Chapter 23

EXPERIMENT-BASED EXAMS AND THE DIFFERENCE BETWEEN THE BEHAVIORAL AND THE NATURAL SCIENCES

Ido Erev and Re'ut Livne-Tarandach
Faculty of Industrial Engineering and Management
Technion, Haifa, Israel

One of the clearest indications of the gap between the behavioral and the natural sciences is provided by the exams used to evaluate students. Typical questions in natural science exams ask examinees to predict the outcome of particular experiments. For example, a question might show a simple circuit and ask, "If the switch is closed at time zero, which of the following curves shows the current through the resistor as a function of time?" Typical questions in the behavioral sciences, on the other hand, ask examinees to state the meanings of particular terms, or to associate them with particular theories.

The current paper proposes and evaluates a method that can facilitate the development of natural science-like prediction questions for the behavioral sciences. We believe that such a change can benefit behavioral scientists in two ways. First, our experience teaching engineering and business students with good backgrounds in the natural sciences suggests that these students tend to find concept-focused multiple-choice exams inappropriate to evaluate their achievements. These students do not understand why they must memorize the structure of abstract theories from which no clear predictions may be drawn. Including in exams questions that focus on predictions would increase the face validity of these exams, improve students' attitudes toward the exams and the course material on which they are based, and so make the students more effective learners. Second, it is also possible that a new focus on predictions in exams, courses, and textbooks will affect mainstream research and will help close the gap between the behavioral and the natural sciences.

The paper is organized as follows. Section 1 evaluates and quantifies the assertion that there is a qualitative difference between exam questions in the natural and behavioral sciences. In this section, we compare the GRE subject exams in Physics and Psychology (exams used by leading universities to evaluate candidates for graduate school; see http://www.gre.org/edindex.html). The results show substantial differences between the two. Whereas nearly all the questions in the Physics exam focus on concrete situations, in the Psychology exam nearly all questions focus on abstract terms.

426

We believe that the focus of behavioral science exams on abstract terms and theories is driven by the fact that an understanding of leading behavioral theories does not ensure accurate predictions of behavior (see related discussion in Heiner, 1983). Thus, exam developers cannot know with certainty the correct answer to a question that requires prediction. At best, they can know what the theory predicts. Section 2 presents one solution to this problem, based on the fact that it is possible to ask questions about specific experiments (laboratory or natural) that have been run. For example, a question might describe a specific experiment and ask the examinee to choose among several possible results. Notice that this focus on concrete situations does not mean that theories are not important. Understanding useful models of human behavior should help students remember the results of important experiments studied in class, and predict behavior in experiments that were not included in the class material.

Section 3 summarizes a pilot (case) study that evaluates the validity (discriminative power) of experiment-based questions. The results show almost no difference between experiment-based and concept-based questions in their ability to discriminate between strong and weak examinees.

Section 4 highlights four problematic properties of experiment-based questions. Here, we suggest that modifying the information students receive (and researchers collect) concerning the value of descriptive models can reduce the negative effect of these properties, and increase the discriminative power of experiment-based questions. Two procedures that can facilitate this goal are discussed.

1. COMPARISON OF GRE SUBJECT EXAMS

In order to evaluate the difference between typical exams in the natural and behavioral sciences, the current section compares the GRE subject exams in Psychology and Physics. We chose to focus on GRE exams because these exams have been carefully developed to evaluate the knowledge taught in undergraduate programs, and they are used to evaluate candidates for top graduate schools.

To help students prepare for the GRE exams (which are developed by the Educational Testing Service, or ETS), the GRE web site (http://www.gre.org/edindex.html) presents a practice book in each subject. The practice books in Psychology and Physics include 214 and 99 questions, respectively. In our analysis, we have divided these questions into three categories: "abstract," "experiment-based," and "mixed." A question was classified as "abstract" if the correct answer is a property of a theory. A question was classified as "experiment-based" if the correct answer is the result (or the likely result) of a particular experiment. A question was classified as "mixed" if it explicitly asks about the relationship between a particular theory and a particular experimental result. Table 1 presents examples of the three types of questions from the GRE practice books in Psychology and Physics.

The right-hand column of Table 1 presents the distribution of questions over the three categories in the two exams. The results reveal a large difference between the two exams. The experiment-based category accounts for 62% of the questions in
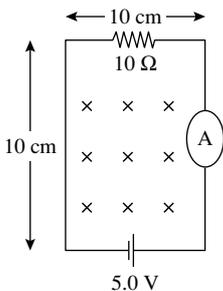
Table 1. Examples of GRE questions in Psychology and Physics that were classified to the different categories. The right hand column presents the proportion of questions in each category.

	Question type	*Example*	*Proportion in GRE test*
Psychology			
	Abstract	Question 1: According to Piaget, the major cognitive attainment of the sensorimotor period is (A) speech perception (B) shape constancy (C) mental representation (D) nonegocentric thought (E) recognition memory	**0.84**
	Experimental based	Question 2: Subjects are presented with a randomly arranged list of animals, fruits, and tools, and then asked to recall the list in any order they wish. Their recall protocols are most likely to show which of the following? (A) The items with the same initial letters occur close together. (B) The items that rhyme occur close together. (C) The items that belong to the same conceptual category occur close together. (D) The items occur in an order highly similar to that used for presentation. (E) The items from only one of the conceptual categories are recalled.	**0.1**
	Mixed	Question 3: Brown and McNeill (1966) read the definitions of uncommon words to subjects and then asked them to supply those words. When asked questions about any words they thought they knew but could not recall, subjects often responded with words that were phonologically similar to the target words. The phenomenon investigated in this experiment is called (A) eidetic imagery (B) proactive inhibition (C) the complexity-of-expression phenomenon (D) the tip-of-the-tongue phenomenon (E) the template model	**0.06**

Table 1. (cont'd)

	Question type	Example	Proportion in GRE test
Physics	Abstract	Question 4: According to the Standard Model of elementary particles, which of the following is NOT a composite object? (A) Muon (B) Pi-meson (C) Neutron (D) Deuteron (E) Alpha particle	**0.09**
	Experimental based	Question 5: 2. The circuit shown above is in a uniform magnetic field that is into the page and is decreasing in magnitude at the rate of 150 tesla/second. The ammeter reads (A) 0.15A (B) 0.35A (C) 0.50A (D) 0.65A (E) 0.80A	**0.62**
	Mixed	Question 6: When the beta-decay of ^{60}Co nuclei is observed at low temperatures in a magnetic field that aligns the spins of the nuclei, it is found that the electrons are emitted preferentially in a direction opposite to the ^{60}Co spin direction. Which of the following invariances is violated by this decay? (A) Gauge invariance (B) Time invariance (C) Translation invariance (D) Reflection invariance (E) Rotation invariance	**0.28**

Physics, and only 10% in Psychology. The abstract category accounts for 84% of the questions in Psychology, and only 9% in Physics.

2. EXPERIMENT-BASED QUESTIONS

Examination of the (few) experiment-based questions in the Psychology practice book reveals that these questions tend to focus on a small set of robust experimental results. These deal with taste aversion, extinction, serial effects in memory, physiological mechanisms of the senses, and perceptual biases. Only one of the 21 questions in social and/or organizational psychology focuses on a specific experiment. This question focuses on the mirror effect.

We believe that the relatively small number of experiment-based questions in Psychology is an indication of the fact that the number of robust experimental phenomena is small. Thus, exam writers seeking to develop experiment-based questions in psychology must struggle with two related problems. First, in many cases the correct answer (the likely experimental outcome) is unknown. Second, even when the exam developers know the likely outcome, the examinee may not be able to derive it; even the best students are expected to err in some cases. This fact is expected to reduce the discrimination power of experiment-based questions.

The method proposed here to facilitate the development of experiment-based questions in Psychology is based on a technical solution of the first problem. We propose to focus questions on specific experiments that have been run. With this focus, the identification of the "correct answer" is easy: it is the obtained (robust) experimental result.

To clarify this technical solution it is constructive to present an example. Question 3 in Table 1 focuses on the relationship of an abstract term to an experimental result (it is a "mixed" question). The following question is an experiment-based modification of this question:

Question 7
Brown and McNeill (1966) read the definitions of uncommon words to subjects and then asked them to supply those words. When asked for words they thought they knew but could not recall, subjects often responded with words that:

    A. Were phonological similar to the target word.
    B. Were semantically similar to the target word.
    C. Had the exact opposite meaning of the target word.
    D. Were names of animals or natural phenomena.

In the original question (Question 3 in Table 1) the correct answer is an abstract term. In the modified question the correct answer is the experimental result (item A) that is summarized by this term.

It is also easy to develop experiment-based questions that involve quantitative predictions. Here is one example:

Question 8

In an experiment conducted by Gneezy, Haruvy and Yafe (2003), groups of six participants were invited to eat lunch in an inexpensive Haifa restaurant. Three conditions were compared. In all three conditions the participants received 70 Israeli shekels participation fee, and were asked to personally state their order on a form. In Condition Individual, each participant paid her bill. In Condition Split, the bill was split between the six group members. In Condition 1/6, each participant paid only 1/6 of her own order. The average order was 37 shekels in Condition Individual, and 57 shekels in Condition 1/6. What was the average order in Condition Split?

A. 37   B. 43   C. 51   D. 57

Notice that in the last question, the derivation of the correct answer (C) requires a good understanding of more than one principle. Good students should know that: People tend to free-ride (so the average order is likely to be larger than 37 shekels), but to a lesser extent than the rational model would predict (that would imply an average order of around 57 shekels). A student who understands the importance of individual differences and the regression effect should conclude that 51 is a more reasonable answer than 43.

The focus on experiments that have been run does not solve the second problem listed above. That is, it is not clear that the information taught in psychology courses is sufficient to ensure that good students will be able to derive the correct answers to experiment-based questions. The next section evaluates the magnitude of this problem.

3. A PILOT (CASE) STUDY

In an attempt to explore the discriminative power of experiment-based questions in Psychology exams, we incorporated questions of this type in five multiple-choice exams in courses that we have taught at the Technion. Two of the exams determined the grades in the course "Introduction to Experimental Psychology," a core course in the undergraduate program in Industrial Engineering that was taught in the fall semester of 2002. The first exam focused on cognitive psychology, the second on social and organizational psychology. Two additional exams focused on the same course and material in the 2003 spring semester. The final exam was the mid term in the elective "Thinking and Decision Making," which we taught in the fall of 2002.

The students in these courses were informed that their grades would be determined based on a new (and experimental) exam format. They were told that most questions in the exams would ask them to predict (or guess) the results of specific experiments. The students were also informed that some of these experiments would be covered in the course material, but that others would not. It was emphasized that in all experiment-based questions, the correct answer is the obtained experimental

result; thus, in the case of experiments not covered in the course, it would be impossible to know the correct answer with certainty. Nevertheless, a good understanding of the course material should ensure good educated guesses and high final grades.

In order to evaluate the discriminative power of the different question formats, we first classified each question in one of three categories: "Abstract or mixed" (as defined in Section 2); "Covered experiment-based" (questions dealing with experiments that were covered in the course); and "New experiment-based" (questions dealing with experiments not covered in the course). A discrimination score was computed for each question, based on the difference between the mean Grade Point Average (GPA) of those students who answered the question correctly and the mean GPA of those who did not. The courses we consider are taken in the fifth or later semester, whereas most courses taken in earlier semesters (which, therefore, determined the relevant GPA) are in math, the natural sciences, and engineering. The standard deviation of the GPA scores over the three courses was 4.82.

Table 2 presents the mean discrimination score in each category for each of the five exams. The results reveal a surprisingly small and inconsistent difference between the three categories. Over the five exams, the differences between the three types of questions are insignificant.

In addition to this analysis, we examined the discriminative power of questions asked in an Introduction to Psychology exam given by a different teacher during the spring semester of 2001. This course was taught in a traditional way, with an emphasis on psychological theories rather than experimental findings. We analyzed

Table 2. Discrimination scores of the different type of questions in the five exams.

Question type	Discrimination scores (standard deviation), and number of questions				
	Cognitive psychology		Social psychology		Decision making
	Fall 2002	Spring 2003	Fall 2002	Spring 2003	Fall 2002
Old experimental based (material covered in the course)	**0.8909** (1.6) N = 29	**1.98** (3.01) N = 18	**.929** (1.27) N = 18	**0.991** (2.21) N = 22	**2.46** (2.717) N = 12
New experimental based (not covered in the course)	**0.185** (1.29) N = 6	**1.86** (2.15) N = 6	**1.456** (2.37) N = 16	**0.335** (2.8) N = 10	**1.52** (3.74) N = 11
Abstract	**1.05** (0.39) N = 3	**0.988** (2.46) N = 10	**.202** (0.706) N = 5	**0.748** (1.39) N = 7	**1.55** (3.68) N = 10

the discriminative power of all multiple-choice questions (17 in number) used in the final exam for this course. According to the classification used above, all the questions in this test were abstract ones. The discrimination analysis shows an average discrimination score of 1.12, and a standard deviation of 1.39. These findings are similar to the average discrimination scores of the abstract questions used in our exam (see Table 2 bottom row).

Another interesting statistic involves the class evaluation (the students' assessment of the quality of the course). In the first semester in which the new method was used, the class evaluations dropped slightly relative to previous semesters. However, in the second semester, the class evaluations rose above the mean of the evaluations in classes using traditional methods.

4. POTENTIAL IMPROVEMENTS

The insignificant difference in discriminative power between our experiment-based and abstract questions is most naturally explained by means of opposing effects that cancel each other out. It seems that some of the unique properties of experiment-based questions have a positive effect on discrimination, while other properties have a negative effect. If this is indeed so, then modifying the problematic properties of experiment-based questions should increase their value. The present section offers a first step in this direction. Here, we identify four problematic properties of experiment-based questions, and discuss ways to reduce the negative effect of these properties.

Biased samples.
The most important problem involves the criteria for publishing experimental results. One of the first things editors look for is "surprising findings" – that is, findings that violate popular models of the type presented in textbooks. As a result, the papers published in top psychology journals – the papers we used to develop the new experiment-based exam questions – represent a biased sample of experiments. In some cases, a good understanding of the textbook models simply would not help students predict the results of these experiments.

Vague boundaries.
A second problem is that most descriptive models focus on relatively narrow sets of empirical results and/or stylized facts. In most cases, researchers (and/or textbooks) pay relatively little attention to defining the set of situations that can be addressed by the proposed model.

Introspection and intuition.
A third shortcoming of experiment-based questions involves the possibility that in certain cases, intuition, introspection and/or personal experience can be used to derive more accurate predictions than the descriptive models taught in the course. For an example, consider the following question:

Question 9:
A study on the use of ear protectors in large factories in Israel (Zohar et al., 1980) shows that typical workers:

- A. Use ear protectors less often than they should according to the safety rules.
- B. Use ear protectors only when instructed.
- C. Use ear protectors more often than they should according to the safety rules.

Most students would guess that the correct answer is A. However, students who take Maslow's (1970) motivation pyramid (one of the models taught in the course) too seriously are likely to err: this model implies that physiological needs and personal safety are always satisfied before addressing other needs.

Creative experimental paradigms.
A fourth shortcoming involves the substantial differences among the various experimental paradigms used to demonstrate different phenomena. Typical paradigms in the behavioral sciences involve many details that are not manipulated during the experiments, including incentives offered the participants; the cover story; general instructions given; the use of deception; the subjects' demographic makeup; and the possibility of "clarification" questions. As noted by Hertwig and Ortmann (2002), these details tend to vary from study to study. Hence, basing an exam question on any experiment not taught in class requires simplifying the description of the paradigm. In our questions we tried to keep the important details, but the distinction between important and minor details is not necessarily clear-cut, and it may be that our simplifications left out important details and so impaired the predictive ability of the good students.

4.1 Standardized descriptive models

One approach that may increase the discriminative power of experiment-based questions is provided in Erev et al. (2004). This paper suggests a procedure for standardizing descriptive models that can be used to reduce the first three problematic properties discussed above.

The suggested standardization process is similar to the standardization of psychological tests (see Anastasi, 1996). It includes two major steps. The first step starts with a translation of the relevant theory (or model) to theory-based point prediction rules. The term "point prediction rules" refers to an equation and/or computer program that uses precise input (the parameters of the situations and the parameters of the model) to make precise predictions of future behavior. The parameters of this precise version of the model are estimated by running experiments. To ensure robust estimates, the experimental conditions are randomly drawn from a well-defined universe of tasks to which the model is assumed to apply. Thus, the first part of the standardization process reduces the need to rely on biased samples of questions and

models with vague boundaries. Instead, it implies a random selection of experimental conditions, and a clear definition of the boundaries of the model.

The second stage of the standardization procedure involves estimating the optimal weighting of the point prediction and "new data." The "new data" consist of a few observations of individuals' behavior in an experiment identical to the experiment to be predicted. To clarify this concept and its relationship to the current context, it is convenient to consider a concrete example. Consider the following question:

Question 10.
In the experiment conducted by Erev et al. (2004), participants were asked to select among 100 hypothetical gambles with one non-zero outcome. One of the problems presented the following pair of gambles:

Gamble 1: Earn $60 with p = 0.80; earn 0 otherwise
Gamble 2: Earn $74 with p = 0.75; earn 0 otherwise

What was the proportion of subjects preferring Gamble 1?

A. 0.09   B. 0.39   C. 0.69   D. 0.99

The correct answer is B (0.39). An example of new data in this prediction task is the observation of the examinee's own (introspective) preferences. For example, an examinee may know that prospect theory with the parameters estimated by Tversky and Kahneman (1992) predict a choice of Gamble 1, but that her own tendency is to prefer Gamble 2. The standardization proposed by Erev et al. allows an estimation of the optimal (least squared) weighting of the two predictors (the model and the new data). The optimal weight to be given to the model is summarized with one statistic: the model's Equivalent Number of Observation (ENO). When combining the model with k new observations, the model weight is (ENO)/(ENO+k), and the data weight is (k)/(ENO+k). Thus, the availability of the ENO statistic addresses the third problem listed above: It provides guidance where predictions based on a model conflict with introspection/intuition. Notice that students do not have to learn ENO values by hurt. Deep understanding of the robust behavioral principles should allow the derivation of accurate estimates of the relevant ENO's.

4.2 Standardized experiments

Hertwig and Ortmann (2002) have argued that the large set of experimental conventions used by psychologists impairs our ability to compare findings and draw general conclusions. To address this problem, they suggest that experimenters should be encouraged to replicate their results in standardized settings. We believe that this idea can help address the final problem discussed above. Replicating the important experimental results in a particular field according to one basic standardized

paradigm would facilitate the development of experiment-based questions. Following such standardization, exam developers would not have to begin each question with a long description of the experiment's unique paradigm. It would be reasonable to expect students to know the basic paradigms.

5. SUMMARY

An analysis of GRE exams highlights an important difference between the natural and the behavioral sciences. Most questions in Physics ask the examinee to predict the results of particular experiments. On the other hand, nearly all questions in Psychology deal with abstract terms. The current analysis clarifies this difference, and proposes two related steps that can lessen the gap.

The first step addresses the difficulty of developing experiment-based questions in the behavioral sciences. We assert that the main stumbling block, from the developer's point of view, lies in identifying questions with unambiguous correct answers. The solution proposed here is technical. It requires focusing each question on a particular experiment that has been run. With this focus the correct answer is crystal clear: It is the observed experimental result. Our analysis suggests that the discriminative power of experiment-based questions based on this technical solution is at par with the discriminative power of more typical abstract questions.

The second step requires some changes in the information collected by researchers and presented to students. We assert that the discriminative power of experiment-based questions can be improved through the standardization of descriptive models and experimental procedures. The standardization of descriptive models as suggested by Erev et al. (2004) is expected to have three benefits: It would allow unbiased selection of experimental tasks; it would clarify the boundaries of descriptive models; and it would provide guidance where models conflict with intuition, introspection and or personal experience. The standardization of experimental procedures (see Hertwig and Ortmann, 2002) is expected to be beneficial in that it would facilitate clear and parsimonious presentations of experiment-based questions.

We believe that the use of experiment-based questions to evaluate students in behavioral science courses is likely to have many attractive outcomes. In addition to making behavioral science exams more similar to those in the natural sciences, this effort will advance the behavioral sciences in substantial ways. A focus on predictions in exams is likely to have a similar effect on courses, on textbooks, and on mainstream research.

ACKNOWLEDGEMENT

REFERENCES

Anastasi, A. (1996). *Psychological testing* (7th ed.). New York: Macmillian.

Brown, R. W. & McNeill, D. (1966). The tip-of-the-tungue phenomenon. *Journal of Verbal Learning and Verbal Behavior*, 5, 325–337.

Erev, I., Roth, A. E., Slonim, R. L. & Barron, G. (2004). Descriptive models as prior beliefs with known equivalent number of observations (ENO). Working paper.

Gneezy, U., Haruvy, E. & Yafe, H. (2003). The inefficiency of splitting the bill. Working paper.

Hertwig, R., & Ortmann, A. (2002). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral & Brian Sciences*, 24.

Maslow. A. (1970). *Motivation and personality* (2nd Ed.). New York: Harper and Row.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 9, 195–230.

Zohar, D., Cohen, A. & Azar, N., Promoting increased use of ear protectors in noise through information feedback, *Human Factors*, 22, 69–79, 1980.