 58\%$) the IDK option. Figure 2 illustrates the trends in the percentage changes of correct, incorrect, and IDK answers in the pre-test assessment for the MCQ with the addition of the IDK option.
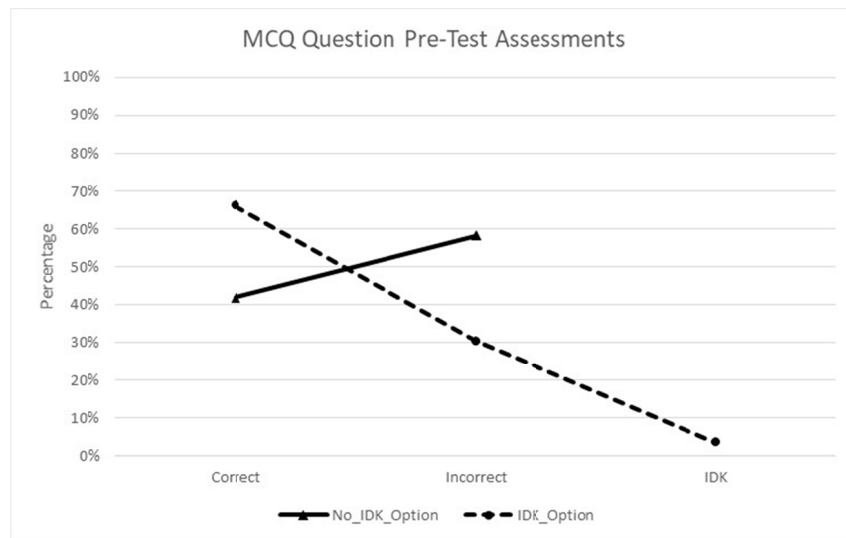

Figure 2. Percentage of questions that were answered Correct, Incorrect and IDK in the pre-test assessment for MCQs

While the difference between two groups of trainees (with or without IDK option) were similar, the results for RQ #2 indicate that, there was a 3% statistically significant difference ($z = 3.59$; $p < 0.05$) in correct post-test MCQ answers with ($P_{Y2} = 83\%$) and without ($P_{N2} = 80\%$) the IDK option. Furthermore, a 4% difference ($z = -4.36$; $p < 0.05$) was observed in the percentage of incorrect post-test MCQ answers with ($Q_{Y2} = 16\%$) and without ($Q_{N2} = 20\%$) the IDK option. The trends in in the percentage changes of correct, incorrect, and IDK answers in the post-test assessment for the MCQ with the addition of the IDK option are illustrated in Figure 3.
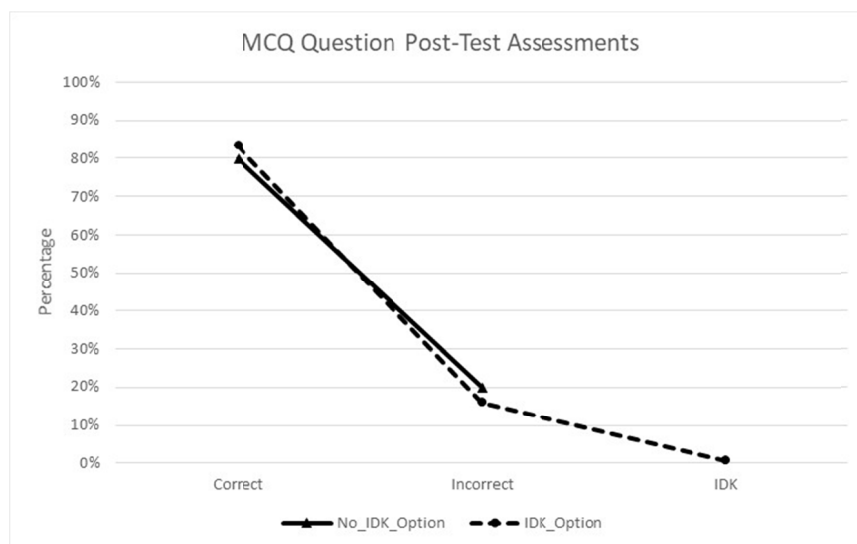

Figure 3. Percentage of questions that were answered Correct, Incorrect and IDK in the post-test assessment for MCQs

The pre-test assessment results for RQ #3 revealed no statistically significant difference ($z = 0.88$; $p > 0.05$) in the percentage of correct pre-test CQ answers with ($P_{Y1C} = 14\%$) and without ($P_{N1C} = 12\%$) the IDK option. Nevertheless, a 63.4% difference ($z = -28.07$; $p < 0.05$) was detected in the percentage of incorrect pre-test CQ answers with ($Q_{Y1C} = 24\%$) and without ($Q_{N1C} = 88\%$) the IDK option. The trends in percentages of correct, incorrect, and IDK answers in the pre-test assessment for the CQ with the addition of the IDK options are illustrated in Figure 4.
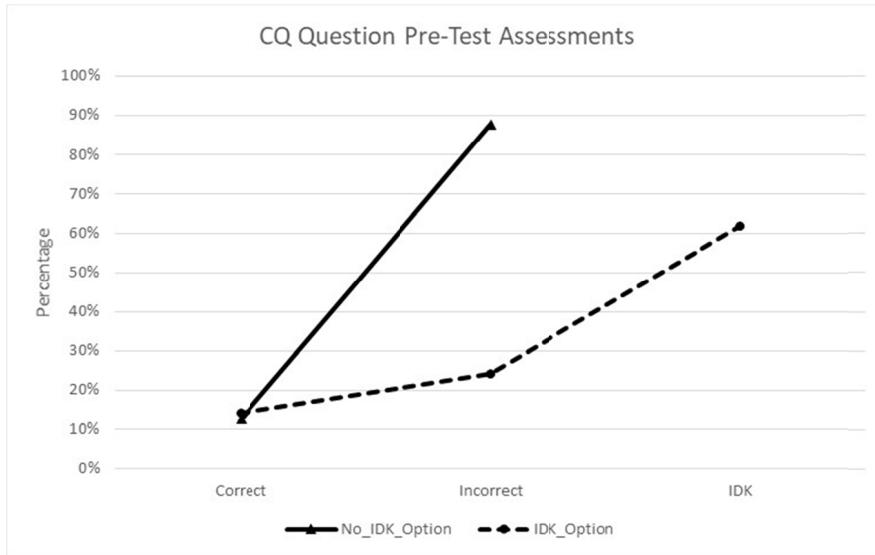

Figure 4. Percentage of Correct, Incorrect and IDK answers for the control question for pre-test assessments

In the post-test assessments there was a statistically significant difference ($z = 5.61$; $p < 0.05$) of 16% in the percentage of correct post-test CQ answers observed with ($P_{Y2C} = 40\%$) and without ($P_{N2C} = 24\%$) the IDK option. In addition, there was a 49% difference ($z = -19.52$; $p < 0.05$) observed in the percentage of incorrect post-test CQ answers with ($Q_{Y2C} = 27\%$) and without ($Q_{N2C} = 76\%$) the IDK option. The trends in the percentage changes of correct, incorrect, and IDK answers in the post-test assessment for the CQ with the addition of the IDK options are presented in Figure 5.
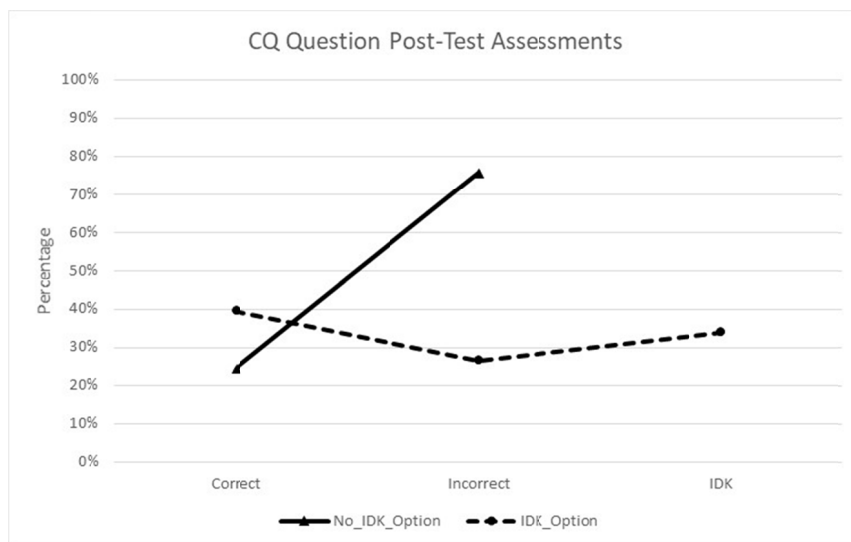

Figure 5. Percentage of Correct, Incorrect and IDK answers for the control question in the various training groups for post-test assessments

Comparing the selection of the IDK option in the CQ between the pre- and post-test assessment results indicated that there was a statistically significant difference ($z = 9.01$; $p < 0.05$) of 28% in the percentage of the IDK answers in CQ between the pre-test ($I_{Y1C} = 62\%$) post-test ($I_{Y2C} = 34\%$) assessments.

In summary, we observe that:

- The addition of the IDK option decreases the percentage of incorrect answers in the pre-test assessment for both the MCQ and the CQ.

- There is a statistically significant reduction in the usage of the IDK option in the post-test assessment for both MCQ and CQ. This is expected for MCQ as the contents were taught in the training session. This is not expected in the CQ and the content was not taught to the participants.

Further analysis of the post-test assessments with respect to RQ #4 revealed some interesting insights that are helpful for understanding the trainees in post-training reaction to IDK option on the test. Table 8 summarizes the percentage and counts of the questions that were answered as IDK in the pre-test assessment and then changed to either correct, incorrect or IDK in the post-test assessment.

Table 8. Change of state for questions answered as IDK in the pre-test assessment

QUESTION TYPE	POST-TEST ANSWER	PROPORTION CHANGED FROM PRE-TEST IDK*
MCQs answered IDK on pre-test	Correct	$P_a = 60\%$ ($n = 52$)
	Incorrect	$P_b = 28\%$ ($n = 24$)
	IDK	$P_c = 13\%$ ($n = 11$)
CQs answered IDK on pre-test	Correct	$P_{aC} = 31\%$ ($n = 91$)
	Incorrect	$P_{bC} = 21\%$ ($n = 61$)
	IDK	$P_{cC} = 49\%$ ($n = 145$)

Note. *Where 'n' is the number of participants.

A statistically significant difference ($z = 4.94$; $p < 0.05$) of 29% in the percentage of answers that changed from IDK in the pre-test assessment to correct in the post-test assessment for MCQ ($P_a = 60\%$) and CQ ($P_{aC} = 31\%$) was observed. However, there was no statistically significant difference ($z = 1.32$; $p > 0.05$) in the percentage of answers that changed from IDK in the pre-test assessment to incorrect in the post-test assessment for MCQ ($P_b = 28\%$) and CQ ($P_{bC} = 21\%$). Finally, a 36% difference ($z = -7.87$; $p < 0.05$) was observed in the percentage of answers that did not change from IDK in the pre-test and post-test assessments for MCQ ($P_c = 13\%$) and CQ ($P_{cC} = 49\%$).

In summary we observe that:

- For MCQs and CQ, 61% and 30% of the participants respectively, changed from IDK in the pre-test assessment to the correct answer in the post-test assessment. This is expected in the case of the MCQ but not expected in the case of the CQ. Thus, it illustrates that some of the participants are able to guess the right answer instead of answering IDK in the post-test assessment.

- For MCQs and CQ, 28% and 21% of the participants respectively, changed from IDK in the pre-test assessment to the incorrect answer in the post-test assessment. This implies that about the same percentage of individuals are not attentive in the training and answer the questions incorrectly in the post-test assessments or choose not to use the IDK option.

- For MCQs and CQ, 13% and 49% of the participants respectively, did not change their IDK choice in the pre-test and the post-test assessment. This implies that for MCQs a small percentage of participants did not learn the concepts taught and were honest in answering IDK in the post-test assessment. For the CQ, a large portion of the participants were honest in answering IDK in the post-test assessment.

- It is of note that in the CQ, 51% of the participants still chose to change their answer from IDK in the pre-test to either correct or incorrect in the post-test even though the concept was not taught. i.e., 51% of the participants would rather guess at an answer in the post-test assessment rather than answer IDK even though they answered as IDK in the pre-test assessment.

4. Discussion

The analysis conducted illustrates some interesting behavioral trends observed in participants with respect to guessing on MCQ pre- and post-training assessments. Several prior studies demonstrated that the concept of adding IDK to only a True/False assessment model helped to minimize guessing on the post-tests (Sanderson,

1973; Newble et. al., 1979; Courtenay & Weidemann, 1985; Hammond, McIndoe, Sansome, & Spargo, 1998; van Mameren & van der Vleuten, 1999; Spears & Wilson, 2010). As mentioned before, the major issue with the previous studies is that their methodology does not allow for true assessment of the training effectiveness. Additionally, since the baseline knowledge was not assessed prior to the training, and control questions were not utilized, it was impossible to separate true learning from guessing on the same group of participants.

The current study design allows these gaps to be filled-in through investigating four main research questions. The first two were specifically addressing the "benefits" of adding an IDK option in pre- and post-test assessments respectively. Based on the results of this study a significant decrease in the percentage of incorrect answers (27%) with the addition of the IDK option to pre-tests is observed. This can simply be explained by a behavioral change, since there is no expectation for a participant to know the correct answer, therefore IDK becomes the best option for the questions about which they have no prior knowledge. In the post-test assessment for MCQs we see a much smaller, although statistically significant, difference (approx. 3%–4%) in the percentage of correct and incorrect answers with the addition of the IDK option.

While it is expected that the proportion of IDK answers on the post-training assessment will be reduced due to gained knowledge, the participants who did not get a perfect score on the post-test did not chose the IDK option instead of guessing. This became further evident while analyzing the response to the CQ and comparing the difference between pre-test and post-test assessments. For an MCQ we expected most of the participants to answer correctly in the post-test assessment if they answered IDK in the pre-test assessment. Nevertheless, for a CQ, we expected most of the participants to answer IDK in the post-test assessment if they answered IDK in the pre-test assessment.

In the pre-test assessment for the CQ, with the addition of the IDK option, we observe no statistically significant difference in the percentage of correct answers but observe a significant decrease (63.4%) in the percentage of incorrect answers. This implies that participants, in the pre-test assessment, are very open to answering IDK to a question to which they do not know the answer. In the post-test assessment we observe a 15.1% increase in the percentage of correct answers and a 50% reduction in the percentage of incorrect answers. Additionally, we observe a 28% reduction in the usage of the IDK option between the pre-test and post-test assessments in the case of the CQ. From years of conducting training for adults in various utility industries, this is completely expected as it would indicate that the participants learned the concepts taught and were able to correctly answer the MCQs in the post-test assessment. However, a concerning observation is that we see a significant reduction in the percentage usage of the IDK option from the pre-test to post-test assessment for the CQ as well. Since this question was not taught during any of the training sessions, it helps expose participant guessing behaviors while answering MCQs.

To quantify how participants who answered IDK in the pre-test assessment for MCQs and CQ changed their answers in the post-test assessment, thus answering research question 4, we observe that the MCQs have a 29% higher conversion from IDK to a correct answer than the CQ. There was no difference in the percentage of conversion from IDK to incorrect answers and participants are 36% more likely to answer IDK again in the post-test analysis in the case of a CQ. This implies that most of the participants are learning the concepts taught if they come into the training session not knowing the concept. The troubling finding is that 51% of the participants who answered IDK to the CQ in the pre-test assessment changed their answer and were willing to guess on the post-test assessment.

The findings with regards to the CQ are at odds with what one would typically expect in a training environment. Since the concept in the CQ is not taught in the class, we would expect a similar percentage of IDK option usage in both the pre-test assessment and the post-test assessment. To get a better understanding of what is occurring in the CQ, the comparisons made between the assessments with and without the IDK option is very telling on participant behavior. In the pre-test assessment, for the CQ, we see that the addition of the IDK option does not impact the percentage of correct answers but helps significantly reduce the percentage of incorrect answers. So, although there is some guessing, it gives an opportunity for the participant to truly express their knowledge level. In the post-test assessment, for the CQ, addition of the IDK option does not have the same impact. There is a significant reduction in the usage of the IDK option, even though the CQ tests a concept that is not taught in the training session. This implies that participants would rather guess at an answer in the post-test assessment than answer IDK, even if they did not know the correct answer. This behavior has been observed and reported among adults and children (Waterman, Blades, & Spenser, 2004, Howie & O'Neill, 1996) and was discussed as a significant impactor of business decisions and reported in a Freakonomics radio podcast (Lechberg, 2014).

The more important interpretation of the overall results is that the addition of the IDK option does not

significantly reduce the amount of guessing in the post-test assessment and is at odds with the findings from the various authors detailed in the literature review (Sanderson, 1973; Newble et al., 1979; Courtenay & Weidemann, 1985; Hammond et al., 1998; van Mameren & van der Vleuten, 1999), who have stated that incorporation of the IDK option minimizes guessing and can be used as an alternate method to formula scoring. The IDK option, however, is quite effective at helping understand the incoming knowledge level of the participants when administered in the pre-test assessment and can be viewed as a powerful tool to help the instructors modify course content and delivery methods to suit the individual class group needs.

One of the limitations of this study is that the results of different groups (with IDK option and without IDK option) are compared. The commonality is that the training content is related to safety in their utility industry and that the CQ in all cases was not taught during the training session. Also, in the current study it was not possible to conduct Computer Based Testing (CBT) for the participants as the training was conducted at various site locations with some level of computer illiteracy, as well as due to the time constraints available to conduct the training which made setting up computers for each training session not a viable option. Finally, in this study it was not possible to use a formula scoring model to minimize guessing mainly due to the confusing nature of the Formula scoring models and the associated risk of confusing the participants. The time constraints in the training sessions was rather short, and it was not possible to clearly explain the Formula Scoring method to the participants in the assessment.

5. Conclusion/Future Direction

This research study investigated and quantified the impact of the IDK option on learning outcomes through MCQ pre- and post-training assessments. A concept called the 'Control Question (CQ)' was introduced in both the pre- and the post-test assessments and is akin to the administration of a placebo treatment since the concept tested by the CQ was not covered in the training sessions. The trends in answers seen in the CQ were compared to those seen in the other MCQs that were taught in the training sessions.

The introduction of the IDK option in the pre-test assessment was observed to statistically significantly reduce incorrect answers by 63% and can be used to help trainers cater the content and delivery to focus on the concepts in which the participants have the largest gaps of knowledge. Nevertheless, the IDK option was not observed to significantly reduce the amount of guessing in the post-test assessment as shown by the change in states measured in the CQ.

Some recommendations that can be derived from this study are:

- Both pre-test assessment before the training and post-test assessment after the training should be administered in order to allow for better assessment of training effectiveness.

- Utilizing MCQs instead of T/F questions decreases the probability of getting a correct answer due to guessing on both pre- and post-test assessments and therefore improves true estimate of learning.

- Conducting the pre-test assessment with the participants prior to the training session and allowing some time to analyze the results before the training may be helpful for the trainers to assess the specific topics that should given greater emphasis during the training.

- Having a dialog on the knowledge gaps to help the training session be more interactive and pertinent to each class will ensure that the trainees get the most out of the training session.

- Being aware that adding an IDK option to the pre-tests was shown to significantly reduce guessing, while on the post-tests the effect was not as pronounced.

- Using a control (placebo) question(s) on pre- and post-tests can be helpful with generating estimates of the probability of guessing and allow better estimates of true learning.

Acknowledgments

References

Alliger, G. M., & Janak, E. A. (1989). Kirkpatrick's levels of training criteria: Thirty years later. *Personnel Psychology*, *42*(2), 331–342. https://doi.org/10.1111/j.1744-6570.1989.tb00661.x

Arciniegas Calle, M. C., Lobelo, F., Jimenez, M. A., Paez, D. C., Cores, S., de Lima, A., & Duperly, J. (2016).

One-day workshop-based training improves physical activity prescription knowledge in Latin American physicians: a pre-test posttest study. *BMC Public Health*, *16*, 1224–1235. https://doi.org/10.1186/s12889-016-3883-2

Arthur Jr, W., Bennett Jr, W., Edens, P. S., & Bell, S. T. (2003). Effectiveness of training in organizations: a meta-analysis of design and evaluation features. *Journal of Applied Psychology*, *88*(2), 234. https://doi.org/10.1037/0021-9010.88.2.234

Bahn, S., & Barratt-Pugh, L. (2012). Emerging issues of health and safety training delivery in Australia: Quality and transferability. *Procedia – Social and Behavioral Sciences*, *62*, 213–222. https://doi.org/10.1016/j.sbspro.2012.09.035

Bersin, J. (2014, Feb. 4). Spending on corporate training soars: Employee capabilities now a priority. *Forbes*. Retrieved from https://www.forbes.com/sites/joshbersin/2014/02/04/the-recovery-arrives-corporate-training-spend-skyrockets/#3e38e97dc5a7

Brogan, D. R., & Kutner, M. H. (1980). Comparative analyses of pretest-posttest research designs. *The American Statistician*, *34*(4), 229–232. https://doi.org/10.1080/00031305.1980.10483034

Burke, L. A., & Baldwin, T. T. (1999). Workforce training transfer: a study of the effect of relapse prevention training and transfer climate. *Human Resource Management*, *38*(3), 227–242. https://doi.org/10.1002/(SICI)1099-050X(199923)38:3<227::AID-HRM5>3.0.CO;2-M

Campbell, D. T., Stanley, J. C., & Gage, N. L. (1963). *Experimental and quasi-experimental designs for research* (No. 04; Q175, C3.). Boston: Houghton Mifflin.

Campbell-Kyureghyan, N., Ahmed, M., & Beschorner, K. (2013). *Measuring Training Impact 1–5*. Paper presented at the US DOL Trainer Exchange Meeting, Washington DC, March 12–13, 2013.

Campbell-Kyureghyan, N., Principe, A. H., & Ahmed, M. (2013). *Effectiveness of first and second tier safety and ergonomics training in power utilities*. Proceedings of the XXVth Annual Occupational Ergonomics and Safety Conference, Atlanta, GA, USA, June 6–7, 2013.

Courtenay, B. C., & Weidemann, C. (1985). The effects of a "don't know" response on Palmore's facts on aging Quizzes. *The Gerontologist*, *25*(2), 177–181. https://doi.org/10.1093/geront/25.2.177

Dimitrov, D. M., & Rumrill, P. D. (2003). *Pretest-posttest designs and measurement of change*. Work-Andover Medical Publishers Incorporated. IOS PRESS.

Frary, R. B. (1988). Formula scoring of multiple-choice tests (correction for guessing). *Educational Measurement: Issues and Practice*, *7*(2), 33–38. https://doi.org/10.1111/j.1745-3992.1988.tb00434.x

Hammond, E. J., McIndoe, A. K., Sansome, A. J., & Spargo, P. M. (1998). Multiple‐choice examinations: adopting an evidence‐based approach to exam technique. *Anaesthesia*, *53*(11), 1105–1108. https://doi.org/10.1046/j.1365-2044.1998.00583.x

Hendrix, L. J., Carter, M. W., & Hintze, J. L. (1978). A comparison of five statistical methods for analyzing pretest-posttest designs. *The Journal of Experimental Education*, *47*(2), 96–102. https://doi.org/10.1080/00220973.1978.11011664

Herbig, M. (1976). Item analysis by use in pre‐tests and post‐tests: A comparison of different coefficients. *Programmed Learning and Educational Technology*, *13*(2), 49–54. https://doi.org/10.1080/1355800760130206

Howie, P., & O'Neill, A. (1996). *Monitoring and reporting lack of knowledge: Developmental changes in the ability to say "I don't know" when appropriate*. Paper presented at the 31st Annual Conference of the Australian Psychological Society, Sydney, Australia.

Kirkpatrick, D. L. (1967). Evaluation of training. In R. L. Craig & L. R. Bittel (Eds.), *Training and Development Handbook* (pp. 40–60). New York: McGraw Hill.

Kontoghiorghes, C. (2001). Factors affecting training effectiveness in the context of the introduction of new technology—a US case study. *International Journal of Training and Development*, *5*(4), 248–260. https://doi.org/10.1111/1468-2419.00137

Lechberg, S. (2014). *The three hardest words in the English language: a new freakonomics radio podcast*. Retrieved June 10, 2017, from

http://freakonomics.com/podcast/the-three-hardest-words-in-the-english-language-a-new-freakonomics-radio-podcast/

Newble, D. I., Baxter, A., & Elmslie, R. G. (1979). A comparison of multiple‑choice tests and free‑response tests in examinations of clinical competence. *Medical Education*, *13*(4), 263–268. https://doi.org/10.1111/j.1365-2923.1979.tb01511.x

Ningdyah, A. E., Greenwood, K. M., & Kidd, G. (2018). A training-model scale's validity and reliability coefficients: expert evaluation in Indonesian professional psychology programs. *Makara Human Behavior Studies in Asia*, *22*(1), 56–66. https://doi.org/10.7454/hubs.asia.2190318

Sanderson, P. H. (1973). The 'don't know' option in MCQ examinations. *Medical Education*, *7*(1), 25–29. https://doi.org/10.1111/j.1365-2923.1973.tb02206.x

Spears, K., & Wilson, M. (2010). *"I don't know" and Multiple Choice Analysis of Pre- and Post-Tests.* Retrieved August 6, 2015, from https://www.joe.org/joe/2010december/tt2.php

van der Linden, W. J. (1981). A latent trait looks at pretest-posttest validation of criterion-referenced test items. *Review of Educational Research*, *51*(3), 379–402. https://doi.org/10.3102/00346543051003379

van Mameren, H., & van der Vleuten, C. P. M. (1999). The effect of a 'don't know' option on test scores: number‑right and formula scoring compared. *Medical Education*, *33*(4), 267–275. https://doi.org/10.1046/j.1365-2923.1999.00292.x

Warr, P., Allan, C., & Birdi, K. (1999). Predicting three levels of training outcome. *Journal of Occupational and Organizational Psychology*, *72*(3), 351–375. https://doi.org/10.1348/096317999166725

Waterman, A. H., Blades, M., & Spencer, C., (2004). Indicating when you do not know the answer: The effect of question format and interviewer knowledge on children's 'don't know' responses. *British Journal of Developmental Psychology*, *22*, 335–348. https://doi.org/10.1348/0261510041552710

Copyrights