# ENTERING EQUATIONS: COMPARISON OF HANDWRITING RECOGNITION AND EQUATION EDITORS

Gabrielle A. Cayton-Hodges
Educational Testing Service
gcayton-hodges@ets.org

James Fife
Educational Testing Service
jfife@ets.org

*Once a novelty, Digitally-Based Assessments (DBA) have become commonplace in the USA. With mathematics, it is often a necessity to include items that require the student to input a mathematical formula, equation, or expression. Many of these responses, especially in the upper grades, cannot be input with a standard keyboard, but must use some type of equation entry. In this study, we compare ninth-graders' entry of mathematical expressions using an equation editor versus using handwriting recognition on a tablet. While neither method is currently without flaws, we discuss the benefits and drawbacks of each as well as potential methods for improvement and the implications for mathematics assessment.*

Keywords: Assessment and Evaluation, Technology

## Introduction

Mathematics has long been a subject of paper and pencils. Scratch work, diagrams, and mark-ups are all parts of solving mathematical problems that need to be addressed as part of developing Digitally-Based Assessments (DBA). One artifact of the paper-and-pencil world is that of the mathematical response itself. Most mathematical formulae, equations, and expressions cannot be input properly using a standard keyboard. Even at the elementary level, a simple fraction requires an equation editor, as most students are used to seeing the numerator directly above the denominator ($\frac{1}{2}$), not the orientation that would result from a sideways slash (1/2). At the middle-school level, exponents, square roots, and $\pi$ further complicate the mix. Finally, at the high-school level, all of these are combined together in various embedded formats that can confuse even those who are comfortable with the individual symbols.

The most common solution to this open-ended response problem is to use an equation editor (see Figure 1).

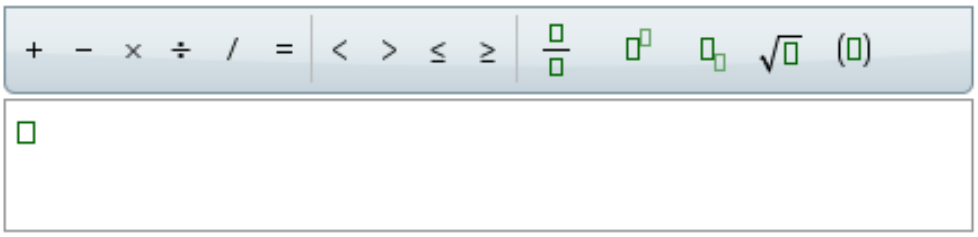


**Figure 1.** Sample Equation Editor Entry Box

While equation editors allow for precise entry of mathematical expressions, they do add an extra burden on the student. These additional difficulties can be both construct-related (i.e., students who struggle with mathematics may struggle more to use the equation editor due to not understanding the various mathematical symbols, orders of entry, etc.; see Noyes, Garland, & Robbins, 2004) and construct-irrelevant (i.e., students who have less exposure to equation

Hodges, T.E., Roy, G. J., & Tyminski, A. M. (Eds.). (2018). *Proceedings of the 40th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Greenville, SC: University of South Carolina & Clemson University.

editors, regardless of mathematical ability, may require extra time to identify and select the proper symbols and where to click or type; see Leeson, 2006). Hargreaves et al. (2004) also showed that students may solve problems differently when presented assessments through different media.

Tablets and other devices that allow for handwritten digital entry could resolve some of this burden, but unless those handwritten responses can be automatically scored with the same ease as the typed equation editor responses, the cost (in both time and money) of scoring the assessment becomes too great for this to be a reasonable solution. Thus, we cannot administer assessments with digital handwritten entry without automated handwriting recognition. In this paper, we discuss a study intended to be the first in a series of experiments aimed at first identifying differences between equation editor entry and automatically translated handwritten entry and then addressing the challenges of each towards the development of a solution that provides the most authentic experience as possible for students that minimizes construct-irrelevant difficulties.

## Study and Research Questions

As stated above, this study describes the first year of a multi-year study aiming to understand and improve equation entry for mathematics assessments. Our goal for the first year of the project was 1) to demonstrate that we can score equations and expressions that have been captured on a tablet (Apple iPad) via handwriting with a stylus, then automatically translated into MathML (Ausbrooks et. al., 2010) via the translator MyScript (Vision Objects, 2017) and 2) to determine if there is a difference in performance when students enter responses on a computer, using an equation editor (WIRIS, see Maths for More, 2017), versus entering responses on a tablet. If we suppose that prior evidence with paper and pencil is an appropriate stand-in for tablets, research would suggest that it is easier for students to handwrite responses on a tablet using a stylus than to enter them on a computer using an equation editor and that students are less likely to make construct-irrelevant errors (see Hargreaves et al., 2004). The ability to administer assessments on a tablet, and to score the responses automatically, will increase flexibility and the attractiveness of digitally based assessments to both stakeholders and test-takers. For this study, we had students copy equations directly from a screen to study only the entry of the equation itself and minimize any effect that the method of entry may have on the solution strategy prior to equation entry.

We designed a study to answer the following research questions:

1. What are the potential causes of errors and variability in score differences with automatically scored equation responses that have been captured on a tablet via handwriting with a stylus?
2. Is there a difference in student performance between responses handwritten on a tablet and automatically translated into MathML and responses typed on a computer using the equation editor WIRIS?
3. Are the error rates of the handwriting recognition comparable to (or better/worse than) the errors students make when entering equations into an equation editor?

## Methods

We developed two parallel forms (termed Form A and Form B) of a mathematics assessment that would each be administered on computer (using an equation editor) and on an iPad (using handwriting recognition). In both forms, students are asked to copy equations directly from the

Hodges, T.E., Roy, G. J., & Tyminski, A. M. (Eds.). (2018). *Proceedings of the 40th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Greenville, SC: University of South Carolina & Clemson University.

screen. The forms had different equations, yet we developed them to be as parallel as possible in terms of mathematical complexity, the necessary use of equation editor templates, and the number of characters. For the first half of the assessment, students are looking at the equation while they copy it, and for the second half, students must hide the stimulus while copying, but they are allowed to go back to view the stimulus as many times as necessary. All students took both forms. One-half of the participants handwrote their responses to the item on the first assessment on a tablet with a stylus and entered their responses to the items on the second assessment on a computer using an equation editor. The other half of participants completed the two assessments in the opposite order. The two groups of students were further subdivided into Form A or Form B and counterbalanced for order. Thus there were four groups, into which participants were randomly assigned (see Table 1).

**Table 1: Counterbalanced design**

| | **First Task: Computer<br>Second: iPad** | **First Task: iPad<br>Second: Computer** |
|---|---|---|
| Form A: iPad<br>Form B: computer | Group 2 | Group 1 |
| Form A: computer<br>Form B: iPad | Group 3 | Group 4 |

In each form, students were asked to copy 20 mathematical equations that ranged from very simple to enter (e.g., 6+2=8) to complex (e.g., one form of the quadratic formula). No equation exceeded in difficulty or complexity that which a student would see as part of a typical Algebra I class. For the computer (equation editor) form, all responses were automatically scored using a proprietary mathematical scoring engine (m-rater) that works through MathML. For the tablet (handwritten responses), all responses were automatically translated into MathML and then scored using the same scoring engine. For this study, we used the WIRIS equation editor, which is widely used in K-12 mathematics assessments and the MyScript Handwriting recognition tool, which has a popular iOS handwritten mathematics calculator and is also used in various mathematical contexts by some leading tech designers.

See Figure 2 for the process of translating the handwritten responses into MathML.

Hodges, T.E., Roy, G. J., & Tyminski, A. M. (Eds.). (2018). *Proceedings of the 40th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Greenville, SC: University of South Carolina & Clemson University.

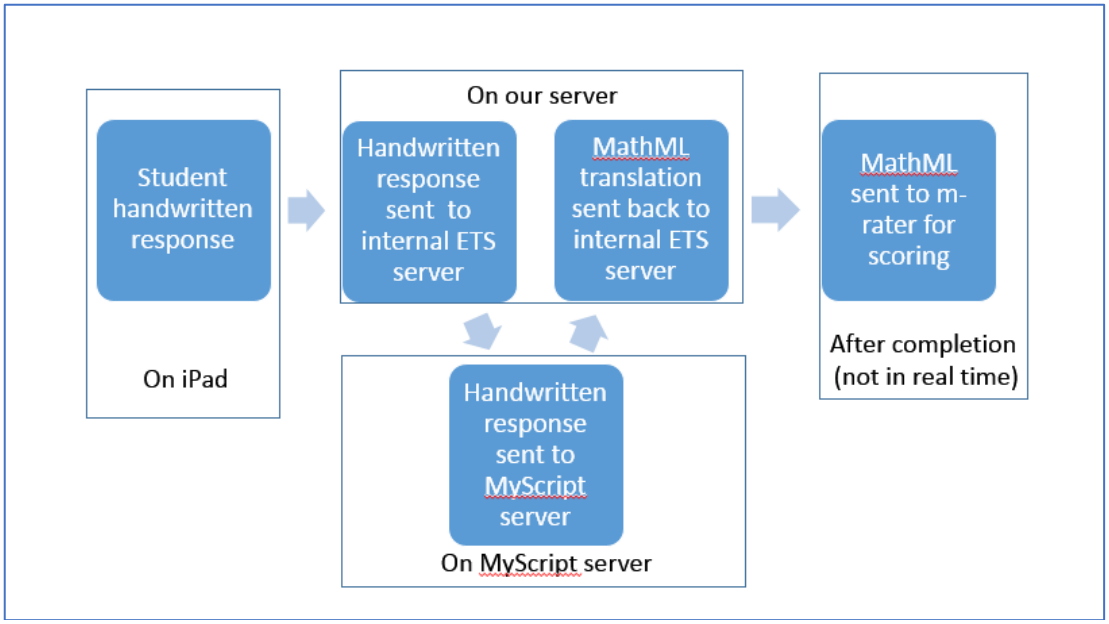


**Figure 2.** Handwriting to MathML process

## Sample and Data Collection

We recruited 9th-grade students from 4 high schools in the USA from racially, culturally, and socioeconomically diverse areas (35% of students qualified for Free/Reduced Lunch of those that reported). Overall, 474 9th-grade Algebra students completed the study (204 Male, 265 Female, 5 Nonreported). The final sample was 45.8% White, 25.9% African American, 10.1% Asian American, 9.7% Hispanic/Latino, and 8.4% other). All students participated during their regular mathematics class and we randomized the experimental groups to which students were assigned within classes (i.e., each class had students placed into each of the four groups). We encountered some connectivity issues discovered after administration whereby 143 students in varying classes and schools did not receive MathML translations from the MyScript server as well as some various missing data with 22 students. Therefore, analyses were conducted with the remaining 309 students so as to maintain the counterbalanced design.

## Results

The students received one point for each response that was an exact match to the target equation (in the case of the handwritten responses, one point was given per response in which the computer translation was an exact match to the target). No credit was given for partial responses. Overall, the automatically-produced scores were higher with the equation editor than with the translated handwriting (Form 1: iPad average 10.96 (out of 20), Computer Average, 15.46 (out of 20); $t(156) = 13.4$; $p < .001$; Form 2: iPad average 10.59 (out of 20), Computer Average, 15.5 (out of 20); $t(145) = 11.5$; $p < .001$).

Hodges, T.E., Roy, G. J., & Tyminski, A. M. (Eds.). (2018). *Proceedings of the 40th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Greenville, SC: University of South Carolina & Clemson University.

It is worth reiterating that these scores are a reflection of the computer translated handwriting not a score obtained from the writing itself. In other words, we are not comparing what a student wrote with what they typed, but rather how well their handwriting translated into the correct scoreable form versus what they typed. Thus, we can state that the student responses entered via equation editor far outperformed the computer translated handwritten responses, but we make no statement as to the writing on the iPad itself and whether it would have been scored correct or incorrect by a human rater. We do this because it is that final translated response that is the subject of the viability of this as an option for large-scale assessment, which is the focus of this multi-year project.

Despite the higher overall scores for the equation editor, not all items are created equal. For some items, scores between the conditions were about equal (and not significantly different) while for others there were large differences. Figure 2 shows the differences for all scores in both Form A and Form B. As can be seen, both forms produced very similar scores per item, no differences between Forms are significant.

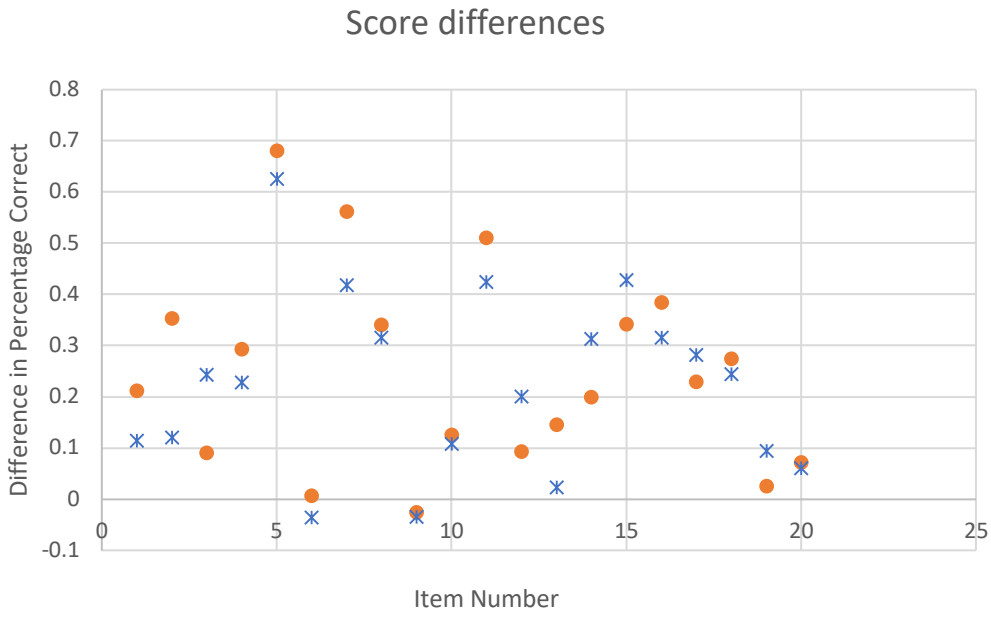


**Figure 3.** Score differences, equation editor minus tablet (Form A in blue square and Form B in orange circle)

As stated in our first research question, we wished to not only look at overall performance, but to better understand the potential causes of the variability in score differences. To do so, we looked at individual examples of high and low difference items. As an illustration, Table 2 lists the three items with the smallest and largest differences in percentage correct over both Form A and Form B.

**Table 2: Items with Greatest and Smallest Differences in Percentage Correct**

| Item | Equation editor mean | Handwritten mean | Difference (Equation Editor minus Handwriting) |
|---|---|---|---|
| **Smallest Differences** | | | |
| $x + y = y + x$ | 87.4% | 86.7% | 0.7 |
| $4 \times 2 = 8$ | 88.7% | 86.3% | 2.4 |
| $y = ax^2 + bx + c$ | 75.4% | 78.0% | -2.6 |
| | | | |
| **Greatest Differences** | | | |
| $28 \div 4 = 7$ | 98.7% | 30.7% | 68.0 |
| $36 \div 4 = 9$ | 92.0% | 29.4% | 62.6 |
| $x(y + z) = xy + xz$ | 84.9% | 28.7% | 56.2 |

In taking a deeper dive into the characteristics of the equations with higher and lower handwriting recognition rates, we found that while variables ($x$ and $y$) were recognized reasonably well, the multiplication symbol (×) was occasionally recognized as an $x$. Additionally, the equation $y = ax^2 + bx + c$ was one of the only items in which the handwritten mean was higher than the equation editor mean, largely because many students had difficulty using the template for the superscript, thus resulting in errors such as $y = ax^{2+}bx + c$ and similar errors.

As can be seen in two of the examples with the greatest difference in response type, the division symbol (÷) had a fairly low recognition success rate, and even though variables were recognized well, once parentheses were used along with the variables, recognition dropped substantially.

However, this is clearly only one part of the story. As part of this exploration of the differences, we felt it was also important to look at timing differences. Our rationale was that if equation editors do allow for higher performance but also take much longer to use, there may be some tradeoff between time and accuracy.

Figure 3 shows the difference in average time taken on the item for both Form A and Form B (computer minus iPad). As can be seen in Figure 3, timing scores for both forms were nearly identical, with both forms having high outliers for items 1 and 11. Both of these items occurred at the beginning of a new section in each form, and thus it appears that the time difference is more a product of the students on computers taking longer to read the directions and begin typing than it is of the item itself. For the rest of the items, differences tended to range between zero and ten seconds, though we did see a great amount of variability between students such that some students had virtually no difference while others took much longer for the computer entry than the iPad. The exploration of individual student characteristics will be the focus of a future analysis.

Hodges, T.E., Roy, G. J., & Tyminski, A. M. (Eds.). (2018). *Proceedings of the 40th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Greenville, SC: University of South Carolina & Clemson University.

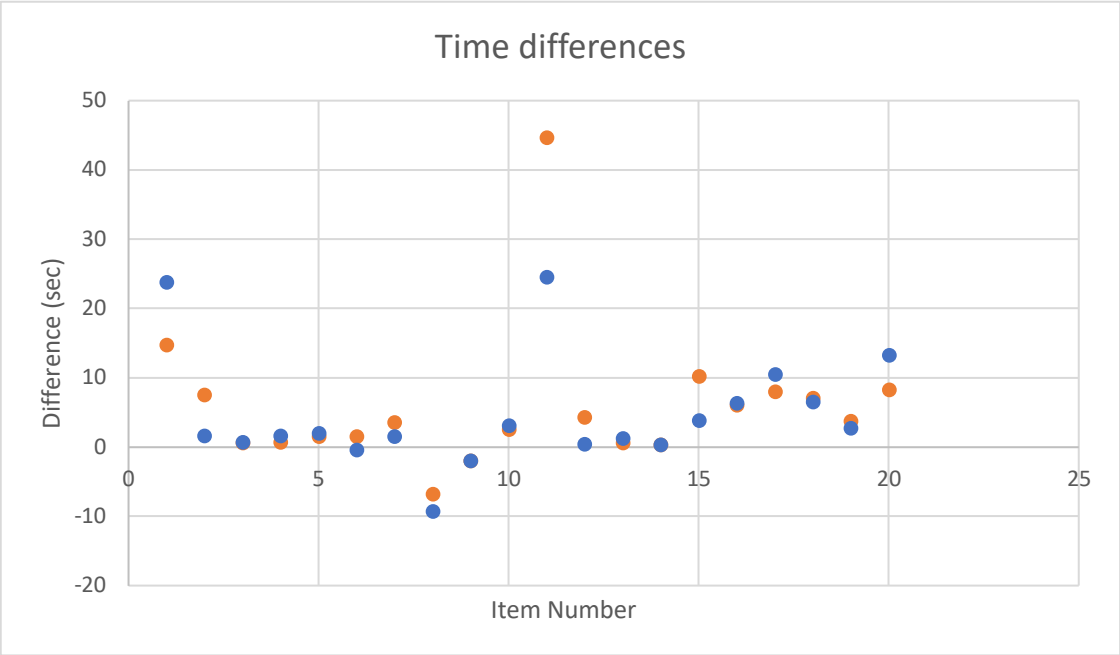


**Figure 4.** Timing Differences, equation editor minus tablet (Form A in blue and Form B in Orange)

Overall, using current technology, automatically produced scores were much higher when equations were entered using a standard equation editor than when using handwriting recognition software. While those equations did take slightly longer to enter, the timing differences were not large enough for this mode of entry to result in more than a few minutes of extra testing time (depending on the number of equations being entered). One caveat is that this was for copying equations. It would be interesting to see if timing differences were similar when students were asked items in which they had to generate equations, as we could see some compounded differences with uses of scratch paper, etc.

## Discussion

What should we conclude about the future of DBA mathematics response entry on the basis of this study? While on the surface, we seem to have found that equation entry is currently superior to handwritten responses, we also uncovered important difficulties with equation editor responses that are particularly troubling considering this is the current preferred response method. For example, equation editor responses ranged between 70% and 95% correct. Considering that students were copying equations, a 70% correct response rate means that 30% of students did not accurately copy an equation using a typical editor that has been the standard of DBA. While these rates did outperform translated handwriting, we see this as more of an indicator that handwriting translation technology needs to be improved so that it can eventually replace the much troubled equation editors, as opposed to an indicator that equation editors are superior for response.

We are currently planning a follow-up study in which students are able to see their handwritten translations in real time and make corrections to their writing to see if this capability enables students to raise their own rates of recognition. We are also looking into ways to limit the lexicon of the recognition software to only include those characters which are part of the Algebra I curriculum (i.e., exclude most Greek letters, integrals, and derivatives). We also plan

to see if a tutorial on how to use the equation editor may improve equation editor responses. These improvements to both of these entry methods would need to be considered along with how much time it may cost the test-taker. For instance, watching a tutorial would be a one-time cost of a few minutes, while the ability to see the handwriting translation could potentially cost test-takers a lot of time if they need to correct every individual item (some more than once).

## Conclusion

Overall, it is true that we should have some concerns about responses to DBA mathematics items that require students to enter mathematical equations or expressions, regardless of entry mode. Equation editor items should potentially allow for some leeway in scoring on responses that indicate the student may have had entry difficulties (e.g., where it appears students have had trouble knowing how and where to use a template). Additionally, we should also not discount the future of handwritten entry, though the technology is not currently up to the state it should be for assessment use. Since DBA mathematics assessment is here to stay, and growing in use, we need to continue this line of work to improve the entry capabilities to the point that they seamlessly allow students to show what they truly know in the mathematics and not be limited by the technology.

## References

Ausbrooks, R., Buswell, S., Carlisle, D., Chavchanidze, G., Dalmas, S., Devitt, S., & Kohlhase, M. (2010). *Mathematical markup language (MathML) version 3.0*. Tech. rep. World Wide Web Consortium (W3C), 2010. url: http://www. w3. org/TR/MathML3.

Hargreaves, M., Shorrocks-Taylor, D., Swinnerton, B., Tait, K., & Threlfall, J. (2004). Computer or paper? That is the question: does the medium in which assessment questions are presented affect children's performance in mathematics?. *Educational Research*, 46(1), 29-42.

Leeson, H. V. (2006). The mode effect: A literature review of human and technological issues in computerized testing. *International Journal of Testing*, 6(1), 1-24.

Maths for More (2017). WIRIS equation editor. url: http://www.mathsformore.com

Noyes, J., Garland, K., & Robbins, L. (2004). Paper-based versus computer-based assessment: is workload another test mode effect?. *British Journal of Educational Technology*, 35(1), 111-113.

Vision Objects (2017). Myscript handwriting recognition engine.