

# **Innovative Applications of Classroom Response Systems: Investigating Students' Item Response Times in Relation to Final Course Grade, Gender, General Point Average, and High School ACT Scores**

**Wilson J. González-Espada & Daniel W. Bullock**  
*Arkansas Tech University*

## **ABSTRACT**

With the introduction of classroom response systems (CRS) in physics classrooms, instructors are now able to examine assessment parameters that are commonly described in the standardized assessment literature but were not previously available. The purpose of this research was to examine the relationship between students' item response time for answering multiple-choice questions posed in an Introduction to Physical Science course offered during the summer of 2004 and variables such as gender, grade point average, ACT scores, and final grades. It was found that (a) for about 1/3 of the participants, there was a significant difference between the average response time for items answered correctly and incorrectly, (b) there was a significant negative correlation between the average response time for items answered correctly and both ACT scores and final score in Introduction to Physical Science, even if item difficulty is accounted for, and (c) there was no significant correlation between the average response time for items answered incorrectly and our variables of interest.

## **INTRODUCTION**

All instructors are involved in three basic but essential tasks: planning, delivery, and evaluation of learning. Each of these tasks must be performed efficiently and effectively. The evaluation of learning, defined as the process of making a value judgment about the worth of a student's product or performance (Nitko, 2004), is especially important since the information instructors gather informs all involved parties about how well the student understands the material presented, which in turn has implications for the student's final grade in the class.

In many college physics and physical science courses, instructors evaluate students with tests or quizzes designed to provide a score proportional to the student's capacity to solve conceptual questions and quantitative problems. Needless to say, those tests and quizzes must be valid, that is, they must measure what they are intended to measure. Test reliability is also important. A test is reliable if on an equivalent test or quiz, a student obtains about the same score. Valid and reliable tests are needed for the student's scores to be both accurate and pedagogically meaningful (Crocker and Algina, 1986).

The validity of instructor-made tests is very important. Instructors must design assessment items that allow students to demonstrate their knowledge of the subject matter covered in class. At the same time, they must decide what type of test they will use (Gulliksen, 1987; Schmitt, 1991):

- Power tests – This type of test combines items of varying levels of difficulty. Students can take as much time as they need to finish a power test. The goal of this type of test is to measure how accurately the student can answer the test.
- Speeded tests – This type of test includes low and medium difficulty items. Students must finish the test in a strict predetermined amount of time. The goal of this type of test is to measure how fast a student responds to it. For speeded tests, item response time (speediness) is an important variable.

In many cases, physics and physical science class tests tend to be partially speeded. Students know that they have a certain amount of time to finish the test but the actual finishing time can be somewhat flexible.

The disadvantage of speeded and partially speeded tests is that the score is not exclusively measuring how accurately the student answers the test, but how accurately the student answers the test *given the additional pressure of a time constraint* (Nitko, 2004). These two factors, response accuracy and response speed, do not measure the same construct and provide separate measures of performance. Research suggests that the validity of a test diminishes when additional variables, such as time constraints, have the possibility of affecting student performance (Scrams and Schnipke, 1997). In order to improve test quality and better assess students, physics instructors must become familiar with the characteristics and limitations of power and speeded tests and their validity and reliability implications.

So far, researchers' study of item response time and its relationship with test scores and test validity has been limited to large-scale standardized tests. Recently, however, educators who have long understood the potential of technology in supplementing the instructional process are implementing new ways of assessing students (Wall, 2000).

During the last decade, inexpensive classroom response systems (CRS) became available for college instructors (Draper et al., 2001). Physics instructors and researchers praise such technologies for their immediate feedback, increased student attention and engagement during lectures, enhanced peer interaction, and positive effect on student performance, especially for large enrollment sections (Beatty, 2004; Bullock et al., 2002; Burnstein and Lederman, 2001; Milner-Bolotin, 2004; Rebello, 2002).

With the advent of computerized testing and other electronic response systems, test and item response times that instructors found very difficult to record during traditional paper and pencil tests, can now be easily and unobtrusively recorded. This new information provides additional knowledge about examinee behavior and increase the validity of tests. In fact, item response time research is growing in popularity along with computerized and electronic test administrations (Schnipke and Scrams, 1999).

This research explores the relationship between the response times of items used as reinforcement and review during lecture, and variables such as gender, grade point average (GPA), America College Test (ACT) scores, and final course grade among students who took an introduction to physical science course during the summer of 2004. More specifically, this study was guided by the following research questions:

- Is there a significant difference between the average response time for items answered correctly and incorrectly?
- Is there a significant correlation between the average response time for items answered *correctly* and variables such as GPA, ACT score, math ACT subscore, gender, and final score in Introduction to Physical Science?
- Is there a significant correlation between the average response time for items answered *incorrectly* and variables such as GPA, ACT score, math ACT subscore, gender, and final score in Introduction to Physical Science?
- Is there a significant correlation between the average response time for items answered correctly *when classified by item difficulty level* and variables such as GPA, ACT score, math ACT subscore, gender, and final score in Introduction to Physical Science?

The information obtained from this and similar research studies has implications in at least three important areas. First, many physics instructors use their experience to “guesstimate” what kinds and how many questions should be included in a period-long test (50 minutes or 90 minutes are usual). An incorrect estimation could have the effect of converting a test to measure accuracy into a test to measure response time for some students, with potentially biasing effects on those students’ test scores (Schnipke and Scrams, 1999). Educators’ understanding item response time and test validity in the context of the local student population should help reduce such bias.

Second, students can be better informed about test-taking strategies on power versus speeded tests. For example, research suggest that on speeded tests, students engage in rapid-guessing behaviors, devoting considerable time to questions at the beginning of the test compared to questions at the end of the test. The analysis of response times allows the instructor to educate students about how to pace themselves better, maintain a constant speed throughout an examination, and allocate time according to the questions’ difficulty levels.

Third, on a cognitive level, when instructors understand item response theory, the results provide the instructor with information about how students processed the item, item difficulty, and even reading comprehension of word problems. Additional research could address individual and group relationships between response time and response accuracy (Schnipke and Scrams, 1999).

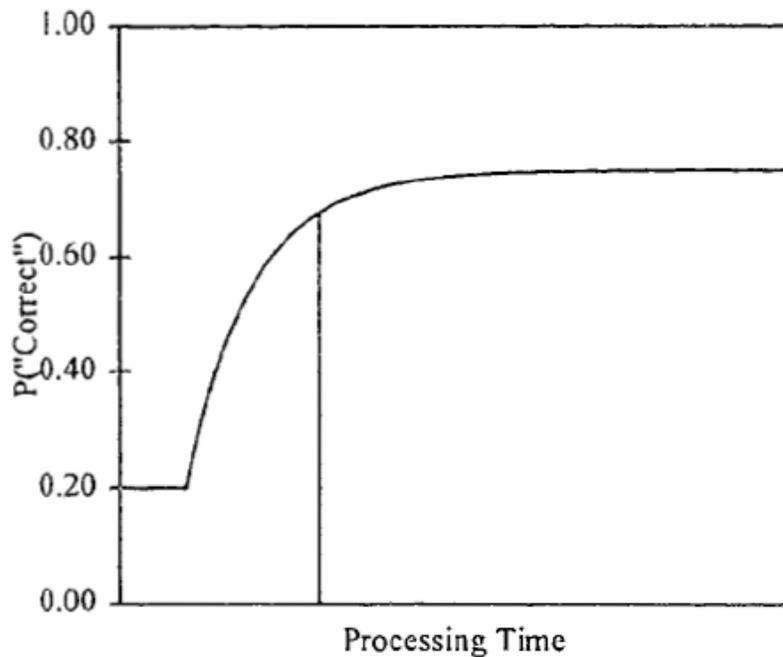
### **Item Response Theory**

According to Item Response Theory or IRT (Baker and Kim, 2004; Hambleton et al., 1991; Sijisma and Molenaar, 2002), the test taker’s knowledge of the subject matter should determine the probability of answering a test item correctly. Instructors should control other parameters as much as possible to reduce bias and increase the validity of the test. Researchers study item response time in large-scale standardized assessment because it is a content-independent factor that violates IRT’s basic assumption by increasing the chances of item bias, defined as “the extent to which there are different probabilities of correct response to an item for examinees of the same ability level” (Douglass, 1981). Moreover, test items that are harder for an individual or group than for another individual or group with the same level of ability are defined as differential items (Schmitt, 1991).

Students working on two simultaneous tasks (response time and response accuracy) on speeded and partially speeded tests have guided researchers such as Thissen (1983) to study statistical models of how students answer these test items. An application of IRT is the speed-accuracy tradeoff function (Scrams and Schnipke, 1997). This function has a slightly different graphical representation for each item and each test-taker, although in general it rests on the following assumptions (see figure 1):

- Examinees require some minimum amount of time in order to respond above levels associated with guessing.
- Performance increases rapidly toward an upper asymptote, representing the test taker’s performance given an infinite amount of time.
- Once a test-taker has approached asymptotic accuracy, further processing time would have little effect on performance.
- In some circumstances (before the student reaches asymptotic accuracy), further processing time has a significant effect on item performance.

**Figure 1**



*Note for Figure 1: Theoretical function relating accuracy to processing time for a particular examinee responding to a particular five-choice item. Accuracy is an increasing function of processing time with a lower asymptote (chance) and an upper asymptote. The vertical line represents the examinee's chosen processing time and the resulting accuracy (Schnipke and Scrams, 1999).*

### **Item Response Time**

Researchers have identified several variables that might be related to item response time. Racial background seems to be a factor that influences how quickly a student completes a series of test items. For example, Schmitt and Bleistein (1987) and Douglass (1981) found lower test scores partly attributable to item response time for African Americans compared with Caucasian students. Schmitt (1991) analyzed Scholastic Assessment Test (SAT) results from about 300,000 test takers and found that African American and Hispanic students have significantly different item response times compared with Caucasian students. Similar results were found for the Graduate Management Admission Test or GMAT (Lawrence, 1993).

The anxiety level of the test taker also affects item response time. According to Bergstrom, Gershon and Lunz (1994), anxiety was significantly related to test response time. Physics students can be made anxious not only by the act of test taking but also by the mathematical and conceptual difficulty of the test, which affects the students' this final grade in the class.

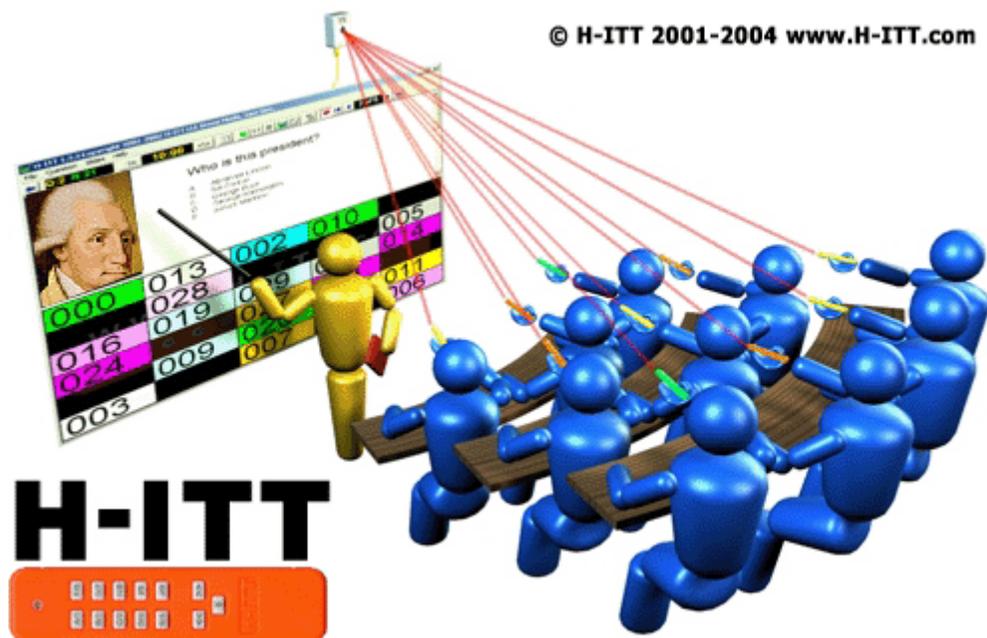
Interestingly, gender seems not to be an important factor in determining how quickly a student completes an assessment in comparison with test scores. Cole (1997) completed a meta-study of 400 tests involving millions of students and found no statistically significant relationship between item response times and test scores for males and females. Schmitt (1991) did not find a relationship between these variables either.

## METHODS

The authors carried this study out over a five-week summer session in an introductory physical science course for non-science majors. Introduction to Physical Science (PHSC 1013) is a three credit-hour course that serves as a general education physical science requirement for most majors at Arkansas Tech University.

The lecture portion of the study was performed in an auditorium with a seating capacity of 65 students. To administer the mini-quizzes, the instructor displayed one multiple-choice question at a time using an overhead projector. Once a question was visible, a timer started automatically in the computer. The instructor instructed the students to answer the question as quickly as they could, making each item a partially speeded one. Three points were awarded for a correct answer, one point for an incorrect answer, and zero for no answer. The two main sources of questions for the mini-quizzes were the test bank that comes with the textbook (Shipman, Wilson, and Todd, 2006) and teacher-made questions.

*Figure 2*



*Note for Figure 2: Hyper-Interactive Teaching Technology (H-ITT) system. Students use transmitters to answer instructor's questions. A receiver collects the responses and sends them to the software for analysis and display. Diagram used by the author with permission.*

The classroom response system used in instruction and data collection is commercially known as “Hyper-Interactive Teaching Technology” or H-ITT. This system was designed to allow instructors to depart from the traditional lecture format, considered by many as pedagogically ineffective in achieving conceptual understanding and challenging misconceptions (McDermott, 1993; McDermott and Redish, 1999; Redish, 1994; Redish and Steinberg, 1999; Thacker, Kim and Trefz, 1994). Instructors promoting a high level of interaction during class

using classroom response technologies such as H-ITT should result in improved attendance and understanding of the physical science content presented.

H-ITT has three components: a transmitter, a receiver, and the software that processes the students' responses (see Figure 2). Each student has a transmitter with a unique ID number and frequency. This transmitter has 10 main keys, each key marked with one letter and one number (A/1, B/2, C/3 ... J/9). Using the transmitter, the students answer verbal or visual questions provided by the instruction, usually in multiple-choice or true-false format. The classroom is wired with two receivers which collect the students' responses and send them to the computer for further processing. The H-ITT software will tally the responses for grading purposes, displays them as a histogram on a screen, and save them for further analysis. This immediate feedback helps the instructor to pace classroom instruction, clarify material that was not well understood, and avoid time-consuming manual grading of quizzes.

In order to encourage participation and ensure accurate data collection, the instructor established a one-to-one correspondence between each transmitter's ID and the student's identification number. Throughout the summer session, students were warned that having more than one transmitter constituted cheating, and headcounts were randomly performed to confirm that they corresponded to the same number of recorded transmitter responses.

With the goal of fostering peer instruction, the instructor gave special attention to the manner in which in-class questions were asked. On occasion, the students were asked a multiple choice question and required to come up with an answer on their own. After the student responses were received, the teacher displayed the histogram. Depending on the result, the teacher would either move on to the next topic or ask the students to reconsider their answer and try to convince their neighbors to agree with their response. Used in this manner, the histogram would prompt peer-instruction. This algorithm of student-to-student teaching, or peer instruction, has been shown to be an effective educational tool in introductory physics (Mazur, 1997) and astronomy (Green, 2003).

Given the limited sample size available ( $n = 27$  students), the researchers were interested in identifying possible relationships between the variables in a broad sense. As a consequence, the significance level for all statistical tests was established at 0.10, that is, the probability of committing a type I error (reporting a relationship between two variables that does not exist) is set at 10%. A Bonferroni correction (Shaffer, 1995) to the alpha level will not be applied to this study because, although there is a probability of chance capitalization, the results of the study might not be conclusive with the sample size available and that significant relationships would have to be confirmed in further studies. In addition, by applying a Bonferroni correction we would increase the chance of making a type II error (overlooking an existing relationship between variables).

## **RESULTS AND DISCUSSION**

### **Demographics**

In terms of gender, there were 10 males and 17 females in this study. 39% of the participants had an overall ACT score of 18 or less, considered remedial in our institution, compared with 61% who had an overall score of 19 or more. The average ACT score was 19.6 points. Their scores in the mathematics section of the ACT were significantly lower than their overall score, with 72% of the participants with sub-scores of 18 or less compared with 28% with sub-scores of 19 or more. The average math ACT sub-score was 17.5 points.

The participants' performance in the class was relatively good, with only one failing grade and one excellent grade (A). There were 11 very good grades (B), 10 satisfactory grades (C) and four deficient grades (D). The average class score was 78.1%. This can be compared with their overall grade point average at the end of the summer session. In general, four students had a GPA of more than 3.50, four students had a GPA between 3.00 and less than 3.50, eight students had a GPA between 2.50 and less than 3.00, eight students had a GPA between 2.00 and less than 2.50, and only three students had a GPA of less than 2.00.

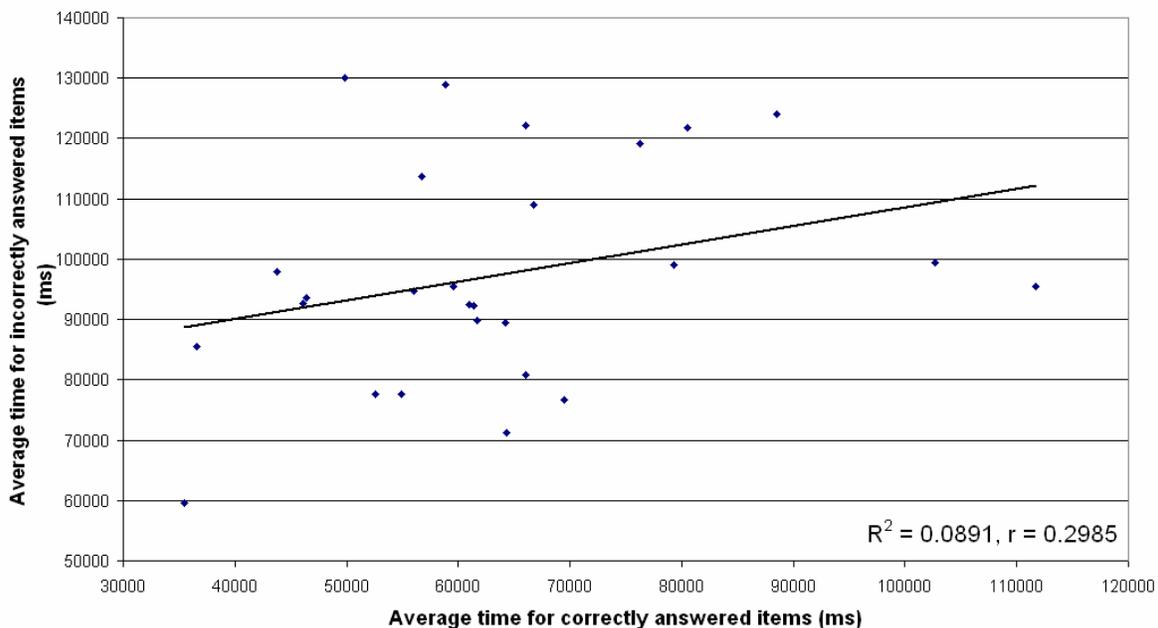
Of all the questions presenting during lecture (n = 74 questions), seven students answered 80% or more correctly, nine students answered between less than 80% and 75% correctly, six students answered between less than 75% and 70% correctly, and five students answered less than 70% correctly. The lowest percentage of correctly answered questions was 60%.

### Item Response Time Correlations

Students, as a group, required on average 63.365 seconds to answer questions correctly, compared with 97.682 seconds for incorrectly answered items. In almost all cases, students took less time to answer correctly (see Figure 3). A positive trend between these variable, although not significant ( $p = 0.1304$ ) suggests that processing time might be consistent between students, that is, faster or slower responders do so regardless of response correctness.

Although a trend between item response time and correctness is apparent, an additional test would establish whether there is a statistically significant difference between the correct and incorrect response times. A number of t-tests were performed for this purpose. It was found that for 33% of the students the difference in response times was statistically significant at an alpha level of 0.10.

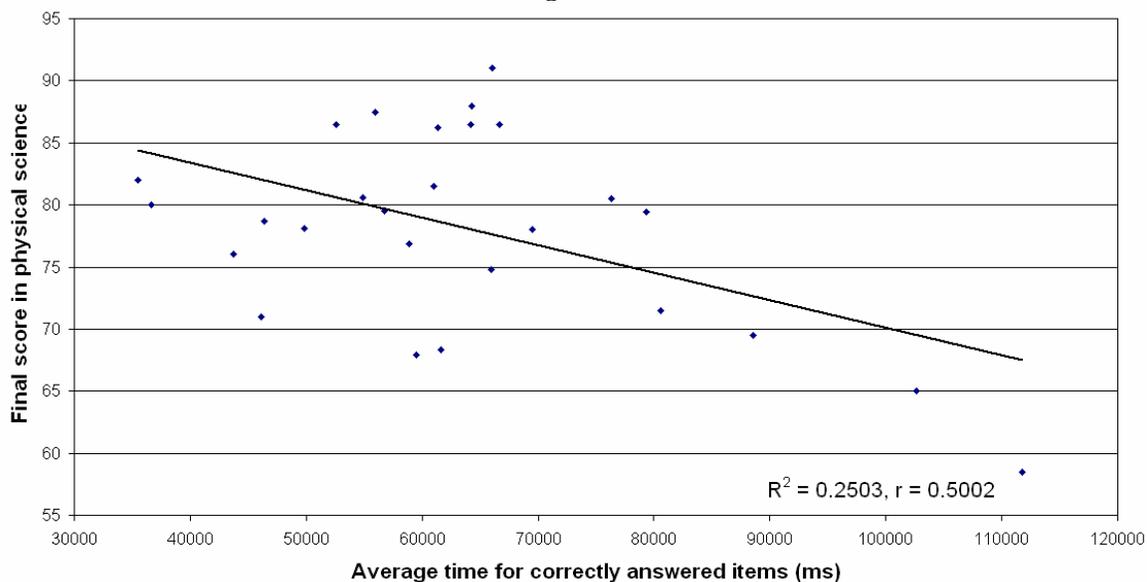
*Figure 3*



*Note for Figure 3: Correlation between the students' average response time for correctly answered item and the average response time for incorrectly answered items.*

The authors performed several correlation tests between the average response times for correctly answered items, the average response time for incorrectly answered items, and independent variables such as final score in the class, GPA, ACT score, math ACT score, and gender. It was found that the average response time for correctly answered question items was negatively correlated with the students' final score in the course ( $r = -0.5002$ ,  $p = 0.008$ ) (See Figure 4). The distribution of average response times and final scores in the class is even more interesting when gender is examined. The correlation between these variables appears to be much stronger for males ( $r = -0.5871$ ,  $p = 0.077$ ) than for females ( $r = -0.3488$ ,  $p = 0.1859$ ). However, examining gender separately reduces the sample size available for the tests, and consequently their power.

**Figure 4**

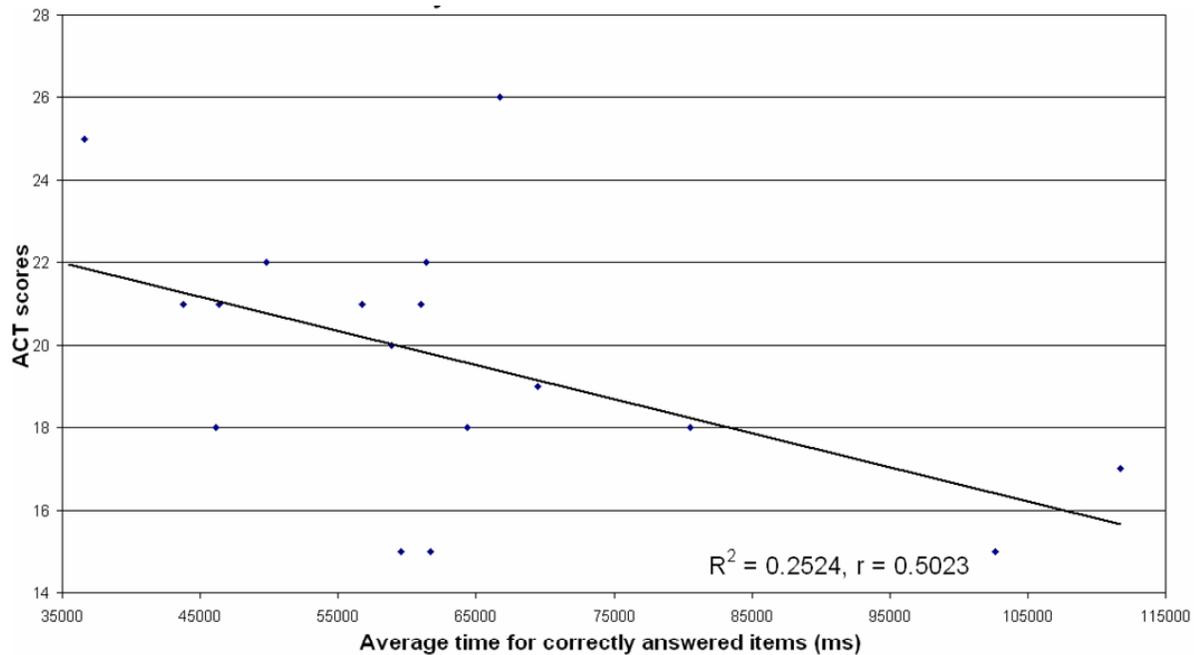


*Note for Figure 4: Correlation between the students' average response time for correctly answered item and their final score in physical science.*

The average response time for correctly answered items is negatively correlated with the students' ACT scores ( $r = -0.5023$ ,  $p = 0.041$ ). Since ACT scores and final grades in Introduction to Physical Science are highly correlated with each other for the sample ( $r = 0.718$ ,  $p = 0.0015$ ), these results reinforces the notion that response time might be related to the processing speed and/or test-taking technique, which is reflected in some measures of achievement (see Figure 5).

Additional tests were performed to examine whether the average response time for incorrectly answered items are correlated to any of our independent variables. We found no significant correlation between the response times and the variables of interest. Consequently, there might be other factors that influence the students' decision in taking longer to answer a question other than our variables of interest.

*Figure 5*



*Note for Figure 5: Correlation between the students' average response time for correctly answered items and their ACT scores.*

In order to provide a better insight into the nature of item response time and the independent variables, we decided to divide the in-class questions into two groups based on difficulty level. To achieve this, the item difficulty for each question, defined as the ratio between the total number of correct responses and the total number of students who attempted the question (Airasian, 2001; Ebel and Frisbie, 1986; Gronlund, 1968), was calculated. Questions were divided into two equal groups and the average response time for correctly answered items was recalculated. It is recognized that analyses based on these response times will be less powerful. However, having an idea of whether the question's difficulty level might be related to our variables makes the analyses informative, although not conclusive.

It was found that the average response time for relatively easy questions was significantly correlated with the average response time for relatively hard questions ( $r = 0.5732, p = 0.002$ ). This result suggests that students who respond relatively quickly tend to do so regardless of the question's difficulty level and students who take longer to respond, do so regardless of how easy the questions are. It is conceivable to conclude that the processing time is a trait that might depend on the students' knowledge and decoding ability more than the difficulty of the task.

In comparing the students' grade point average with the newly calculated response times, it was found that there was no correlation between the average response times for relatively easy questions and GPA ( $r = 0.019; p = 0.9264$ ). Interestingly, a statistically significant relation was found between the average response times for relatively hard questions and GPA ( $r = -0.4012; p = 0.0383$ ). This result, consistent with the literature on assessment, suggests that relatively difficult items are better at discriminating students' ability and knowledge compared with easy items (Airasian, 2001; Ebel and Frisbie, 1986).

In addition, statistically significant correlations were found between the students' final score in the course and both the average response times for relatively easy questions

( $r = -0.3948$ ;  $p = 0.0396$ ) and the average response times for relatively hard questions ( $r = -0.4802$ ;  $p = 0.012$ ). Also, statistically significant correlations were found between the students' ACT scores and both the average response times for relatively easy questions ( $r = -0.5147$ ;  $p = 0.0353$ ) and the average response times for relatively hard questions ( $r = -0.4459$ ;  $p = 0.0735$ ), not a surprising result given our more general analysis.

## LIMITATIONS AND FUTURE RESEARCH

This study establishes the possibility of a relationship between response times and final score in the class, GPA, ACT score, math ACT score, and/or gender. This was done, to a certain degree, given a relatively small sample size (27 students and 74 questions) and the selection of a very general statistical tool.

Given the limited nature of this study, a number of questions were left unanswered. First, it is important to determine whether the recently discovered relationships between response time and both ACT score and final score in Introduction to Physical Science are present in future replications of this study with a much larger sample populations, possibly several sections of Introduction to Physical Science over the course of at least two semesters. Needless to say, the in-class items used for all sections and both semesters must be identical for a comprehensive analysis. Investigating item response time during regular semesters, as opposed to summer sessions, might also provide a more representative sample given the fact that the chances of having a self-selected group during the summer is probably higher compared with enrollment in regular semester-long courses.

Second, the relationship between response times and other variables can be further elaborated by recording data for different levels of response times (very fast, fast, average, slow, very slow). The researchers consider the current data as partially speeded: students were instructed to answer in the least time possible, but there was not a specific time limit imposed per item. By assigning such time limits and varying them for different groups and different semesters, it would be possible to better determine to what extent item response time and response accuracy are related.

Third, for the purpose of this study, the authors calculated the students' average response times for all items and for two data subgroups based on item difficulty. Dividing the items into two groups is a rather crude way of examining item difficulty in relationship with other variables of interest. A better approach for future research is to use item difficulty coefficients (Haladyna, 1994) for individual in-class questions as a variable to be examined in more detail.

Fourth, very little was said in this study about the composition of the items. How many words did each item have? What was the cognitive level of each item according to Bloom's (1956) taxonomy? Were there more quantitative or qualitative items? Were quantitative problems contextual or "plug-and chug"? This was done purposefully because it was recognized that these variables would be impossible to study with such a small sample size. Future research should surely address item characteristics and how they are related to student performance and response time, in addition to item type (multiple choice, true or false, other).

With a larger sample size, researchers can apply more complex and powerful tests, like multiple regression analysis, to a more specific and diversified number of variables. These tests will determine more precisely the nature of the relationships and interrelationships between variables and will be pursued by the authors in the near future.

## CONCLUSION

This study answered the following research questions: (a) For about 1/3 of the participants, there was a significant difference between the average response time for items answered correctly and incorrectly, (b) there was a significant negative correlation between the average response time for items answered correctly and both ACT scores and final score in Introduction to Physical Science, even if item difficulty is accounted for, (c) there was no significant correlation between the average response time for items answered incorrectly and our variables of interest, and (d) there were several correlations between relatively hard items and some of the variables of interest, but these are not as conclusive due to a reduction in question sample size during the analysis.

The ability of physics and physical science teachers who use CRS to collect data on student item response times opens an innovative and promising area of research in science education assessment by the application of quantitative educational assessment theories, such as item response theory, in the examination of test validity, item bias, student knowledge, construct-independent examinee characteristics, and other instrument-related variables.

### Contributors

*Dr. Wilson J. González-Espada is an Associate Professor of Physical Science/Science Education at Arkansas Tech University. He teaches courses in physical science, physics, multicultural science education, and science methods. Originally from Puerto Rico, he completed undergraduate studies at the University of Puerto Rico and graduate studies at the Interamerican University of Puerto Rico and The University of Georgia. Dr. González-Espada has contributed extensively to professional journals, magazines, newspapers, and professional meetings in the areas of physics education and multiculturalism.*

*Dr. Daniel W. Bullock is an Assistant Professor of Physics at Arkansas Tech University. He teaches courses in both physical science and physics. He completed his undergraduate studies at Arkansas Tech University and graduate studies at the University of Arkansas. Dr. Bullock has published extensively in professional journals in the fields of nanoscience and, more recently, physics education.*

### References

- Airasian, P. W. (2001). *Classroom assessment: Concepts and applications*. Boston, MA: McGraw Hill.
- Baker, F. B., and Kim, S. (2004). *Item response theory: Parameter estimation techniques*. Marcel Dekker, New York.
- Beatty, I. (2004). *Transforming student learning with classroom communication systems*. UMPERG/Research Bulletin of the EDUCAUSE Center for Applied Research, 2004, issue 3.
- Bergstrom, B., Gershon, R., and Lunz, M. E. (1994). *Computer adaptive testing: Exploring examinee response time using hierarchical linear modeling*. Paper presented at the 1994 meeting of the National Council on Measurement in Education, New Orleans, LA.
- Bloom, B. S. (1956). *Taxonomy of educational objectives: The cognitive domain*. Longman, London.
- Bullock, D. W., LaBella, V. P., Clingan, T., Ding, Z., Stewart, G., and Thibado, P. B. (2002). Enhancing the student-instructor interaction frequency. *The Physics Teacher*, 40, 535-541.
- Burnstein, R. A., and Lederman, L. M. (2001). Using wireless keypads in lecture classes. *The Physics Teacher*, 39, 8-11.

- Cole, N. S. (1997). *The ETS gender study: How females and males perform in educational settings*. Educational Testing Service, New Jersey. ERIC document #424337.
- Crocker, L., and Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Harcourt Brace Jovanovich.
- Douglass, J. B. (1981). *Item bias, test speededness, and Rasch tests of fit*. Paper presented at the 1981 meeting of the American Educational Research Association, Los Angeles, CA. ERIC document #200632.
- Draper, S., Cargill, J., and Cutts, Q. (2001). *Electronically enhanced classroom interaction*. Proceedings of the 2001 Conference of the Australasian Society for Computers in Learning in Tertiary Education, Melbourne, Australia. ERIC document #467932.
- Ebel, R. L. and Frisbie, D. A. (1986). *Essentials of educational measurement, 4<sup>th</sup> ed.* Englewood Cliffs, NJ: Prentice Hall.
- Green, P. J. (2003). *Peer instruction for astronomy*. Upper Saddle River, NJ: Pearson Education.
- Gronlund, N. E. (1968). *Readings in measurement and education*. London, UK: Collier-Macmillan Ltd.
- Gulliksen, H. (1987). *Theory of mental tests*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Haladyna, T. M. (1994). *Developing and validating multiple-choice test items*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hambleton, R. K., Swaminathan, H., and Rogers H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.
- Lawrence, I. M. (1993). *The effect of test speededness on subgroup performance*. Educational Testing Service, New Jersey. ERIC document #386494.
- Mazur, E. (1997). *Peer instruction: A user's manual*. Upper Saddle River, NJ: Prentice Hall.
- Milner-Bolotin, M. (2004). Tips for using a Peer Response System in a large introductory physics class. *The Physics Teacher*, 42, 47-48.
- Nitko, A. J. (2004). *Educational assessment of students, 4<sup>th</sup> edition*. Upper Saddle River, NJ: Merrill.
- Rebello, N. S. (2002). *Using K-State online and in-class personal response system in a large enrollment physical science class*. Paper presented at the 2002 Convergence of Digital Learning Conference, Kansas State University, Manhattan, KS.
- Schmitt, A. P. (1991). *Differential speededness and item omit patterns on the SAT*. Paper presented at the 1991 Meeting of the National Council on Measurement in Education, Boston, MA. ERIC document #389716.
- Schmitt, A. P., and Bleistein, C. A. (1987). *Factors affecting differential item functioning for black examinees on SAT analogy items*. Educational Testing Service, New Jersey.
- Schnipke, D. L., and Scrams, D. J. (1999). *Exploring issues of test taker behavior: Insights gained from response-time analysis of the LSAT*. Law School Admission Council, New Jersey, 1999. ERIC document #467811.
- Scrams, D. J., and Schnipke, D. L. (1997). *Making use of response times in standardized tests: Are accuracy and speed measuring the same thing?* Paper presented at the 1997 meeting of the American Educational Research Association, Chicago, IL. ERIC document #409357.
- Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, 46, 561-584.
- Shipman, J. T., Wilson, J. D., and Todd, A. W. (2006). *Introduction to Physical Science, 11<sup>th</sup> edition*. Boston, MA: Houghton Mifflin Co.
- Sijtsma, K., and Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: SAGE Publications.
- Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss and R. D. Bock (eds.) *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York, NY: Academic Press, pp. 179-203.
- Wall, J. E. (2000). *Technology-delivered assessment: Diamonds or rocks?* Washington, DC: Office of Educational Research and Improvement. ERIC document #446325.