Staying With Initial Answers on Objective Tests: Is it a Myth?

Ludy T. Benjamin, Jr., Texas A&M University
Timothy A. Cavell, Louisiana State University
and William R. Shallenberger, III, Texas A&M University

Since 1928, at least 33 studies have been published concerning a number of issues surrounding answer-changing behavior on objective tests. Although results in these studies have sometimes been at variance, the one consistent finding is that there is nothing inherently wrong with changing initial answers on objective tests. In fact, the evidence uniformly indicates that: (a) the majority of answer changes are from incorrect to correct and (b) most students who change their answers improve their test scores. None of the 33 studies contradicts either of those conclusions.

Most of the research in this area has been aimed at testing the accuracy of “first impressions” in test-taking. This bit of academic folk wisdom is typically stated as the belief that one should not change answers on objective tests because initial reactions to test questions are intuitively more accurate than subsequent responses.

Prevalence of the Belief and Potential Sources. That this belief is widespread among test-takers is supported by a number of studies. Surveys of students’ attitudes toward the results of answer changing are fairly consistent in their outcomes revealing that most students (between 68% and 100%) do not expect changed answers to improve their score. Indeed, approximately three out of every four of these students felt answer changes would lower their score. The percentages of students reporting an expected improvement have ranged from 0% to 32% (Mathews, 1929; Foote & Belinky, 1972; Lynch & Smith, 1975; Mueller & Shwedel, 1975; Ballance, 1977; Smith, White & Coop, 1979).

Because a majority of test-takers believe that changing answers will not improve their scores, the obvious question to ask is where they acquired such a belief. One potential source would be manuals or articles on “how to take tests.” Actually very few books on test-taking strategies have been identified as either perpetrators of the belief of “first impressions” or as proponents of answer changing. Interestingly, which side of the fence these strategists were on was often not clear. For example, Huff’s (1961) position was both pro and con, depending on the article in which he was cited (Jacobs, 1972; Pascale, 1974; Davis, 1975; Lynch & Smith, 1975; Stoffer, Davis, & Brown, 1977). For the record, Huff’s position is a qualified “yes” to answer changing, except in “a case that is very close to sheer guess” (p. 36). However, there have been some blatant examples cited of advice contrary to all the extant research. In defense of these strategies, though, it should be said that they rarely recommended complete abstinence from answer changing, rather they strongly qualified those occasions when it seemed beneficial. Millman, Bishop, and Ebel (1965) have written that a tendency to make judicious changes in one’s answers is a characteristic of “test-wiseness.” More recently, Mehrens and Lehmann (1978) have urged teachers to disabuse students of the misconception, but Crocker and Benson (1980) advise classroom teachers “specifically not to discourage response changing” (p. 239).

Because a half-dozen studies have shown the prevalence of the belief among college students, it is possible that it is reinforced by their instructors. No investigator to date has looked at faculty opinions. That oversight prompted our survey of faculty at Texas A&M University in the Colleges of Education, Liberal Arts, and Science using a brief questionnaire that inquired about belief in the outcome of answer changing on objective tests. Those faculty who indicated that they used objective tests were also asked about any instructions they gave to students about answer changing. Like students, most faculty (55.2%) believe that changing the initial answer will lower scores. Only 15.5% of those responding said they thought answer changing would improve a student’s score. Interestingly, the majority of the faculty in that group (77.7%) were members of the College of Education (see Table 1). Apparently they are more familiar with the many studies in this area, which is not surprising because much of this research has been published in the educational literature.

Do faculty members give instructions to their students about changing answers? Of those responding, about a third (32.7%) indicated that they did, and of that group, nearly two-thirds (63.1%) warned their students not to change their answer because the likely outcome would be a greater number of wrong answers. The remainder essentially cau-

Table 1. Faculty Responses to Question about the Outcome of Changing Initial Answers on Objective Tests

<table>
<thead>
<tr>
<th>College</th>
<th>N</th>
<th>Improve the Test Score</th>
<th>Hurt the Test Score</th>
<th>No Change</th>
<th>Don’t Know</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education</td>
<td>23</td>
<td>30.4%</td>
<td>52.2%</td>
<td>13.0%</td>
<td>4.3%</td>
</tr>
<tr>
<td>Liberal Arts</td>
<td>19</td>
<td>5.3%</td>
<td>52.6%</td>
<td>5.3%</td>
<td>36.8%</td>
</tr>
<tr>
<td>Science</td>
<td>16</td>
<td>6.2%</td>
<td>62.5%</td>
<td>12.5%</td>
<td>18.8%</td>
</tr>
<tr>
<td>All Faculty</td>
<td>58</td>
<td>15.5%</td>
<td>55.2%</td>
<td>10.3%</td>
<td>19.0%</td>
</tr>
</tbody>
</table>

The common advice to not change answers appears to be a mistake, but before that is certain, additional information is needed.
tioned their students to be judicious in changing their answers as they reassessed their first responses. No instructor, even those who indicated their belief that answer changing would improve scores, advised students to change answers on that basis.

College faculty may contribute to the initiation and/or maintenance of the belief, but they are not the sole source as indicated by studies noting the prevalence of the belief in high-school students and even some non-student populations. Although no investigator has asked the question, it is likely that many test-takers acquire the admonition on answer changing from their peers.

The Attitude-Behavior Question. Given the discrepancy between existing data that belie the validity of the belief regarding changing initial answers and the evidence for its persistence, one may rightfully question the influence of a belief which does not seem to affect answer-changing behavior. At least four authors have dealt with this question, although each in a different manner. Mathews (1929) encouraged his students to make changes, for doing so was “more apt to raise than lower their final scores.” The effect of this encouragement upon total changes and/or benefits from changing answers is indiscernible, however, due to an imposed penalty for guessing (which probably inhibited some answer-changing behavior) and the absence of a comparison group. After having first surveyed students’ attitudes toward answer-changing results, Foote and Belinky (1972) provided feedback to the same individuals in the form of the percentages of items on two multiple-choice tests changed from wrong to right (W/R), right to wrong (R/W), and wrong to wrong (W/W). Despite the fact that for the two tests combined, the percentages (55% for W/R changes, 21% for R/W changes, and 23% for W/W changes) disagreed with students’ expectations (only 32% expected to gain), answer-changing behavior on two subsequent tests was not noticeably altered, in terms of both the number of answers changed and results of those changes.

Likewise, Jacobs (1972) found no relationship between perceived outcomes from changing answers and net gains made. One study (Ballance, 1977) was designed expressly to look at the effect of students’ expectations on their answer-changing behavior. Having grouped subjects according to their stated belief, Ballance concluded there was no “evidence to support the conjecture that beliefs asserted by a student have an effective relationship with the number of answer changes.... [or] fluctuation in test score resulting from answer-changing.” (p. 165).

Assuming that reported expectancies are a valid measure of actual expectancies, one might then argue that answer-changing behavior was poorly correlated with students’ expectations. Implicit in this argument is the added question of the utility of debunking a “myth” which has no impact on test-taking behavior. However, in response it should be noted that on the average about 16% of all test-takers do not change any of their answers, and of those who do change answers, the average number of responses changed is around 3%. In addition, given the amount of selective perception that seems to exist among students with respect to their own answer-changing results, expectancies may still be exercising some inhibitory effects upon answer-changing behavior. Bath (1967) has pointed out that students seldom remember those items that they changed and got correct. Would positive expectations concerning answer-changing results alter answer-changing behavior? In the one study which grouped students by their expectancies, Ballance (1977) found no differences among the groups in terms of their answer-changing behavior. However, considering the small number of subjects (N = 12), conclusions here must be rather tentative. Thus, it may be the case that changing one’s belief about the consequences of one’s behavior is a difficult task, but it does not necessarily follow that answer-changing behavior is independent of the beliefs surrounding the consequences of this behavior.

Returning to the notion of selective perception among students, one could view that as a much more plausible explanation for the maintenance of this widely reported belief of “first impressions” than that of misleading test-taking manuals or occasional misinformed teacher’s instructions (Lynch & Smith, 1975; Stoffer, et al., 1977; Smith, et al., 1979). Moreover, there would have to be at least equally widespread ignorance of the data which refute this myth. In fact, this apparent ignorance has extended into the circle of those who could be expected to know better, that is, the researchers in this area. For example, Foote and Belinky (1972) in an article published after no less than ten other studies in this area, reported finding “no published research on this question” (p. 667).

Empirical Studies of Answer Changing. Given the persistence of this belief among faculty and students, it is hoped that the literature review provided here may serve to make faculty aware of the accumulated evidence on answer changing and cause them to evaluate their test-taking instructions. Such awareness may eventually alter student expectations on this question. However, the principal purpose of this review is to provide a groundwork for the future directions in research discussed at the close of this article. Although research on this topic has spanned more than 50 years, the collective knowledge gained from this research is disappointingly small. This state of affairs is due to problems of experimental design, narrowness of the questions researched, and duplication of previous studies, presumably due to ignorance of their existence. As an aid in this review, Table 2 provides a summary of test format, test content, types of subjects, and results obtained in 33 studies published between 1928 and 1983.

In this review we will analyze the research in terms of techniques for assessing answer changing, the extent of answer changing, the consequences of those changes, as well as the various subject and item variables that have been studied.

Measurement Strategies to Determine Answer Changes. Researchers have employed varied strategies in an attempt to measure accurately both the number and results of changed answers on objective tests. The vast majority, however, have simply relied upon their own ability to detect changes as indicated by observed erasures and crossouts. Several authors reported using this basic procedure (Mallinson & Miller, 1956; Bath, 1967; Johnston, 1975; Crocker & Benson, 1980). Some studies reported using interobserver agreement to check the degree of accuracy associated with
the measurement of changes (Archer & Pippert, 1962; Foote & Belinky, 1972; Stoffer, et al., 1977; Beck, 1978; Smith, et al., 1979). Typically these studies reported high interobserver reliability (e.g., 99.5%), and several chose to include only those items for which there was 100% agreement among observers. Several studies (Foote & Belinky, 1972; Davis, 1975) sought to enhance accuracy of detection of changes by back-lighting answer sheets. In one study (Archer & Pippert, 1962) this approach also included checks by another independent judge. The high interobserver agreement provides support for a number of authors who have commented on the ease and unambiguousness of detecting answer changes (Hill, 1937; Jarrett, 1948; Archer & Pippert, 1962; Foote & Belinky, 1972).

In other studies some type of obtrusive measure was used to get at the "true" number and direction of answer changes. For some, students were given special instructions (e.g., "Do not erase," "Draw a line through your first judgment that I may know what it is."). As a result of these instructions, answer-changing behavior may have been inhibited. Thus Lehman (1928) reported a relatively low rate of changers (roughly six percentage points below the median reported figure of 84%), but Lamson (1935) reported the lowest rate of items changed (2.2%) among all studies. One study (Smith, et al., 1979) which required students to use a pen resulted in 6.1% of the items being changed, one of the highest percentages reported, excluding two highly deviant observations produced apparently by special answering

### Table 2. Summary of Conditions and Results in Studies on Answer Changing

<table>
<thead>
<tr>
<th>First Author (date)</th>
<th>Test Typea</th>
<th>Itemb</th>
<th>Test Content</th>
<th>Type of Ss</th>
<th>Percent of changes</th>
<th>W-R/R-W</th>
<th>Percent of 4th</th>
<th>Changers</th>
<th>Gainers</th>
<th>Losers</th>
<th>Samers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lehman (1928)</td>
<td>Ach</td>
<td>Educ.</td>
<td>Col.</td>
<td></td>
<td>2.9 66.4 33.5 N/A</td>
<td>1.98</td>
<td>78.0</td>
<td>51.0</td>
<td>36.0</td>
<td>13.0</td>
<td></td>
</tr>
<tr>
<td>Lowe (1929)</td>
<td>Ach</td>
<td>EdPsy</td>
<td>Col.</td>
<td></td>
<td>3.2 65.3 32.3 N/A</td>
<td>2.13</td>
<td>48.8</td>
<td>29.8</td>
<td>16.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mathews (1939)</td>
<td>Ach</td>
<td>EdPsy</td>
<td>Col.</td>
<td></td>
<td>2.6 52.4 21.1 26.5</td>
<td>2.49</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mathews (1939)</td>
<td>Ach</td>
<td>PsyEd</td>
<td>Col.</td>
<td></td>
<td>2.2 65.6 34.4 N/A</td>
<td>1.91</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hill (1937)</td>
<td>Ach</td>
<td>Educ.</td>
<td>Col.</td>
<td></td>
<td>2.5 58.9 41.1 N/A</td>
<td>1.43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Berrien (1939)</td>
<td>Ach</td>
<td>Psy.</td>
<td>Col.</td>
<td></td>
<td>2.8 71.3 28.7 N/A</td>
<td>2.49</td>
<td>96.1</td>
<td>66.9</td>
<td>23.0</td>
<td>10.1</td>
<td></td>
</tr>
<tr>
<td>Jarrett (1948)</td>
<td>Ach</td>
<td>Psy.</td>
<td>Col.</td>
<td></td>
<td>3.9 69.3 16.6 14.1</td>
<td>4.18</td>
<td>89.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jarrett (1948)</td>
<td>Ach</td>
<td>Psy.</td>
<td>Col.</td>
<td></td>
<td>6.2 48.9 23.0 28.1</td>
<td>2.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reile (1952)</td>
<td>Ach</td>
<td>Sci.</td>
<td>SS</td>
<td>Col.</td>
<td>48.0 26.7 25.1</td>
<td>1.80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mallinson (1956)</td>
<td>Ach</td>
<td>Educ.</td>
<td>Col.</td>
<td></td>
<td>3.3 44.5 16.5 39.0</td>
<td>2.70</td>
<td>86.2</td>
<td>68.1</td>
<td>24.6</td>
<td>7.2</td>
<td></td>
</tr>
<tr>
<td>Best (1975)</td>
<td>Ach</td>
<td>Psy.</td>
<td>Col.</td>
<td></td>
<td>3.8 54.5 21.6 23.9</td>
<td>2.52</td>
<td>95.6</td>
<td>76.9</td>
<td>4.6</td>
<td>18.5</td>
<td></td>
</tr>
<tr>
<td>Jacobs (1972)</td>
<td>Ach</td>
<td>ChildDev</td>
<td>Col.</td>
<td></td>
<td>56.0 24.0 20.0</td>
<td>2.33</td>
<td>88.0</td>
<td>64.6</td>
<td>15.9</td>
<td>19.5</td>
<td></td>
</tr>
<tr>
<td>Crocker (1972)</td>
<td>Ach</td>
<td>Psychol.</td>
<td>H.S.</td>
<td></td>
<td>4.3 59.8 20.1 20.1</td>
<td>2.97</td>
<td>93.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Johnston (1975)</td>
<td>Ach</td>
<td>Educ.</td>
<td>Col.</td>
<td></td>
<td>5.5 58.0 21.0 21.0</td>
<td>2.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lynch (1975)</td>
<td>Ach</td>
<td>Educ.</td>
<td>Col.</td>
<td></td>
<td>5.4 57.6 20.3 22.1</td>
<td>2.83</td>
<td>74.5</td>
<td>69.6</td>
<td>30.3</td>
<td>2.9</td>
<td></td>
</tr>
<tr>
<td>Mueller (1975)</td>
<td>Ach</td>
<td>Educ.</td>
<td>Grad.</td>
<td></td>
<td>2.5 56.5 26.9 16.7</td>
<td>2.10</td>
<td>83.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smith (1976)</td>
<td>Ach</td>
<td>Educ.</td>
<td>Col.</td>
<td></td>
<td>3.7</td>
<td>3.00</td>
<td>60.0</td>
<td>66.5</td>
<td>17.0</td>
<td>18.0</td>
<td></td>
</tr>
<tr>
<td>Eccles (1977)</td>
<td>Ach</td>
<td>GED</td>
<td>Col.</td>
<td></td>
<td>4.0 46.5 19.2 34.0</td>
<td>2.43</td>
<td>84.8</td>
<td>60.8</td>
<td>11.4</td>
<td>12.7</td>
<td></td>
</tr>
<tr>
<td>Edwards (1977)</td>
<td>Ach</td>
<td>Grad.</td>
<td>H.S.</td>
<td></td>
<td>2.5 66.7 6.7 26.7</td>
<td>10.00</td>
<td>60.7</td>
<td>67.0</td>
<td>17.0</td>
<td>13.0</td>
<td></td>
</tr>
<tr>
<td>Stoffer (1977)</td>
<td>Ach</td>
<td>Psy.</td>
<td>Col.</td>
<td></td>
<td>3.8 65.0 20.0 16.0</td>
<td>2.83</td>
<td>60.3</td>
<td>72.0</td>
<td>14.0</td>
<td>14.0</td>
<td></td>
</tr>
<tr>
<td>Stoffer (1977)</td>
<td>Ach</td>
<td>CritRep</td>
<td>U.S.A.F.</td>
<td></td>
<td>3.1 64.5 17.0 18.5</td>
<td>3.79</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beck (1978)</td>
<td>Ach</td>
<td>Reading</td>
<td>3rd Grd</td>
<td></td>
<td>2.8 58.0 18.7 23.3</td>
<td>3.10</td>
<td>75.5</td>
<td>10.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Johnstone (1978)</td>
<td>Ach</td>
<td>Psy.</td>
<td>Col.</td>
<td></td>
<td>3.2 61.7 20.8 17.4</td>
<td>2.96</td>
<td>74.7</td>
<td>10.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smith (1979)</td>
<td>Ach</td>
<td>Psy.</td>
<td>Col.</td>
<td></td>
<td>6.1 62.3 25.6 12.1</td>
<td>2.44</td>
<td>86.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crocker (1980)</td>
<td>Ach</td>
<td>Math.</td>
<td>7th Grd</td>
<td></td>
<td>4.0</td>
<td>57.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sitton (1980)</td>
<td>Ach</td>
<td>Psy.</td>
<td>Col.</td>
<td></td>
<td>6.1 25.0 13.3</td>
<td>2.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vidler (1980)</td>
<td>Ach</td>
<td>Crit.Rep</td>
<td>Col.</td>
<td></td>
<td>3.8 61.5 38.5</td>
<td>1.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Range (1982)</td>
<td>Ach</td>
<td>Psy.</td>
<td>Col.</td>
<td></td>
<td>3.7 55.7 20.0 24.5</td>
<td>2.80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Skinner (1983)</td>
<td>Ach</td>
<td>Psy.</td>
<td>Col.</td>
<td></td>
<td>4.0 51.5 26.3 22.3</td>
<td>2.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*aAch = Achievement; Apt = Aptitude; T = Timed
bTF = True-False items; MC = Multiple choice items
cRatio of W-R changes to R-W changes
dOutcome of answer changing: "Gainers" are those with net increases; "Losers" are those with net losses; "Samers" are those answer changers whose net score does not change
nNot clearly specified in the study
fFigure combines data from true-false and multiple-choice tests
gSpecial testing instructions and procedures appear responsible for this value
formats. That this figure supports an argument that many erasures go undetected is plausible but certainly not evidence enough.

Special answer sheets were used in three studies. Typically these sheets included spaces for marking subsequent as well as initial, answers. Edwards and Marshall (1977) also used tests with carbon backing (for detecting erasures) and imposed a penalty for answer changing. In their 1929 study, Lowe and Crawford required that students read over the entire test before marking any answers in the "second answer" space. Jacobs (1972) used a slide projector to present test items. Items were exposed for only 30 to 45 seconds, depending on their word count. After all slides had been viewed and attempted, mimeographed copies of the test were passed out with red pencils for making answer changes. Considering that the median reported percentages of all items which are changed is 3.3%, one would suspect each of these last two strategies to be a factor in their substantially larger percentages—10.2% and 32.7%, respectively.

In short, the use of obtrusive strategies for detecting answer changes cannot be justified in light of the evidence. Reports of high interobserver reliability in unobtrusive procedures, the possibility of altered change rates using obtrusive procedures, and the apparent ease of detecting answer changes all argue against the use of obtrusive measures.

**Extent of Answer-Changing Behavior.** Most people change at least one or more answers on exams. The data on percentage of changers are somewhat clouded by the fact that some studies have reported results per exam, but other studies averaged figures across exams. This problem notwithstanding, data from 15 studies indicate that the percentage of persons changing one or more answers ranged from 57% to 96% with an estimated median 84%. This estimate of the median of the combined studies is simply the median of the reported medians. For the combination of other research, the same procedure is used.

In terms of the proportion of items which were changed, the amount was consistently very small except, as indicated earlier, in those instances where an obtrusive experimental measurement strategy was employed. Excluding those two studies which reported such highly inflated percentages, the data from the 28 studies reporting this figure ranged from a low of 2.2% to a high of approximately 9.0%. The estimated median percentage of items changed was only 3.3%. These percentages are based on a large number of response observations ranging from a low of 1,800 to a high of 144,370 with the median number of responses being approximately 12,000. These percentages on the extent of answer-changing behavior are comparable to those reported in an earlier article by Mueller and Wasser (1977), even though the current review contains data from an additional eighteen studies. In short, the data on this question are very consistent in terms of the proportion of persons changing answers and the proportion of answers changed.

**Consequences of Answer-Changing Behavior.** That most individuals improved their scores and that most item response changes are from wrong to right (W/R) are two observations found consistently throughout the literature on answer-changing behavior. For purposes of discussion of these data we will define "gainers" as individuals who improve their score through changes, "losers" as those whose score is lowered after making changes, and "samers" as those whose score is neither raised nor lowered as a consequence of changing answers. Data on the proportions of these outcome categories should only be compared with respect to the item formats used on exams. True-false items, unless ultimately left blank, involve only W/R and R/W changes, whereas multiple-choice items also allow for W/W changes. Moreover, the nature of the item formats may entail response tendencies which differ in their propensity to be changed in a particular direction (W/R, R/W, W/W). Thus, not only is it the case that one who changes an initially wrong answer on a true-false test marks a correct answer, and such is not necessarily true for multiple-choice tests, but it is also the case that multiple-choice and true-false items may involve recall tasks of different natures with correspondingly different results. This distinction is most critical when discussing answer change results by item responses since there is no W/W category for true-false item change results.

Two studies have reported response changes by individual subjects taking tests involving only true-false items, and one of those (Lowe & Crawford, 1929) listed only the percentage of students who made only W/R changes (49%) and only R/W changes (16%). Consequently, those individuals who made both changes (35% of all changes) but were also possible gainers, losers, or samers were not included in the data. In the other study, Lehman (1928) found 51% of all changers were gainers, 36% losers, and 13% samers. Two studies which used individuals as the unit of analysis involved tests with combined multiple-choice and true-false formats (Berrien, 1939; Mueller & Shwedel, 1975). The respective percentages from the two studies were as follows: 67% and 65% improved their score, 23% and 17% lost points, and 10% and 18% did neither.

A general discussion of the consequences of answer-changing behavior by students taking multiple-choice tests (see Table 3) is hindered by the fact that some studies, rather than report the proportion of changers who gained,

**Table 3. Range and Estimated Median Values for Six Studies on Answer Changing Outcome in Multiple-Choice Tests**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Students who Gained</th>
<th>Students who Lost</th>
<th>Students who Stayed the Same</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highest</td>
<td>76.9%</td>
<td>30.3%</td>
<td>18.5%</td>
</tr>
<tr>
<td>Lowest</td>
<td>47.0%</td>
<td>2.4%</td>
<td>2.9%</td>
</tr>
<tr>
<td>Estimated Median</td>
<td>67.5%</td>
<td>15.0%</td>
<td>14.0%</td>
</tr>
</tbody>
</table>

etc., chose to report the proportion of total test-takers who experienced the consequences of answer changes. Using this format, Johnston (1978) and Best (1979) reported virtually identical results with 75% of all subjects gaining, 10% losing, and 15% either staying the same or changing no answers.

The percentages of changers (based on six studies) who gained points on multiple-choice tests ranged from 47% to 77% with an estimated median of 67%. One can see that had Johnston (1978) and Best (1979) excluded those stu-
students who had made no changes, their figures would have been even higher, and thus some of the largest percentages reported. The percentages of losers ranged from 2.4% to 30% with an estimated median percentage of 15. The estimated median percentage of students who neither improved nor hurt their grade, but did change answers, was 14%, with a range of samers from 3% to 16%.

Recognizing the sparse data on true-false items, there nevertheless seems to be an increase in the proportion of individuals who gain changing item responses as the test formats change from true-false to multiple-choice. This fact is illustrated by the median ratios of numbers of gainers to losers for true-false (1.43 to 1), combination (3.9 to 1), and multiple-choice formats (5.3 to 1). Regardless of test format, all studies concluded that the majority of students improve their score by changing initial answers.

Answer changing results as reported by item responses are typically stated in terms of the percentages of total changes which were W/R, R/W, and WW changes (see Table 4A). For true-false tests, W/R changed ranged from 59% to 71% with an estimated median of 66% from the five studies reporting. The one study (Lynch & Smith, 1975) that used both types of items in one format and which looked at these proportions found 57% of all item responses changes were in the W/R category. Twenty studies used tests having only multiple-choice items. Percentages ranged from 44% to 69% with an estimated median of 58% for W/R changes. For R/W changes, the true-false item studies had percentages ranging from 29% to 41% with an estimated median of 33%, for the one combined format, 27% of the changes were R/W, whereas multiple-choice formats produced R/W change percentages ranging from 7% to 26% with an estimated median of 20%. Of course, no data exist for WW changes on true-false tests, but the one combination format produced 17% WW changes. For multiple-choice item WW changes, percentages ranged from 12% to 39% with an estimated median of 23%.

Figures on the ratios of W/R to R/W changes offer a good summary of item response change results from the different test formats. Studies using true-false items had ratios ranging from 1.43 to 1 to 2.49 to 1 with an estimated median of 1.98 to 1. Mueller and Shwedel (1975) used a combination format but did not report on the breakdown of percentages

for each change category. However, they did find a ratio of W/R to R/W changes of 5.3 to 1 compared to a ratio of 2.1 to 1 in a similar study reported by Lynch and Smith (1975). Combining the data from these two studies produces an estimated median ratio of 3.7 to 1. Multiple-choice item response changes produced W/R:R/W ratios (see Table 4B) ranging from 1.80 to 1 to 10 to 1 (the next highest being 4.18 to 1) with an estimated median of 2.77 to 1. Ignoring the figures from the combination formats (based on only two studies using unstated proportions of true-false to multiple-choice items), and focusing on the median ratios of true-false and multiple-choice item response change results, allows the following conclusions: Approximately every three changes on the true-false tests examined resulted in two correct and one incorrect answers whereas every four changes on multiple-choice tests have produced three correct and one incorrect answers. This second ratio does not include WW answer changes because those changes do not hurt the students' performance.

Subject Variables. Many studies have attempted to differentiate between changers and non-changers and among gainers, losers, and samers based on some external criterion. Typical variables investigated have been academic ability, sex, and several personality variables.

Academic Ability. The variable most often used to indicate a student's ability has been the student's test scores: sometimes scores for the entire course and sometimes scores for the particular exam under study. Of course, there exists a confounding of any relationships proposed to exist between test scores and numbers and types (W/R, R/W) of revisions made due to the fact that the former is partly a product of the latter. Lehman (1928) was the first to note this relationship and others have followed suit (e.g., Mueller & Shwedel, 1975; Best, 1979) attempting either to circumvent or minimize the importance of this problem. Lehman re-examined tests using initial answers only and found only "two or three" students whose position changed with respect to the median. Mueller and Shwedel chose to look at the relationship between test scores and WW changes in order to assess independently which students made the greatest gains in answer changing. They reported a significant negative correlation (−.22) between total scores and WW changes. Thus, higher-scoring students made fewer WW changes. Best emphasized the unimportance of such confounding after observing the small proportion of items changed per test compared to the number of points necessary to alter one's grade. Best's observations, while they may be valid in the context of his study, are not necessarily generalizable to all studies of this sort.

The problems of using test scores as a subject/classification variable notwithstanding, several investigators have looked at the possible relationship between a student's test score and his or her answer-changing behavior. If one were to view extensive answer changing as a sign of test-taking uncertainty, then it is not surprising to find that most authors report an inverse relationship between test scores and the numbers of revisions made by students: Six studies reported a statistically significant negative relationship (Lynch & Smith, 1975; Mueller & Shwedel, 1975; Stoffer et al., 1977; Johnston, 1978; Best, 1979; Sitton, Adams & Anderson, 1978).
of these studies showed a slightly higher frequency of answer changing by females and one study (Skinner, 1983) reported the only statistically significant sex difference, finding that females changed twice as many answers as males. With regard to gains made from answer changing, Reile and Briggs (1952) reported that females profited less, but Bath (1967) found that females made larger gains than males. Both studies have problems. Reile and Briggs suggested that their findings might be confounded by some pre-existing performance differences which restricted the potential for gains by females, but an analysis of Bath's data on sex differences simply does not support his conclusion. In short, if sex differences do exist, they seem to be in the frequency of answer changing and not in the gains made from those changes.

Personality. Several writers have suggested that personality variables may contribute to the separation between changers and nonchangers and between gainers and losers. To date, only three investigators have reported research in this area. In an *a posteriori* analysis of subjects' scores on Rotter's I-E scale, Stoffer, et al. (1977) found no relationship between the I-E measure and the number of revisions a student made. Sitton, et al. (1980) and Range, et al. (1982) have correlated a host of personality variables with frequencies and outcomes of answer changing. They report that anxious students change more answers than non-anxious ones, and that non-depressed students are more successful in their answer changes. One variable that may be related to answer changing is impulsivity (Mueller & Shwedel, 1975), however it has not been investigated to date.

Item Variables. Researchers have only occasionally looked at the characteristics of test items in their efforts to detect some distinguishable pattern of answer-changing behavior. Item difficulty and item position are the variables that have received the most attention.

Item difficulty has been defined in varying ways by researchers. Most often, p-values (i.e., the proportion of students who answered an item correctly) were used, either derived from the sample data or from previously studied samples. Jacobs (1972) and Beck (1978) used previously obtained p-values to establish item groups based on level of difficulty. Lynch and Smith (1975) and Jackson (1978) apparently based their correlational analysis of item difficulty on p-values, also. Best (1979) formed his easy-difficult item grouping based on whether items fell above or below the median percentage correct for all test items. Berrien (1939) and Pascale (1974) chose to rate whole tests as either “difficult” or “easy,” the former doing so intuitively and the latter doing so empirically based on the mean p-values of the two tests employed. As with test scores, measures of item difficulty derived from the sample data may be open to confounding if linked to measures such as net profit due to answer changing. Still it seems unlikely that item difficulty values would be substantially affected by response changes (Best, 1979).

Generally it can be said that difficult items are more likely to be changed than easier ones. Most of the studies indicate trends in that direction, however, only four (Lynch & Smith, 1975; Beck, 1978; Jackson, 1978; Vidler & Hansen, 1980) have reported results that reached statistical significance.
Of course for poorer students, that is, those who are ill-prepared for the test, many more items will be seen as difficult (Lynch & Smith, 1975). That observation can be used to explain the fact that those students typically change more of their answers.

Data on the relationship between item difficulty and profit are conflicting, and the reason for these conflicts is not at all apparent. Jacobs (1972) found most net gains involved changes of items which were either moderately or highly difficult. Similarly, Pascale (1974) found a significant positive correlation between a test's difficulty and the number of additional correct answers due to revisions. Beck (1978), however, reported that changes on easy items were significantly more likely to be correct than changes on hard items. Controlling for test performance, Best's (1979) data provide further evidence on the opposing side based on a comparison of W/R:R/W ratios for difficult (2.47) versus easy items (4.28).

It seems reasonable to conclude that item position might be an important variable. Most tests have a time limit, so review of the answers may not proceed beyond the earlier test items. In addition, seeing items later in the test may provide a context in which answers for earlier items might be reviewed and changed. Thus, on an intuitive basis, one could argue that items at the beginning of a test would be changed more frequently than items at the end. Only two studies have examined the effect of item position on answer-changing behavior and both found that fewer changes were made near the end of the test (Reiie & Briggs, 1952: Jackson, 1978). Jackson also examined alternatives within an item subgroup using partial correlation. Results showed item position within a test subgroup of "some specific stimulus material" produced significant zero-order correlations with the number of answers changed. The trend was for earlier items to have more changes, an effect described by Jackson as due to novelty. No one has looked at the relationship between item position and net gains from changing.

In addition to level of difficulty and position, two other item variables have been researched. Reiieing and Taylor (1972) attempted to relate items which involved "analytical reasoning" to answer-changing behavior but found no relationship to exist with either the amount of answer changing or the net gain enjoyed. Jackson (1978) did find some evidence (significant results for one of the three samples he tested) that items of low discriminating power tend to be changed more than answers to other items.

Although the research on the influence of item variables on answer-changing behavior is fairly sparse, Jackson's view that changed items are often misinterpreted items is given some credence by the results. Item difficulty, novelty, position, and discriminability would all seem to affect, to varying degrees, the propensity of an item being changed.

Conclusions and Future Directions. After more than a half-century of research on this topic, what do we know about answer changing? Based on the 33 studies reviewed in this article we know that: (a) only a small percentage of items are actually changed, (b) most of these changes are from wrong to right answers, (c) most test-takers are answer changers, and (d) most answer changers are point gainers. These findings would appear to refute the widespread belief about the consequences of answer changing as well as questioning what effect, if any, that belief may have on actual behavior.

Having established that most test-takers change answers on objective tests in spite of their purported reasons, what are their reasons for making those changes? Amazingly, no researcher to date has attempted seriously to answer this question. Obviously there are different reasons for changing an answer, for example, marking the wrong space on the answer sheet or simply misinterpreting the item (e.g., not seeing the negative in the stem of the item). In these cases it is most unlikely that the test-taker will believe the assertion about the correctness of initial answers, and thus will readily make those changes. Another reason for changing an answer is that the test-taker is unsure of the answer; indeed, it may have been a guess. In this situation, answer changing may be pursued with greater reluctance. In fact, it is possible that in this latter case, the myth may be no myth at all.

In short, the research to date has treated all answer changes as though they were the same. That procedure does not permit a real answer to the questions, indeed, it ignores the crucial question on this topic. Future research on answer changing must begin with a way to segregate answer changes with respect to the reasons for those changes and should focus on those answer change situations in which the answer is altered due to some cognitive, evaluation. Indeed, studies of the decisional process itself could shed more light on answer changing. In the research to date we are limited to studying the items that were actually changed, and are unable to know how many other items were considered but left unchanged. Both of these issues, that is, the reasons for change and the consideration of items that were not changed, could be studied by an inquiry procedure immediately following the test. Such a procedure would be awkward under most testing conditions, which is to say that researchers may need to construct their own testing situations specific to the problem, rather than relying exclusively on extant samples and tests that are part of college courses.

A related variable to be considered in this research is the confidence the student has in a particular answer. What confidence level has to exist before a student would be willing to change an answer? Only Skinner (1983) has looked at that question. He reported that students felt a different answer should have a high probability (74.6%) of being correct before a change was made. Further, how is confidence level as a variable affecting answer changing related to the student's level of belief in the caveat about answer changing? These questions are the most basic in this area, yet they have been largely ignored in the extant research.

Almost all of the studies to date have used regular course exams to evaluate answer changing. These are achievement tests which are typically administered under a time limit, although they are not speed tests in the strictest interpretation of that label. Students may have time to answer all of the questions, but it is unlikely that they have the opportunity to review their answers thoroughly, thus adding to the potential for answer change. Although it was impossible to discern in all cases, we could not identify a
single study in the 33 reviewed here which used a testing procedure that allowed test-takers as much time as they wanted to complete the test. Time constraints are always a critical variable in test taking, and in research assessing answer changing one ought to be concerned that students have the opportunity to change answers. It seems quite plausible to assume that rates and, perhaps, consequences of answer changing could be affected by timed versus power testing conditions.

Other test variables are also potentially important issues in this research. Do test-takers perform differently in terms of answer changing on aptitude versus achievement tests? To date only two studies (Beck, 1978; Vicler & Hansen, 1980) have used an aptitude test. Similarly, these studies have used tests that are largely norm-referenced and few that are domain-referenced. Answer changing may or may not differ between those types of tests. It is impossible to answer that question based on the research to date, but we believe it is a question worth pursuing. Researchers in this area need to state the purpose of the tests they are using. Although testing purpose is often implied, a more definitive statement would add to the understanding of this literature. Further, penalties for guessing should be studied as they affect the frequency of answer changing and the outcome of those changes.

Another area for future research concerns the apparent discrepancy between what students say about answer changing and what they actually do, an issue which we referred to earlier as the "attitude-behavior question." So much evidence in psychology has indicated a strong correspondence between belief and action that it calls for investigation of this discrepancy. Yet such an analysis cannot be answered unless we know why students change answers, and under what conditions. That information is mandatory if we are to understand the way in which the belief operates.

Further, we would encourage additional research focusing on the interrelationships of item variables (such as difficulty and nature of item distractors) and cognitive strategies of test takers. Research to date has emphasized academic ability, using mostly confounded measures, while ignoring some better measure of both cognitive ability and style. Among the subject variables, we believe that cognitive variables are likely to be far more meaningful than others studied thus far, such as sex or certain personality traits.

In conclusion, this review has noted a remarkable consistency of results across 33 separate investigations. That kind of consistency is rare in psychology and the authors are duly impressed with the harmony of those data. That consistency of results has caused a number of researchers in this area to argue that the statement about the accuracy of first impressions is a myth. That conclusion may be accurate, but we do not believe that it is justified on the evidence provided to date. Until researchers are able to segregate answer changing responses, thus separating those that are the result of "second impressions" from other kinds of changes, the conclusions of this research must be considered tentative.

The question of whether answers should be changed or not is an important one, not only for test-taking strategies, but for an understanding of the psychological processes involved in that behavior. Our hope is that future research will seek to systematically answer the questions raised in this review.

References

Archer, N. S., & Plopper, R. Don't change the answer! An expose of the perennial myth that the first choices are always the correct ones. Clearing House, 1962, 37, 39-41.


Best, J. B. Item difficulty and answer changing. Teaching of Psychology, 1979, 6, 228-230.


Jarrett, R. F. The extra-change nature of changes in students' responses to objective test items. Journal of General Psychology, 1948, 38, 243-250.


Lamson, E. E. What happens when the second judgment is recorded in a true-false test? Journal of Educational Psychology, 1935, 26, 223-227.

Lamson, H. C. Does it pay to change initial decisions in a true-false test? School and Society, 1928, 28, 456-458.


Muller, D. J., & Shwedel, A. Some correlates of net gain resultant from answer changing on objective achievement test items. Journal of Educational Measurement, 1975, 12, 251-254.
Undergraduate Psychology Research Conferences: Goals, Policies, and Procedures

Alan L. Carsrud, University of Texas at Austin
Joseph J. Palladino, Indiana State University Evansville
Elizabeth Decker Tanke, University of Santa Clara
Lyn Aubrecht and R. John Huber, Meredith College

The purposes of this paper are to: (a) examine the assumptions underlying undergraduate research in psychology, (b) present survey data that describe the current state of undergraduate psychology research conferences, (c) explore the roles undergraduate research conferences play in science education, and (d) provide suggestions for faculty members who would like to organize undergraduate conferences in such a way as to maximize the contributions to the educational process.

Assumptions About Science Education. Psychology educators have traditionally assumed that a knowledge of research methodology and concepts is essential. "Past conferences on undergraduate education in psychology have affirmed the importance of instruction in statistics and methodology in the major program. Consideration of the role that methodology plays in a frontier science indicates that the emphasis has been appropriate" (Kulik, Brown, Vestewig, & Wright, 1973, p. 55). Cole and Van Krevlen (1977) surveyed psychology departments in small liberal arts colleges and found that "Courses in statistics, research design and methodology were by far the most commonly identified as appropriately required of all psychology majors" (p. 165). Furthermore, their respondents clearly believed that knowledge of methodology is "an important contribution . . . to a student's liberal arts education" (p. 167). In a survey of psychology baccalaureates, Davis (Note 2) found that courses in statistics and experimental design were most frequently cited as useful by the respondents in their present occupations, even though the occupations were not directly related to psychology.

An examination of the "typical" undergraduate psychology curriculum indicates that the suggested emphasis on research methods is not always realized in practice. Many psychology departments require only one or two methodology courses for completion of a major. The typical content course in psychology provides little direct experience with the majority of techniques and methods used to acquire the facts studied. Commonalities in design and analyses exist regardless of the content area; for example, almost all studies use control groups and certain statistical tests. A scientist-teacher of invertebrate physiology or inorganic chemistry would probably not teach the content of an area without giving students direct experience with the research techniques that generated those facts. The opposite is the case for many psychologists who teach content courses with little direct contact with the methodology involved. As Campbell and Fiske (1959) have noted, the majority of variance accounted for in most studies is related to the methods used rather than the concepts being studied, yet we often do not teach this to undergraduates.

In addition, the particular techniques taught in research methods courses are frequently limited by the interests and/or training of the instructor. Thus, instructors may ignore,