An Empirical Study on the Nature
of Trick Test Questions

by

Dennis M. Roberts
Penn State University

February 1993
Note: The author wants to thank Dr. Henry Clark, at Northern Arizona University, for his assistance.

Copies of materials are available from the author.
Abstract

This study attempted to better define trick questions and see if students could differentiate between trick and not-trick questions. Phase 1 elicited definitions of trick questions so as to identify essential characteristics. Seven components were found. Phase 2 obtained ratings to see which components of trick questions were considered to be most crucial. The intention of the item constructor and having multiple correct answers received highest ratings. Phase 3 presented a collection of statistics items, some of which were labeled on an a priori basis, as being trick or not-trick. The analysis indicated that examinees were able to statistically differentiate between trick and not-trick items, but the difference compared to chance was small. Not-trick items were more successfully sorted than trick items, and trick items that were classified as intentional were sorted about as well as non-intentional items. Evidence seems to suggest that the concept of trickiness is not as clear as some test construction textbook authors suggest.
A frequent complaint of students is that tests, particularly course exams, are "tricky". This may refer to the test as a whole, some specific test item or, even in some cases, a tricky "professor". While the terms "trick test" and "trick test item" are often heard, the perceptions of what these terms mean are not quite so apparent. Some might claim that a question that is ambiguous is "tricky" even though the ambiguity was not intentional. Others may argue that unless there were some deliberate intent on the part of the test constructor to mislead or fool the examinee, then the items or test would not be "tricky". But, do the factors of ambiguity or intent to mislead on the part of the test constructor make any difference?

One would think that test construction books would frequently speak about this issue. However, such is not the case; very few references were found on this topic. Ebel (1979) suggested that items constructed to catch examinees off their guard (i.e.; examinees who are not alert) rather than test some specific knowledge, resort to "trickery". According to this perspective, intention of the item writer would play a major role in the definition of a trick question. Hopkins, Stanley, and Hopkins (1990) refer to trick items as those that insert some trivial word (that may be easily overlooked by the examinee) or confusing use of the negative, both of which are designed to catch the student napping. More recently, Thorndike, Cunningham, Thorndike and
Hagen (1991) describe an item as being "tricky" when some word that appears to be irrelevant to the content of the item actually turns out to be crucial. These references to trick questions appear to suggest the same basic point: the test constructor is deliberately trying to see if the examinee is paying attention to some minute detail in the question.

Further support for the lack of discussion of trick questions in measurement literature is a taxonomy of item writing rules by Haladyna & Downing (1989a). They developed a list of 43 rules based on an extensive examination of 46 leading textbooks in educational measurement. Of the 46, only 7 mentioned trick questions and all 7 argued against such practice. However, according to criteria developed by Haladyna and Downing (less than 10 textbook citations), the relative infrequent citation suggested a lack of interest in this as an important item writing rule. Unfortunately, perceived lack of importance could be due to an unclear definition of trick questions.

There also is little research that bears on this issue. In a survey article on research on item writing rules, Haladyna and Downing (1989b) found no empirical literature bearing on the impact of trick questions on psychometric characteristics of tests. However, there are some tangential references to trick items in research papers. Nield and Wintre (1986) examined attitudes and use of a test response technique
called "explain option" where examinees could explain why they made a certain response to a certain item on the test. Tests would be scored conventionally but, if the explanation of the response given to the wrong answer was strong enough, the authors would give credit for the item. Credit would also be taken away if a student gave a poor explanation to an item answered correctly. As one part of this study, students were asked to comment on various types of test formats ranging from essay to true-false. In their responses, students indicated that items of the multiple-choice and true-false variety tended in many cases to be confusing, picky and "tricky" compared to other types of test questions. Although "tricky" was not defined, there seemed to be some linking between recognition types of items and "trickiness".

Dodd and Leal (1988) explored a similar test format where students could justify their answers to items they considered to be ambiguous or "tricky". Again, examinees who made a convincing defense of an answer to an item, even if the selected response was incorrect, could obtain credit for that item. Unlike the scoring system of Nield and Wintre (1986), credit was given for convincing explanations but no penalty was assessed for poor explanations to correct responses. Results of the survey asking students about their opinions of answer justification as a test format were positive and similar to those of Nield and Wintre. Dodd and Leal argued that the answer justification format may help to
modify students' perceptions from that the instructor will, with at least some test items, try to trick the examinees into making mistakes, to a person who is genuinely concerned about assessing the students' true state of knowledge.

Although the topic of trick questions seems to be of particular interest to students, it is surprising to see that no studies have been reported that focus directly on the definition of trick items or to see if students could actually differentiate between what presumably are trick questions from not-trick questions. If one uses a dictionary definition of trick: "... an action or device designed to deceive, etc. ..." (Webster's New World Dictionary, 1959), one would assume that any definition of trick questions would necessarily have to include some element of deception. Even Haladyna and Downing (1989a) define trick items as those which mislead or deceive (emphasis added) examinees into answering incorrectly. But, if deception is present, is it due to poor item writing (accidental) or intentionally placed (deliberate)? What will people report on this issue?

The purpose of the present investigation was to explore the nature of trick questions, and to find out what the essential characteristics are that make an item or test "tricky". While a few authors briefly mention their conception of what constitutes a trick question, that does not mean that these definitions are generally held. A first step
An Empirical Study on

in providing clarity on this matter would be exploring whether there is some common view of what trick questions are, and their defining characteristics. The present study progressed through three phases: Phase 1 attempted to elicit open-ended definitions of trick questions; Phase 2 attempted to elicit more precise ratings related to which elements best constitute a definition of trick questions; and Phase 3 attempted to see if people could properly sort a mixture of trick and not-trick questions (based on previously found definitional characteristics).

**PHASE 1**

**Methods**

During Phase 1, a survey form was developed that requested 174 college students (86% graduate, 14% undergraduate) and 41 college faculty to indicate what their definitions were for trick questions. The form asked: "What do you think a "trick" question is?" and also had them indicate which type of item (multiple-choice, true-false, essay, etc.) first comes to mind when thinking of trick questions. Other survey items included: how difficult tests tend to be where you find trick questions, and whether trick items are accidentally or deliberately placed on tests.
No particular systematic sampling plan was used. Forms were circulated to faculty acquaintances at Penn State University and Indiana University of Pennsylvania with the request to administer the forms to some of their students and also to circulate the forms to other faculty colleagues. No specific instructions were given as to how to sample students or faculty. Therefore, no claim is made that the sample is highly representative of the general population of college students or faculty, though none of the present data suggest that they are not. The major purpose of Phase 1 was to see if the definitions provided by the faculty and students would allow some common components to be derived that would better differentiate between trick and not-trick questions.

Results

One question asked respondents to identify which type of item came to mind when thinking about a trick question. There was a total of 226 responses. Note that some people responded to more than one type of item without claiming that "all types" were possible. Multiple-choice, true-false, matching, essay, short answer, and any type were the categories. Of these, 48% selected multiple-choice, 20% selected true-false, less than 1% selected matching and essay, 3% selected short
answer, and 27% selected any type. Thus, with nearly 70% selecting either multiple-choice or true-false, it appears that the commonly used selection-type items are most likely to be perceived as being "tricky". Another item inquired about the difficulty level of tests on which trick questions appear. The rating scale ranged from 1 for very hard to 7 for very easy. The distribution was somewhat positively skewed (Mean = 3.18, Median = 3.00, SD = 1.62). Thus, respondents indicated, on the average, that tests they considered to be tricky tended to be relatively difficult. Finally, the last quantitative item related to whether trick questions appeared on tests accidentally or deliberately (1 = accidental, 7 = deliberate). This distribution was seriously negatively skewed (Mean = 5.00, Median = 6.00, SD = 1.90) with more than 100 of the total number of responses being either a value of 6 or 7. Thus, according to this sample, the perception was that trick items appeared on tests because they were deliberately placed on tests. One might infer from this that the factor of "intentionality" is perceived as being important to the definition of trick questions. In interpreting the results of this item, it is important to note that this item was on the reverse side of the survey form and was responded to after having given their definitions of trick questions. Thus, the appearance of this item should have had minimal influence on respondents definitions of trick questions.
An Empirical Study on

Of primary importance to Phase 1 were the open-ended definitions of trick questions and whether or not common elements would be identified across the responses. It is important to note that no specific content domain was alluded to as a context in which responses were to be made and, as noted above, these definitions were given before encountering the accidental/deliberate rating item. To quantify these data, each definition was examined and the essential components of the definition were recorded. From the definitions, 7 major themes reappeared across many of the definitions. It is important to note, however, that most reported definitions contained only one or two of these components; no definition contained all seven. The themes that appeared, with brief explanations, are as follows.

**Intention:** Items are considered to be tricky if the test constructor intended the item to be confusing or misleading. Thus, it is not the content of the item per se but rather the intention of the item writer.

**Trivial Content:** Items are considered to be tricky if the content of the item is trivial and unimportant but that the trivial point is the focus of the correct answer.

**Too Fine Answer Discrimination:** Content that is discussed at one level of precision (e.g., approximate areas under the normal
An Empirical Study on curve) and then tested at a much finer level of discrimination (e.g., decimal areas from an area table) thereby having answer options that call for distinctions that were never discussed during instruction.

**Stem Includes Noise Material:** Items with long stems, that contain information totally unrelated to the required skills to answer the item, are considered to be tricky. This component is similar to what is referred to in the literature as window dressing.

**Multiple Correct Answers:** Items that have answer choices that look so much alike (first two places of answers the same, but different third and fourth place decimals) but have a very subtle and slight difference are considered to be trick items.

**Opposite Principle:** Items that measure knowledge of content in the opposite way from which it was learned are considered to be trick items.

**Highly Ambiguous:** An item is considered to be tricky if even the smartest student essentially has to guess at the correct answer. This is not window dressing but probably refers to a very poorly written item that is so ambiguous that no one can intelligently understand what is called for in the item.

As can be seen, some of the categories appear to be independent
of the content of the item in that unless other information is known, one could not assess whether the item falls in the trick category or not. For example, the intention of the test constructor would generally not be knowable to the examinee, and the content of the item would not necessarily give any clue to the item writer's intent. Also, an item that tested a principle in the opposite way from which it was taught would not be knowable simply based on the content of the item. However, some of the categories may be determined by examining the items. For example, one may be able to examine the content of the stem of the item and the answer alternatives and make some determination as to whether there was some noise information in the stem. The same could be said for other categories such as the fineness of the discriminations amongst the answer choices, although even here, it may not be clear cut. If the instructor emphasized fine discriminations, then such options on an item may be valid. Thus, some items may be classified as falling into one or more of the categories based on content alone while placing other items into the trick category may require information external to the question itself.

PHASE 2

Methods
An Empirical Study on

Phase 2 attempted to isolate, in more precise terms, which of the categories identified in Phase 1 were considered to be most important to the definition of trick items. To explore this, two different component rating strategies were employed. In both cases, the categories identified in Phase 1 were given along with brief descriptions of what these categories meant. For the first rating strategy, each of the categories was listed on the left side of the page with a 7 point rating scale listed on the right side of the page. The scale values ranged from 1 for "not important" to the definition of a trick question to 7 which meant "very important" to the definition of a trick question. In this system, all items could have been rated as being very important or not important. In the second component rating method, a paired comparison technique was used. In this case, every pair of categories was presented asking the rater to select one or the other as being the most important to the definition of a trick question. With 7 categories, there were 21 possible pairings. As with any paired comparison procedure, if one consistently selects one particular category over all others, then some categories will necessarily receive higher ratings than others. All categories cannot be rated high or low. Both rating forms also contained an item asking (on a scale of 1 to 7) the extent to which the definition of trick questions depends on whether the item writer deliberately tries to
confuse or mislead the examinee. It is important to note that this item was presented after the ratings were completed; thus, presumably had little influence on the ratings of the trick item categories. The main reason for having two different rating methods was to attempt to show that the ratings were not method dependent.

Rating forms were circulated to a different set of faculty and students at Penn State University. Overall, 98 rating forms were obtained with 55 being the simple rating method and 43 were paired comparisons. Of the 98, there were 19 faculty and 79 students. All but one of the students were at the graduate level, and primarily came from two different statistics courses in the College of Education. Since there was a large disparity between the n's for students and faculty, data for both groups were aggregated together.

Results

Table 1 presents the descriptive statistics on the 7 categories using both methods of ratings. For the simple ratings, note that no category had a mean that was in the "not important" region on the 7 point scale. At least some people indicated that every category played some role in the definition of trick questions. However, also notice that the categories with the highest ratings, 5+ for Intention and Multiple Correct Answers, were considered to be the most important
factors in determining whether an item was tricky. In both instances, it would appear that the apparent deliberateness on the part of the item writer must be assumed. At the other end of the scale, the categories of Trivial Content and Opposite Principle were rated, relatively speaking, as being least important.

For the paired comparison data, there were 21 pairings that could be made, two categories at a time, from a base of 7 categories. If one category was selected over every other category, that category would have a score of 6; preference of that category to all of the other six. If one category was not preferred over any other category, that score value would be 0. Thus, scores for the categories using the paired comparison technique had a possible range from 6 to 0. It is important to realize that the simple rating and paired comparison values are not directly comparable in an absolute sense although, relatively speaking, higher values and lower values have similar meanings.

The two categories rated as being the most important were the Intention and Multiple Correct Answers categories, the same as was found using the simple rating method. Likewise, the categories found to be least important were the same as found to be least important using the simple ratings, Trivial Content and Opposite Principle. To better assess the consistency in the rankings across both methods, a Pearson correlation was calculated and found to be .98. It appears that both
methods identify the same ordering of categories, in terms of importance.

One final source of evidence of importance from Phase 2 was the item about how important it was, for the definition of trick questions, that the item writer had deliberately attempted to mislead or confuse the examinee when constructing the item. For both rating groups, the responses were seriously negatively skewed (Simple ratings Mean = 5.46, Median = 6, SD = 1.50; Paired comparisons Mean = 5.37, Median = 6, SD = 1.22) with most responses occurring at the high end of the scale. This evidence again suggests that the intention of the test item writer is perceived as being an important determiner of whether the item would be classified tricky or not.

PHASE 3

Methods

Using the components of trick questions identified in Phases 1 and 2, and the content of elementary statistics, a collection of trick and not-trick items was generated. Clearly, selecting statistics as the content domain represents only one small area in which test items can be constructed; therefore, no assumption is made that the results found here are comprehensive across other content domains. However, despite the image that statistics tests may have, items do cover a wide range
of item content types such as calculation, definitions, word problems, etc. Although it may seem easy to produce a collection of items like these, especially trick items, such was not the case. If one makes an assumption that the intention of the item writer is important to making an item "tricky" and depending on the context of instruction, it is not that easy to generate items that can be unquestionably viewed as being tricky. Several examples of trick items that were generated and the reasons why they were considered to be tricky, are as follows.

**Trick Item 1**: A researcher collected some data on 15 students and the mean value was 30 and the standard deviation was 3. What is the sum of X if the variance is 9 and the median is 2?

A. 3  B. 30  C. 15  D. 450

This item was seen as being tricky for 2 reasons. First and foremost, the actual point focused on, obtaining the sum of X, seems to be emphasizing trivial content. Second, even if that is a legitimate point to test on, the stem here contains considerable window dressing that is irrelevant to obtaining the correct answer. While it may not be totally unreasonable to include some extraneous material in stems, especially if the objective of the item is to see if such material can
be differentiated from the relevant, generally this is considered to be a poor item writing strategy.

**Trick Item 2**: For finding the $Y'$ regression equation, the standard deviation of $X$ is 23.34, the standard deviation of $Y$ is 18.27 and the correlation between $X$ and $Y$ is .394. What is the $b$ or slope value for finding the $X'$ value given a $Y$ value?
A. .308   B. .503   C. .783   D. 1.278

This was seen as tricky because it sets up the stem referring to the $Y'$ equation when it actually asks a question related to the reverse $X'$ equation. Testing the opposite principle from which was probably taught appears to be deliberately misleading the examinee.

**Trick Item 3**: Someone gives a test and also collects some attitude data on a group of parents in a city. The test has an average value of 30 with a variability value of 4 while the attitude average score is 13 with a variance of 18. The relationship between the two variables is .4. What is the regression equation using $X$ to predict $Y$?
A. $Y = .9 + .2 X$   B. $Y = -32 - 1.3 X$
C. $Y = 3.2 + 3.2 X$   D. None of the above
An Empirical Study on

This item was seen as tricky in that the stem is so ambiguous that even a good student would not have any reasonable idea of what to do. What is X and what is Y? Is the average referred to the mean? Is the relationship of .4 referring to a correlation? One can visualize most of the class coming up to the instructor asking for considerable clarification.

In addition to the set of trick items generated, an approximately parallel set of not-trick items was also generated. By parallel, it is meant that similar content was contained in the both the trick and not-trick questions. Two examples of not-trick items that were used are given below.

**Not-Trick Item 1**: For which of the correlations would it be MOST likely that the relationship between X and Y is curvilinear?
A. .09  B. .32  C. .92  D. .72

This item was viewed as being a direct test of the notion that the closer to 0 is the correlation, the more likely it may not reflect a straight line relationship and therefore could more possibly be something else like a curvilinear relationship.

**Not-Trick Item 2**: What is possible between the size of the
An Empirical Study on variance (V) and the size of the standard deviation (SD)?

A. V larger then SD      B. V less than SD
C. V equal to SD         D. All of the above are possible

This item was seen as testing the fact that the variance is not always larger than the SD.

From an initial set of 40 items, 20 trick and 20 not-trick, 5 items were eliminated as being too close to call by the author and one other faculty evaluator (see Note 1). This was done subjectively and means that an item that was written as being tricky bordered on not being tricky, or the reverse. There was no attempt to deliberately eliminate 5 items; that is simply the number that was arrived at by further inspection. Then, from the remaining collection of 35 items, a random set of 25 items was taken (to make the evaluation task by the students more manageable) by the use of a set of random numbers. This resulted in 13 trick items and 12 not-trick items being used. Finally, the 25 items were then randomly re-ordered so that there was no particular order of appearance of the trick and not-trick items across the set.

Students were informed that they were to read each item and then decide, on a sliding scale, whether they felt the item was tricky or not-tricky. Students were instructed that it was not their task to work
the questions, but rather, simply make an assessment as to whether the items were tricky or not. However, it is reasonable to assume that some students did attempt to work some of the problems (some booklets showed problem calculations, for example), although no data were collected on this factor. They were also asked not to refer to any text material nor to work with any other student. It is very important to note that there was no information provided about the categories of trick items that were generated in Phases 1 and 2. Students had to rely on whatever conception they had about trick test items when making their evaluations. This was the main point of the study: are students, using their own definitions of trick items, able to differentiate between trick and not trick questions?

Data were gathered from a total of 3 graduate classes of introductory statistics from Penn State University and Northern Arizona University, from their respective Colleges of Education. The total sample size was 106. Data were deliberately gathered near the end of the semester, so the topics covered in the collection of trick and not-trick questions (based on preplanning) had already been introduced earlier in the course. While the best sample of participants perhaps would be students who were actually learning the specific material on the test (which contained some trick and not-trick items), the next best sample appears to be students who have recently covered the
relevant material, although there may not be an exact congruity between the items and the way it was taught. To use students who were not familiar with the material covered by the items seems totally inappropriate given the context of the study. Having students evaluate whether items are tricky or not implies that they have a certain familiarity with the content to make an intelligent assessment.

The method of data collection was to describe the study and then ask for volunteers in the statistics classes. Volunteers were then given the packet of material and asked to complete the task outside of class and then return the completed answer sheets indicating their responses to the 25 items.

Results

The first result to be examined was the overall performance of students and whether they were able to correctly identify items as being trick or not-trick. For each of the 25 items, the student was to respond with a numerical value ranging from 4 = I definitely think it is a trick item, to 1 = I definitely think it is not a trick item. Response categories of 3 and 2 had the terms "I somewhat think ... " attached and represented a less intense level of response. There was no neutral or middle category. However, these ratings are simply the
An Empirical Study on perceptions of the students. To score the responses, a determination was first made whether the student's response was correct in the sense that it matched the "key". The key referred to is whether the item had been previously labeled as being trick or not-trick. If the student's response matched the key (matched saying ratings of 4 or 3 when the item was categorized as being trick, or 2 or 1 when the item was categorized as being not trick), it was deemed correct and then a value of 1 was assigned. If the match was not made, the score value was 0. Based on this analysis, the mean (out of a possible maximum score of 25) was 14.604 with a standard deviation of 2.657. Using a null hypothesis value of 12.5 (expected chance score based on complete random guessing; p = .5 for correct or incorrect response), the t value was significant (t = 8.15, df = 105, p < .01). Thus, from a statistical viewpoint, students were able to respond better than the expected chance score of 12.5. Even if one computes effect sizes using the standard deviation of the actual data (2.657) or the expected chance distribution standard deviation if all students were responding completely at random (binomial value of N*p*q = 25*.5*.5 = 4), one obtains values of .8 and .53 respectively. In the effect size literature, these are not small values. Despite this however, given that 25 was the maximum possible score, the absolute difference between the chance score and the mean student response was small, and suggests
that students were not able to differentiate very well.

To further explore the finding that ability to differentiate between trick and not-trick items was marginal, KR 21 was calculated on the data and the reliability value was .15. Tests of this approximate length covering similar material in the author's statistics classes typically have alpha values between .7 and .8 when scoring items for content knowledge. Thus, it is difficult to argue that the low KR 21 value is due to the poor quality of the items for measuring course achievement. However, if items are viewed as being used to measure the construct of "ability to differentiate between trick and not-trick questions", students were not reliably able to perform that task. Of course, as with any internal consistency measure, low values are difficult to interpret since the problem could be due to problems with the items or the respondents. Generally, though, low reliability values suggest problems with the items. Perhaps there were problems with certain types of items but not others. That is, was there a differential discrimination rate (success at sorting) between trick and not-trick items?

To test this hypothesis, a comparison was made between the correct response rate of the items labeled as trick or not-trick. There were 13 trick items and 12 not-trick items. If examinees were having an approximately equally difficult time sorting trick and not-trick items,
then the means for both types should have been approximately the same. Note: if there had been a larger difference in the numbers of trick and not-trick items, these would have been changed to proportions, and the hypothesis would have been in terms of equal proportions. Given the same approximate numbers of items, a simple difference in means test was performed. The mean correct sort rate for trick items was 5.453 (SD = 2.687) whereas, for not-trick items, the mean was 9.151 (SD = 2.382). Thus, for the category (not-trick) that actually had fewer items, the success rate was much higher. The difference between the means was significant using a dependent t test (t = -8.8, p < .01). Therefore, while the overall "hit" rate, determining whether items were trick or not-trick, was only marginally different than chance expectation, there clearly was a differential "hit" rate depending on whether the items were trick or not-trick. Items categorized as trick were much harder to correctly sort.

Since trick items were not correctly identified as often as not-trick items, could this sorting difficulty depend on the definitional characteristics on which the trick items were constructed? Since the literature definitions of trick questions and the Phase 1 and Phase 2 responses suggest that "intentionality" may be a major factor in determining whether an item is tricky or not, is it more difficult for examinees to correctly sort trick items of the intentional variety?
An Empirical Study on

Keep in mind that an intentional trick item is one that during Phases 1 and 2 was defined and rated as being where the item constructor has deliberately placed some trivial word in the item, that turns out to be crucial for correctly answering the question. Thus, not all trick items are of the intentional variety. In essence, in the intentional case, the item constructor is trying to see how alert the examinee is rather than assessing his or her real knowledge. Of the 13 trick items, 5 were classified as being of the intentional variety whereas the remaining 8 were not. Examinees response sheets were examined again, using only the trick items, and scores for intentional and non-intentional items were calculated. Since there were 5 and 8 of these items respectively, scores were converted into proportions. The mean proportion for intentional trick items was .3906 (SD = .271) whereas the mean proportion for the non-intentional items was .4375 (SD = .2258). Given that these means are based on the same people, a dependent t test on the difference between the two proportions was run (t = -1.81, NS). While one may hypothesize that intentional trick items would be more difficult to discriminate, and the proportion was in the predicted direction, the difference was not significant.

Discussion
Since the practical and empirical literature on the topic of trick test questions is very limited, it was difficult to formulate any firm hypotheses or expectations as to what would be found when exploring the nature of trick questions. From the test construction books that did make mention of trick items and from the taxonomy of item writing rules developed by Haladyna and Downing (1989a), the definitions seem to hinge around the notion that item writers sometimes focus (though not exclusively) on embedding a trivial word in an item that is crucial to an examinee being able to correctly answer the item. An item that presents material in the stem as though the Y' regression equation is being sought, but then asks, almost as an aside, for the slope in the X' equation, is almost hoping that the examinee will miss X' and read it as Y'. Or, an item that embeds a "not" or "no" somewhere in the final question part of the stem (particularly without emphasizing the word by underline or caps), in many cases, is hoping to catch or mislead the examinee by having him or her slide over the negative term. In both cases above, items would be classified as being tricky since it appears that the item writer was deliberately trying to fool the examinee. Clearly, test construction texts frown on this practice as being counter productive if one is attempting to assess whether the examinee does or does not understand the content material. Unfortunately, the empirical literature to establish the impact of this
practice does not exist.

The present study first attempted to provide empirical information on the perceptions of what trick questions are by obtaining definitions and ratings from students and faculty. By doing this, several dimensions of trick questions were identified such as intention on the part of the test constructor, items being highly ambiguous, and items that require discriminations in the answer options at a much finer level than was provided during instruction. Based on this information, it appeared that "intentionality" was a key factor. Then, based on these dimensions, items were constructed that were of the trick and not-trick variety, and administered to students to sort into trick and not-trick categories. Results showed that students could statistically differentiate but the difference compared to chance was small. In addition, the not-trick items were much more successfully sorted than were the trick items. It was also found that trick items that were classified as being constructed to intentionally fool the examinee were sorted about as well as the non-intentional items.

The current research certainly has limitations. First, like many survey research studies, samples are not the most representative of the desired target population. No active attempt was made to sample all levels and types of college students and faculty. Second, studies need
to make decisions about content material to use and, in the present case, materials were only drawn from the statistics domain. It is certainly possible that the results obtained here would not apply across all other domains. Finally, one may question the assumption made (currently) that the construction and classification of items into trick and not-trick categories is valid. It is certainly recognized that if there are problems in this area, then the results have limited value. However, despite these limitations, the results do suggest some possibilities about trick questions, which are as follows.

First, textbook definitions of trick questions are probably too narrow. Intention on the part of the item writer seems to be a theme in these definitions, and so is it a theme in responses of the students and faculty surveyed in this study. However, there were other dimensions identified that did not necessarily have anything to do with an item writer deliberately constructing an item to fool the student. For example, some item writers are better than others at the item construction task, and therefore poor item writers may construct highly ambiguous items, although they did not intend to do so.

Second, if intention truly is an important ingredient in the definition of trick questions, it is generally not possible for an objective item reviewer (even an examinee) to be able to uncover the intention of the item writer from the content of the item itself. It
would be a different story if the correct answer were known, but that is not the case. Correct answers are only known after the fact, not while a person is working on the item. Therefore, if intention is important, and the content of the item does not usually provide a clue to that intention, then the perceptions of examinees as to whether the item is tricky or not will perhaps be based on other contextual factors (perceived difficulty of the item, length, amount of numerical information provided in the stem, etc.). Since contextual factors may not be the true key to differentiating the defining characteristics of trick items, is it any wonder that examinees are unreliable at the sorting task and can barely sort items into trick and not-trick categories at much beyond the chance level?

Third, it may simply be the case that while examinees often complain that items or tests are tricky, examinees really do not know what constitutes a tricky item. Thus, there may be a perception of what constitutes a trick item but, that perception may be highly individual and vary substantially across examinees. If trickiness is actually multi-faceted, and not unitary as the few textbook authors who discuss this topic may suggest, then it may be difficult to construct a measure or test of whether examinees can tell the difference between trick and not-trick items.

The topic of trick questions deserves further attention despite
An Empirical Study on failing to meet the criterion of importance defined by Haladyna and Downing (1989a). Given that examinees often complain of tests or items being tricky, it would be worthwhile to explore whether the field can clarify this issue. Certainly, at the least, the present study should be extended to see if similar results are obtained using items from different content domains: history, science, etc. Also, if trick items can be more clearly conceptualized and defined, then test constructors should be able to reduce that extraneous variance from the perceptions of the quality of tests. However, the present data do not support the contention that the current definition of trick questions is clear. Future work should focus on trying to clarify the definition, if possible, and then see if trickiness is a function of factors like item difficulty, length of items, or complexity. In addition, is the ability to differentiate between trick and not-trick items a function of general ability? When all is said and done, it may be that trickiness is nothing more than the perception that some items are much harder than others. Anecdotal support for this possibility comes from the observation that difficult tests or items are the ones that seem to elicit most of the complaints and comments about tests being tricky. However, if it is true that some item constructors do, in fact, deliberately construct some items to intentionally mislead examinees, this suggests perhaps a breach of testing ethics in that tests are
An Empirical Study on
meant to fairly assess what examinees know without tomfoolery.
Hopefully, additional research will shed some light on these matters.
References


TABLE 1
Descriptive Statistics on Categories for Simple Ratings and Paired Comparisons

<table>
<thead>
<tr>
<th>Category</th>
<th>Simple Ratings</th>
<th>Paired Comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Intention</td>
<td>5.53</td>
<td>1.50</td>
</tr>
<tr>
<td>Trivial Content</td>
<td>4.11</td>
<td>2.01</td>
</tr>
<tr>
<td>Too Fine Disc</td>
<td>4.87</td>
<td>1.87</td>
</tr>
<tr>
<td>Noise in Stems</td>
<td>4.30</td>
<td>1.72</td>
</tr>
<tr>
<td>Multiple Correct</td>
<td>5.31</td>
<td>1.79</td>
</tr>
<tr>
<td>Opposite Principle</td>
<td>4.16</td>
<td>1.88</td>
</tr>
<tr>
<td>Highly Ambiguous</td>
<td>4.96</td>
<td>1.97</td>
</tr>
</tbody>
</table>

NOTE 1: N simple = 55; N paired comparisons = 43
NOTE 2: Simple ratings based on values from 1 to 7; paired comparisons value represents average number of times given stimulus was preferred over all others