Running head: APOPHENIA AND MULTIPLE CHOICE EXAM PERFORMANCE

Effects of Apophenia on Multiple-Choice Exam Performance

*Stephen T. Paul
Samantha Monda
S. Maria Olausson,
and
Brenna Reed-Daley
Robert Morris University

*Correspondence concerning this article should be addressed to Stephen T. Paul, Social Sciences, Robert Morris University, 6001 University Boulevard, Moon Township, PA 15108-1189. E-mail: paul@rmu.edu.

Abstract

There is a broad literature on the various issues related to effective exam construction applicable to both on-ground and on-line course delivery. These guidelines tend to support rather close contact between the instructor and the exam. However, to remain competitive, both textbook and course management providers have developed technologies to automate many aspects of exam construction. As test construction becomes automated, the possibility of inadvertently deviating from demonstrated or intuitive guidelines increases. Two experiments were conducted to examine the degree to which apophenia (perceiving patterns in random data) might negatively influence multiple choice exam performance among college students. Experiment one indirectly demonstrated the extent to which certain answer patterns seemed to be tolerated among students (maximum of three repeated answers) in comparison with what might be expected from randomly generated exams from BlackBoard™. Experiment two directly examined the effects of answer patterns on exam performance. Participants' performance declined as the underlying answer patterns became more obvious, and this effect appeared to be particularly strong for the upper-level psychology students. The importance and implications of these findings with regard to automated test construction were discussed and a recommendation is provided.

Keywords: APOPHENIA, BLACKBOARD, EXAM CONSTRUCTION, MULTIPLE CHOICE

Effects of Apophenia on Multiple-Choice Exam Performance

It has been argued that the ability to distinguish patterns (i.e., "sameness" from "differentness") holds adaptive value for humans and animals (Wasserman, Young, & Cook, 2004). Indeed, the apparent need to identify patterns is so strong as to produce relatively frequent (and newsworthy) accounts of pareidolia such as the "face on mars" and religious images burned into grilled cheese sandwiches. The broad term for identifying or perceiving patterns in random or meaningless data – where such patterns are neither present nor intended, is apophenia (cf. Carroll, 2003). To the extent that students perceive underlying patterns (whether intentional or not) within the answer keys of multiple choice exams, outcomes unrelated to content knowledge seem likely.

Although there are numerous exam formats available, multiple choice exams are particularly popular among students and teachers for differing reasons. Students tend to prefer multiple choice questions because recognition based performance is often superior to recall based performance (Hart, 1965). For teachers, advantages of multiple choice exams include ease of construction, as many instructor resources linked to textbooks include test generators which automate a great deal of the process of test building. This benefit extends to online learning management systems such as Blackboard^{tm} given that most textbook publishers provide textbook specific test-bank modules that can be imported directly into course shells. In addition, multiple choice exams are almost always less effortful and less time consuming to score than written response exams. However, despite these apparent advantages, a great deal of research has shown that there are many pitfalls to avoid when constructing effective multiple choice exams (Haladyna & Downing, 1989; Haladyna, Downing, & Rodriguez, 2002; Hogan & Murphy, 2007).

Among these pitfalls are the potential construction biases that can emerge when instructors over-represent or under-use certain response options. As Mitchell (1974) showed, some response alternatives tend to be either over-represented (often "C") or under-represented (often "A") in exams. More recently, Attali and Bar-Hillel (2003) have shown that both test takers and test makers appear to be biased in favor of answer choices located centrally among the alternatives in a ratio of 3 or 4 to 1. This bias will produce answer keys that are unbalanced in the sense that not all answer choices are equally represented. However, evidence regarding whether this concern is important or not is both mixed and outdated. For example, Wevrick (1962) found that positional bias led to participant response sets. Yet, neither Hopkins and Hopkins (1964) nor Jessell and Sullins (1975) were able to demonstrate similar response sets.

In striking contrast to positional bias, some concerns have been raised over deliberately balancing the answer key. Bar-Hillel and Attali (2002) note that systematic attempts to derive answer keys that produce equal representation (as appears to have been done for SAT exams) may be exploited by savvy test takers. Thus the argument against balanced answer keys is that, in such cases, test takers can use answer counting strategies to improve the odds of guessing correct answers. Of course, effective use of such a strategy presumes a relatively high degree of content knowledge among exam takers (i.e., weaker students would not clearly benefit from answer counting because their counts would be inherently inaccurate). Nonetheless, Bar-Hillel and Attali argue that random assignment of correct responses to response position should be the preferred technique as answer counting would not benefit students of any caliber. Such a tactic is made relatively easy given the widespread availability of electronic exam generators offered by nearly all of the mainstream textbook publishers or built into the online delivery systems (e.g., Moodle™, Edmodo™, Blackboard™, Canvas™, eCollege™).

Multiple-choice exams may be popular in online courses simply due to their ease of implementation. The online market is rapidly expanding and colleges and universities are working to develop offerings to match (or sometimes to generate) enrollment interest even at the community college level (Brent, 2010). Online multiple-choice exams can be constructed automatically by specifying the desired number of items to be drawn randomly from a larger pool of relevant questions. An advantage of employing such options is that it results in a unique exam for each student taking the course which makes it more challenging for students to engage in academic dishonesty (Chiesl, 2007).

However, once instructors transition to, or in any other way come to rely on, automatically randomized exam construction, a potential danger emerges. Specifically, a random determination of items provides the opportunity for correct answer responses to inadvertently result in answer key "patterns" (i.e., apophenia) that violate test-taker representations of randomness. This is not a trivial danger if two conditions are true. First, if it is the case that inadvertent patterns are more likely to occur by chance than students believe, apophenia has a greater potential to hurt performance. Second, if students believe that the perceived pattern is more trustworthy than their grasp of content, they may be more likely to adopt the strategy of changing correct answers to incorrect responses in order to eliminate the offending pattern. For example, students who observe an unexpected but valid pattern (e.g., ABCDE, or AAAAA) that *seems* "non-random" may feel pressure to modify their responses to better match pattern expectancy. Consequently, correct responses will be changed to incorrect responses.

As a testament to how sensitive students may be to inadvertent patterns, consider that Carlson and Shu (2007) demonstrated a change in perception of any events that occurred more than two times in a row. In this case, the authors examined the finding that three events in a row

are perceived as unusual somehow. This is often associated with winning or losing streaks (Burns & Corpus, 2004; Vergin, 2000) and has more recently been examined within the interpretive context of the timing of such events (Sun & Wang, 2010) that people perceive "streaks" in random data (what we would call apophenia in the present article).

Carlson and Shu (2007; experiment two) presented sequences of events to participants, varying the number of outcome repetitions that occurred (2, 3, 4, and 5 times in a row). The events Carlson and Shu used were coin tosses, rolling of a six-sided die, drawing red (suited) cards from a deck of black and red suited playing cards, and drawing poker chips (red) from a collection of 120 colored chips (20 each of six different chip colors that included red). Participants rated the "streakiness" of the possible outcomes among the different events. Findings showed that participants rated outcomes of three or higher repetitions as significantly more "streaky" than events of two repetitions, regardless of the likelihood of such outcomes among the events used. In other words, although repetitions of three like outcomes are more likely among coin tosses than the tossing of a single die, participants appeared to treat the multiple events equivalently across event types. If a "streak" is perceived by students as something unusual (or unintended) taking place, then they may have reason to question their response choices on a multiple-choice exam should the same answer be used three times (or more) in a row. Presumably, based on Carlson and Shu's findings, this would be independent of the number of answer alternatives provided for each question on the exam.

In order to test the extent to which perceived violations of randomness in test answer patterns occurs and influences test takers, two studies were undertaken. The first study was designed to conceptually replicate the Carlson and Shu (2007) findings within a more exam-like context. That is, experiment one examined the degree to which students were willing to tolerate

(allow) apparent deviations from expected random patterns. In addition, random patterns (streaks) were examined among exams generated by BlackBoard™. Experiment two was designed to directly assess the effects of such deviations from perceived randomness on actual exam performance. Of some importance is whether students have the cognitive resources to spare in detecting such patterns while processing the content and meaning of exam items in a test-taking context. In which case, "patterns" of exam answers would not be problematic because they would not be attended.

Based on the research of Carlson and Shu (2007), we expected that students in experiment one would be less likely to generate response patterns exceeding three answer repetitions compared with the BlackBoard™ simulation of the same task. In other words, as Carlson and Shu found, people appear to be insensitive to the actual probabilities of the events they observe. Therefore it is unlikely that the outcome sequences generated by the students will match the outcomes generated by whatever random selection is used by Blackboard™ (which we are assuming is representative of other similar delivery platforms).

In experiment two, we predicted that students would adopt strategies of answer-changing or answer-avoiding to evade exam answer patterns that seem to violate expectations (cf. Bar-Hillel & Attali, 2002). Such a strategy would be reflected in lowered exam scores as test takers would be most likely to change correct answers to incorrect answers (or select wrong answers) in order to degrade the perceived "non-random" patterns.

Experiment 1

Method

Participants. Sixteen college-aged, predominantly white undergraduates (approximately equal numbers of women and men) from a small private university in western Pennsylvania took

part in the present study. Participants were recruited from general psychology classes and were given extra credit in their courses for volunteering. An additional 16 virtual participants were generated by BlackBoard™ (detailed below). The relatively low number of subjects was determined to best reflect the average number of students typical to on-line courses at the University (i.e., numbers that would provide a realistic sense of the likely dangers of apophenia during test-taking during realistic or common use).

Materials. A VisualBasic program was written that displayed a grid of "bubbles" similar to what would be found on a standard 100 item bubble-sheet multiple choice answer form (see Figure 1). All white dots (radio buttons) were initially blank. As a radio button was clicked (selected) for each item, a black dot appeared in the center of the dot and the immediately surrounding area was highlighted in gray to simulate paper and pencil response sheets. Only one of the five possible radio buttons per item could be highlighted at any time to prevent multiple responses.

A pool of 1000 questions was created and uploaded to BlackBoard™ to serve as a resource from which to derive random answers that would be found on a typical exam. There were 200 questions representing each of the five possible answer choices. The 16 randomly generated exams were used to represent answer patterns characteristic of the randomization features that are likely to be encountered when typical test generation programs are employed.

Design. This study utilized a 2 (Participant) x 3 (Item Repetition) mixed design in which Participant (human versus BlackBoard™) was treated as an independent groups variable and Item Repetition (two, three, four-plus) was treated as a repeated measure category. The primary dependent variable was the number of item repetitions that occurred for each participant. Additional examinations included the overall proportion of answer choices ("A" through "E")

selected as well as the number of forward ("ABCDE") or backward ("EDCBA") response sequences observed.

Procedure. After providing their consent to participate, volunteers were seated in front of a computer monitor and mouse. On the screen appeared instructions for the participant to read. The specific instructions were, "Multiple choice exams are often created using computer programs which select random questions. For each item, students must select the correct response from up to five possible choices. If students do not know the correct answer, they must guess." Beneath these instructions was a list of bullet point directives designed to emphasize the participant's task expectations. These included, (1) Whether a correct answer is placed as answer A, or B, or C, etc. is determined randomly. This study will examine how well you are able to make random guesses. (2) For the next 100 trials, you must try to match the random answers that the computer will generate. Followed by, (3) If you have any questions, please ask them now, otherwise, once the experimenter tells you to begin, you may click on the START button below.

Essentially, participants were responsible for filling in 100 "bubbles" in a fashion that they deemed to be "random". No feedback was given as to whether they guessed the same response as the computer. They merely had to provide 100 "random" responses. When participants completed the 100 trials (which took approximately 10 minutes), they were debriefed. In addition, BlackBoard™ was instructed to generate 16 exams of 100 questions and answers (randomly pulled from the 1000-item test bank) for comparison with the human data.

Results

A 2 x 3 (Participant Type [human, computer] x (Repetition Strings [2, 3, 4+]) mixed analysis of variance (ANOVA) was conducted on the mean number of answer repetitions that occurred across all participants. In this analysis, Participant Type was the independent groups

factor, while Repetition Strings was the repeated measures factor. The results indicated a significant main effect of Participant Type, $F(1,30) = 7.96$, $p < 0.01$, $\eta^2 = 0.024$, in which human participants produced fewer repetitions ($M = 3.65$, SD = 4.78) than computer participants ($M = 5.29$, SD = 5.82). The analysis also produced a significant main effect of Repetition Strings, $F(2,60) = 225.50$, $p < 0.01$, $\eta^2 = 0.759$, in which the mean number of repetitions decreased as the string length increased (i.e., there were more instances of two identical answers in a row, than three identical answers in a row, which occurred more often than four-plus answers in a row). Finally, as depicted in Table 1, the interaction effect was also significant, $F(2,60) = 7.88$, $p < 0.01$, $\eta^2 = 0.027$.

An assumption of randomness in the present study is that none of the answer options should be selected more or less often than any other option (i.e., each should be represented in the data about 20% of the time). To examine the degree to which this assumption was violated, additional analyses of the relative proportions of item responses selected across human and computer participants was also performed relative to the expected 20% proportions.

The proportions for both human and computer data are presented in Table 2. In terms of the analyses, no significant finding occurred for the BlackBoard™ generated data, $\chi^2(4, N = 16) = 0.52$, $p > .05$. That is, each of the five possible response options was selected about 20% of the time by the computer. However, the human data revealed a significant deviation from what would have been expected by chance, $\chi^2(4, N = 16) = 96.87$, $p < .01$. Consistent with Mitchell (1974), human participants tended to over-represent “B” and “C” response alternatives and under-represent both “D” and especially “E” response alternatives.

A final inspection of the data was performed to identify the number of instances in which orderly sequences (i.e., “ABCDE” or “EDCBA”) occurred between human and computer

participants. There was no instance whatsoever among the datasets produced by the computer; whereas five instances were observed among the human data. However, it seems notable that all of these instances occurred within only two participants' data (1 and 4 instances).

Discussion

The primary goal of this study was to provide participants with an opportunity to generate what they believed to be representative of a random sequence of examination answers. These procedures were expected to reflect participants' implicit beliefs as to what aspects of answer patterns were and were not likely to appear in comparison to actual randomly generated patterns. The data revealed two major points of interest in this regard. First, human participants did not behave like the electronic (BlackBoard™) participants in that electronic generated answers included a higher proportion of all repeated response strings than human generated answers. In other words, consistent with what was predicted based on the findings of Carlson and Shu (2007), and extending those findings, humans tended to generally avoid answer repetitions relative to what was generated electronically (random responses). Presumably this is because strings of like responses are noticeable as "streaks" which would seem to participants as unlikely to occur by chance alone. Second, because all repetition strings were under-represented among the human participants, their data over-represent the number of single answers (non-repetitions) compared with the electronically generated data.

Given the present findings, it seems reasonable to expect that as students become more aware of embedded repetitions ("streaks") in an exam, the more likely they may be to perceive such events as violating the expected random pattern of the exam answers. Consequently, they may be more likely to change or select answers to limit the occurrences of such streaks. If so, then exams that contain longer answer repetition strings should produce worse performance than

exams that do not contain as lengthy repetitions. However, such a prediction requires that exam-takers *notice* such repetitions.

Experiment 2

Method

Participants. Tested were 192 traditionally-aged predominantly white undergraduates (approximately equal numbers of women and men) from a small private university in western Pennsylvania. All participants were recruited from psychology classes and some earned extra credit for volunteering.

Design. The study used a simple one-way design in which the independent variable, Exam-Type, had three levels (i.e., random-pattern, short-pattern, long-pattern) and was manipulated between subjects. The dependent variable was the percent correct score earned on the exam.

Materials. A 32-item general psychology exam was developed from the test-bank of a popular psychology text (Wood, Wood, & Boyd, 2006). There were four response options for each item, and the arrangement of correct answers was such that each option occurred eight times (i.e., eight "A" answers, eight "B" answers, etc.). There were three versions of the exam. The random-pattern version resulted in no obvious underlying correct response pattern. Order of answers was determined randomly by pairing each of the 32 item response options (eight "A" responses, eight "B" responses, etc.) with a random number via the random function embedded in the Microsoft Excel program and then sorting the items according to the random number (smallest to largest). Inspection of the sorted order revealed that there were never more than two repetitions of any single response alternative. This same random order was used for all participants in this condition. The short-pattern version was organized so that correct responses

occurred as ABCD continuously throughout the exam. That is, the correct answer for items 1, 5, 9, 13, etc. was always “A”, the correct answer for items 2, 6, 10, 14, etc. was always “B”, and so forth. For the long-pattern version of the exam, correct responses were organized such that the first eight answers were all “A” while the second eight answers were all “B” and so on.

All exam questions were presented in the same order for all participants and organized to match the order of material presented in the psychology text. The only change made across exams was the order of correct response alternatives. In other words, the organization of response alternatives was the only modification made to each exam item. In addition, all response alternatives were independent of one another (there were no “all of the above,” “none of the above,” etc. options).

Procedure. The thirty-two-question tests were distributed to psychology classes during class time. All students were also given a bubble sheet upon which to record their responses (included in the exam packet). The three exam-types were distributed randomly within each of the classes in which the study was conducted.

Results

An independent groups one-way analysis of variance was conducted on mean exam scores for all three conditions. The results indicated a significant main effect of exam type, $F(2,189) = 3.29$, $p < 0.05$, $\eta^2 = 0.033$. The analysis revealed that performance improved as the underlying answer patterns became less obvious.

In hindsight, it occurred to us that students with a greater knowledge of the material to be tested would be more likely to detect possible answer patterns in the exams compared with students whose responses were less accurate. Consequently, these students could be more susceptible to handicapping due to answer changing strategies compared with less

knowledgeable test-takers. To examine this hypothesis, a post-hoc analysis of just the upper division participants (i.e., 116 psychology majors enrolled in courses more advanced than general psychology) was performed. The analysis indicated an even stronger relationship, $F(2,73) = 5.89$, $p < 0.01$, $\eta^2 = 0.139$, which tentatively confirmed the greater knowledge hypothesis. The means for each condition in both analyses are presented in Table 3.

Discussion

As predicted, performance declined as the deviation from expected patterns of randomness became more pronounced. This outcome may be contrasted with the findings of Jessell and Sullins (1975) who examined runs of 7 and 14 correct responses (all "B") embedded at various points within a 60 item exam. Although the random version of the test produced the highest average grade, Jessell and Sullins did not find statistical support for the conclusion that answer patterns affected performance. It is important to note, however, that performance within the 7 and 14 answer runs was not reported nor was it compared with the non-run portions of the exams. Because the experimental manipulation only involved either 12% or 23% of the total portion of the exam, it would not be surprising if students taking the exam never became aware of the embedded answer patterns. One or two responses that violated the answer-runs would likely be enough to mask the pattern to exam takers. This hypothesis would be supported if performance between the random and non-random portions of the exams used by Jessell and Sullins (1975) was greater for better students compared with less skilled students. That is, a negative correlation between performance in the random and non-random portions of the exam would be expected. The point here is that the better students would be more likely to notice a pattern and therefore be more susceptible to second-guessing their responses than students who did not notice the patterns (as any underlying patterns in their exams would be masked by poor

performance). To examine this possibility in the present study, upper level participants were examined separately. This examination showed that upper-level (better performing) students appeared to be particularly disadvantaged by "non-random" patterns in the answers.

Conclusions

Students demonstrate assorted heuristics during test-preparation as well as test-taking that, for some, are presumably designed to compensate for less than adequate preparation (cf. Hong, Sas, & Sas, 2006). In addition, differences in test-taking strategies have been demonstrated across academic performance levels (McClain, 1983). In fact, McClain suggested that the strategies used by higher-performing students might benefit students who tend not to perform as well on multiple-choice exams. Based on the present findings, it seems likely that higher performing students would be particularly disadvantaged compared with lower performing students by engaging in answer switching behaviors that target perceived deviations from randomness in answer patterns. This is probably because higher performing students are more likely to become aware of deviations from "randomness" which would otherwise be masked by the incorrect answers typically produced by lower performing students. In other words, answer patterns that appear to deviate from what students expect would only emerge if the responses making up these patterns were correct. A single wrong answer embedded in such a pattern might be enough for a lower performing student to remain oblivious of otherwise offending sequences. Additional research is needed to follow up on and validate this hypothesis. In particular, additional personality correlates might be expected to interact with this strategy. For example, more confident students might be willing to tolerate deviations from "randomness" compared with less confident students.

It is worth noting that the present research touches upon the literature examining people's perceptions of randomness. The general finding appears to be that the perception of randomness is better, but not much better, than the production of randomness (cf. Nickerson, 2002). In the present study, "randomness" appeared to become violated when response repetitions increased, which apparently affected students' response choices. The most common explanation for why people seem unable to behave randomly is that people are susceptible to the representativeness bias (heuristic). That is, long strings of identical outcomes appear to deviate from what seems representative of a "random" outcome. Yet, statistically, such outcomes (assuming independence, etc.) are as likely as any other outcome of equal number of trials. Apparently, then, people fail to understand and demonstrate randomness. However, as Hahn and Warren (2009) argue, and as may be particularly relevant in the present case, such an explanation may be misleading. Instead, it may be that students in the present study have a very good grasp of the likelihood of various outcomes. In general experience with exams, students probably rarely ever encounter long strings of identical responses, and indeed, such strings are to be avoided during exam construction according to Haladyna (1994). So perhaps a better explanation of the underlying response bias operating might be the availability heuristic. In which case, if long strings are noticed, students assess the likelihood of long strings of identical responses based on their experience with exams. That is, their bias may be aligned with actual probabilistic likelihoods, rather than idealized likelihoods. However, to the extent that automated exam generators become more popular, such a bias may ultimately disadvantage test-takers.

An obvious recommendation based on the present research would seem to be to equip automated test generators with filters that prevent, minimize, or compensate for the occurrence of inadvertent sequences of three or greater identical responses when they generate their items.

However, we do not believe that such an approach would truly eliminate the problem. Consider, for example, an answer key resulting from such a filter which would allow the following sequence: AABAA. Students could artificially create a long-string pattern by incorrectly selecting “A” instead of the correct response “B” in the above sequence. Such a pattern would surely trigger an apophenia-based interpretation that could result in changing more than one of the previously correct “A” responses. Instead, rather than focusing on the test-maker, we suggest a focus on the test-taker. Specifically, we recommend an education based solution in which students are (1) alerted to the inaccuracy of their beliefs regarding randomness as well as (2) strong admonishment to focus on relevant exam taking strategies (content knowledge preparation) rather than superficial strategies such as perceived patterns, answer counting, etc. that should normally, and under most circumstances, be irrelevant to exam success.

References

Attali, Y., & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement, 40*(2), 109-128.

Bar-Hillel, M., & Attali, Y. (2002). Seek whence: Answer sequences and their consequences in key-balanced multiple-choice tests. *The American Statistician, 56*(4), 299-303.

Brent, C. (2010). Online education in community colleges. *New Directions for Community Colleges, 150*, 7-16.

Burns, B. D., & Corpus, B. (2004). Randomness and inductions from streaks: "Gambler's fallacy" versus "hot hand". Psychonomic Bulletin & Review, 11(1), 179-184.

Carlson, K. A. & Shu, S. B. (2007). The rule of three: How the third event signals the emergence of a streak. *Organizational Behavior and Human Decision Processes, 104*(1), 113-121.

Carroll, R. T. (2003). *The skeptic's dictionary: A collection of strange beliefs, amusing deceptions, and dangerous delusions*. Hoboken, N.J: Wiley.

Chiesl, N. (2007). Pragmatic methods to reduce dishonesty in web-based courses. *The Quarterly Review of Distance Education, 8*(3), 203-211.

Hahn, U. & Warren, P. A. (2009). Perceptions of randomness: Why three heads are better than four. *Psychological Review, 116*(2), 454-461.

Haladyna, T. M. (1994). *Developing and Validating Multiple-Choice Test Items*. Hillsdale, NJ: Erlbaum.

Haladyna, T. M. & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*(1), 37-50.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-334.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology, 56*(4), 208-216.

Hogan, T. P., & Murphy, G. (2007). Recommendations for preparing and scoring constructed-response items: What the experts say. *Applied Measurement in Education, 20*(4), 427-441.

Hong, E., Sas, M., & Sas, J. C. (2006). Test-taking strategies of high and low mathematics achievers. *The Journal of Educational Research, 99*(3), 144-155.

Hopkins, K. D., & Hopkins, B. R. (1964). Intraindividual and interindividual positional preference response styles in ability tests. *Educational and Psychological Measurement*, *24*(4), 801-805.

Jessell, J. C, & Sullins, W. L. (1975). The effect of keyed response sequencing of multiple choice items on performance and reliability. *Journal of Educational Measurement, 12*(1), 45-48.

McClain, L. (1983). Behavior during examinations: A comparison of "A," "C," and "F" students. *Teaching of Psychology, 10*(2), 69-71.

Mitchell, W. E. (1974). Bias in writing objective-type examination questions. *The Journal of Economic Education, 6*(1), 58-60.

Nickerson, R. S. (2002). The production and perception of randomness. *Psychological Review, 109*(2), 330-357.

Sun, Y., & Wang, H. (2010). Perception of randomness: On the time of streaks. *Cognitive Psychology, 61*(4), 333-342.

Vergin, R. C. (2000). Winning streaks in sports and the misperception of momentum. *Journal of Sport Behavior, 23*(2), 181-197.

Wasserman, E., Young, M., & Cook, R. (2004). Variability discrimination in humans and animals: Implications for adaptive action. *American Psychologist, 59*(9), 879-890.

Wevrick, L. (1962). Response set in a multiple-choice test. *Educational and Psychological Measurement, 22*(3), 533-538.

Wood, S. E., Wood, E. G., & Boyd, D. (2008). *Mastering the World of Psychology*, 3rd ed. Boston, MA: Allyn and Bacon.

Table 1

Mean number of occurrences of 2, 3, and 4+ item repetition strings by type of participant.

	Human			Computer		
Repetition Strings	N	Mean	SD	N	Mean	SD
2 (e.g., AA)	16	8.94	1.19	16	13.00	2.75
3 (e.g., AAA)	16	1.63	0.42	16	2.44	0.89
4+ (e.g., AAAA)	16	0.38	0.72	16	0.44	0.63

Table 2

Relative proportion of item responses by participant type (assuming 0.20 = unbiased selection).

	Human			BlackBoard™		
Response Option	N	Mean	SD	N	Mean	SD
"A"	16	0.204	0.035	16	0.205	0.048
"B"	16	0.236	0.041	16	0.196	0.043
"C"	16	0.263	0.057	16	0.201	0.051
"D"	16	0.176	0.033	16	0.196	0.037
"E"	16	0.121	0.056	16	0.203	0.033

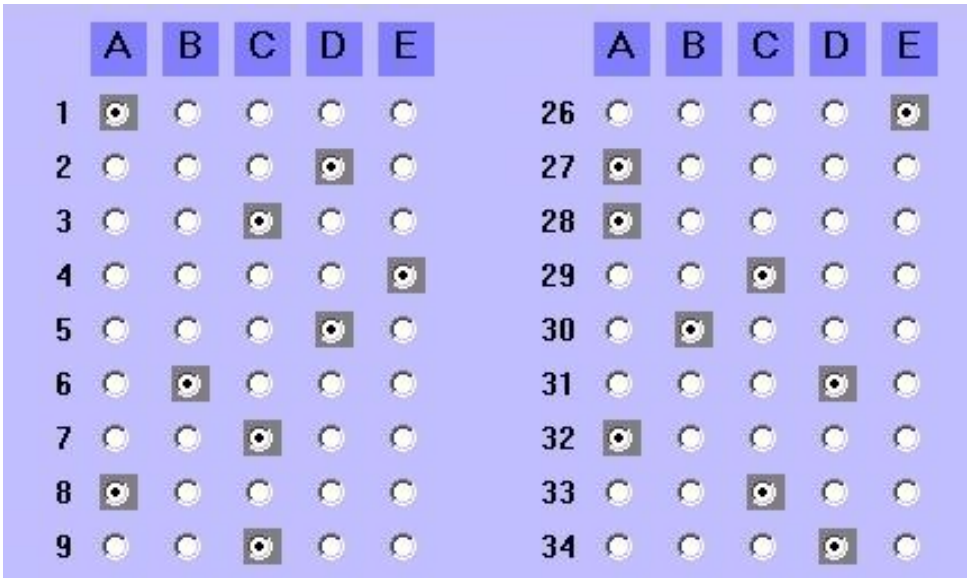
Table 3

Effect of exam type (all participants vs. upper-level participants).

	All Participants			Upper-Level Participants		
Condition	N	Mean	SD	N	Mean	SD
Long Pattern (AAAA)	64	41.9	11.0	25	44.5	7.07
Short Pattern (ABCD)	64	45.9	12.9	26	49.4	9.51
Random Pattern (CBDA)	64	47.1	12.6	25	52.8	8.80

Figure Caption

Figure 1. Partial screenshot of layout used for experiment one. Participants were presented with four columns of numbered (1-100) “ABCDE” response options, of which a portion of the first two are seen here. Clicking on a white dot resulted in the dot being filled with a black dot and highlighted in gray to simulate the pencil and paper versions of multiple choice answer sheets (bubble sheets). No dots were filled when the participants first viewed the computer form, and participants could change answers for any item simply by clicking on another (empty) white dot.



	A	B	C	D	E		A	B	C	D	E
1	●	○	○	○	○	26	○	○	○	○	●
2	○	○	○	●	○	27	●	○	○	○	○
3	○	○	●	○	○	28	●	○	○	○	○
4	○	○	○	○	●	29	○	○	●	○	○
5	○	○	○	●	○	30	○	●	○	○	○
6	○	●	○	○	○	31	○	○	○	●	○
7	○	○	●	○	○	32	●	○	○	○	○
8	●	○	○	○	○	33	○	○	●	○	○
9	○	○	●	○	○	34	○	○	○	●	○