International Journal of Social Science and Economic Research

ISSN: 2455-8834

Volume:06, Issue:06 "June 2021"

SELF-ASSESSMENT OF UNDERSTANDING: WE DON'T ALWAYS KNOW WHAT WE KNOW

John Leddo, PhD[1], Ava Clark[2] and Emma Clark[2]

[1]Director of Research at MyEdMaster.
[2]researchers at MyEdMaster.

MyEdMaster, LLC, Herndon, Virginia, USA

DOI: 10.46609/IJSSER.2021.v06i06.005 URL: https://doi.org/10.46609/IJSSER.2021.v06i06.005

ABSTRACT

Personal assistants and educational software both convey information to their users. Personal assistants allow users to ask questions but generally do not check to insure that users understood the answers. Educational software generally does not allow users to ask questions and typically assesses understanding of content through testing. The present project explores whether simply asking people if they understand information they are given is an accurate way to assess their understanding. 18 middle school students and 19 adults were taught topics in quadratics on a Google Form. After each piece of instruction, they were asked whether or not they understood what they were taught. Upon completion of the instruction, they were given problems to solve based on the lessons. Participants' self-assessments of whether or not they understood what they learned were matched to whether or not they gave the correct answers to the corresponding problems. Results showed that when indicating that they understood the lessons, students correctly answered associated problems 66.2% of the time and adults correctly answered them 73.6% of the time. These percentages were both statistically higher than a 50-50 coin flip in terms of their diagnositicity, but also significantly lower than 99% (near perfect diagnositicity). When indicating that they did not understand the lesson, students gave the wrong answer 62.5% of the time and adults gave the wrong answer 90.4% of the time. This percentage for students was significantly lower than the 99% level but not statistically different from 50%. The percentage for adults was significantly higher than 50% but significantly lower than 99%. Results suggest that self-assessments of understanding are somewhat diagnostic of true understanding (or lack thereof) but not sufficiently diagnostic to be relied on by themselves.

Key Words: Personal assistants, educational software, question answering, self-assessment

 Copyright © IJSSER 2021, All rights reserved

ISSN: 2455-8834

Volume:06, Issue:06 "June 2021"

Introduction

Personal assistants and educational software share the important function that both attempt to increase their users' knowledge. While both convey information to their users, personal assistants such as Siri or Google Assistant are more flexible than typical educational software in that they respond to direct questions posed by their users. It seems self-evident that educational software should be equally responsive, just as human tutors and teachers are. Accordingly, we have incorporated such a question-and-answer capability into the artificial intelligence (AI)-based educational software we are developing (Leddo et al., 2019).

The next logical phase of incorporating question-and-answer capabilities into personal assistants and educational software is to insure that users understand the information they are given. Just as there are different ways to ask a question, so, too, are there different ways to answer a question. The present project is part of a series of investigations to determine how best to answer people's questions to maximize their understanding of the answers and overall learning of the questions' subject matters.

Devil's advocates may wonder why the customary practice of providing users with direct answers to questions is not sufficient. There are two compelling answers to this. First, research suggests that people do havepreferred methods of receiving information (John et al., 2016). Second, our previous research found that when students were given the option to pick the format they wanted their questions to be answered in (i.e., informational, using a real-world example, explaining cause and effect principles, or stating the goals that are served) their performance was double that of students who were given only informational answers (Leddo et al., 2021).

Given that people have preferences for how they want questions to be answered and how questions are answered makes a difference in how well people learn, the next question is to determine how best to answer questions to maximize learning. One solution, and this may be part of the overall method for addressing this issue, is that the type of question being asked may dictate the best way to answer the question. For example, the Leddo et al. (2021) study found that informational answers were overwhelmingly preferred when questions about facts were asked. However, when people asked "why?"-type questions, the preference shifted to other types of answers.

We believe that, while answers to questions can be optimized, we do not want to leave it to people themselves to search through answer types and pick the one they want. This process seems to be cumbersome. Rather, we believe machine learning can play a valuable role here by determining the best way to answer questions based on parameters relating to type of question and characteristics of the questioner (e.g., age, level of existing knowledge on the topic). In

 Copyright © IJSSER 2021, All rights reserved

ISSN: 2455-8834

Volume:06, Issue:06 "June 2021"

order for machine learning to work in this capacity, two things are required. First, we must specify the range of parameters that may affect an appropriate answer. Second, we need a measure of user understanding that can be used to train the machine learning model, so that it can determine whether the user did understand the answer.

Determining whether a person understood an answer to question may not be an easy task. For example, educational software typically assesses whether people understood the information they were given by giving them post-tests and measuring performance. While such assessments may be accurate gauges of understanding, these tend to be time consuming, and it is unlikely that people will want to endure such assessments each time they ask a question. In fact, the prospect of facing an assessment after having their questions answered may even deter people from asking them. A somewhat less burdensome assessment approach may be to use Cognitive Structure Analysis (Leddo and Cohen, 1989). In Cognitive Structure Analysis (CSA), a person is asked questions to probe the different types of knowledge associated with a concept (facts, procedures, problem solving plans, cause and effect relationships). Based on the respondent's answers, a representation is created of what the person knows about that topic. CSA can be implemented in dialog format, meaning the assessment can be delivered conversationally and with limited time commitment. Moreover, CSA has been shown experimentally to be highly diagnostic of what someone knows as CSA's assessment of a person's subject matter knowledge correlates .88 with his or her problem-solving performance (Leddo and Sak, 1994).

While using a technique such as CSA to assess whether a person understands the information he or she has been given, the devil's advocate may simply ask, "Can't you just ask people if they understood what you told them? After all, people do that to each other all the time." This does seem to be a reasonable question to ask, especially since there is a history of asking people to describe how they solve problems and using those protocols as valid representations of knowledge (Ericsson and Simon, 1984).

The flip side of this coin is research done by Nisbett and Wilson (1977). In a paper, aptly named "Telling more than we can know", Nisbett and Wilson report extensive research demonstrating that people are not always good judges of what they know or how they arrive at the conclusions they do. Accordingly, we believe it would be unwise to take it on faith that people's self-assessment of whether or not they understood information they were given is completely accurate. Rather, the reliability of self-assessments should be investigated empirically, which is the purpose of the present study. In this study, both adults and middle school students were taught algebra topics. After each topic, they were asked to indicate whether or not they understood the lesson material. They were then given problems to solve relating to those lessons. The research questions posed in the present study are the extent to which self-

 Copyright © IJSSER 2021, All rights reserved

ISSN: 2455-8834

Volume:06, Issue:06 "June 2021"

assessments of understanding of the information given in the lessons are diagnostic of problem-solving performance and whether this diagnosticity differs for adults and students.

Method

Participants

The participants were 18 middle school students and 19 adults recruited from the Northern Virginia area in the United States. Middle school students were screened based on whether they had already learned the Algebra 1 topic of quadratics. If they already knew, their data were excluded from the analysis.

Materials

Materials used in the experiment were delivered on a Google Form. The first page of the Form collected data about whether the participant was a child or adult and whether the person already knew quadratics and parabolas. Subsequent pages presented instruction on features of polynomials, using FOIL (first, outer, inner, last) to multiply binomials to produce a trinomial, factoring quadratics and graphing parabolas. After each major teaching point, the Google Form asked the user to indicate whether s/he understood the material or was confused. Once the instruction was complete, there was a 10-question post-test that covered all the topics. Three of the questions were multiple choice and seven required the participant to type in an answer.

Procedure

Participants were tested individually. Each was given the link to the Google Form. Each instructional page consisted of a self-contained lesson on one of the above topics and the last page was devoted to the post-test.Participants were allowed to complete the Google Form at their own pace. Once they completed the Google Form,participants hit the submit button and their answers were scored.

Results

Fortunately, each participant answered each question on the Google Form, providing the analysis with a complete set of data. Participants' responses on the Google Form were analyzed for two things: whether they indicated that they understood or did not understand each lesson component and whether they correctly answered the questions linked to the lessons. This analysis created four possible combinations: participants said they understood and gave the correct answer, they said they understood and gave an incorrect answer, they said they did not

 Copyright © IJSSER 2021, All rights reserved

ISSN: 2455-8834

Volume:06, Issue:06 "June 2021"

understand and gave an incorrect answer, and they said they did not understand and gave a correct answer.

Of these four possible outcomes, understand/correct answer and not understand/incorrect answer show a match between a participant's self-assessment of his or her understanding of the lesson and his or her answer to the test question. Understand/incorrect answer and not understand/correct answer show a mismatch between a participant's self-assessment of his or her understanding of the lesson and his or her answer to the test question. Given that there were a total of 10 lessons and associated questions, each participant contributed 10 data points to the overall analysis.

The data points for the four possible outcomes were tallied across student and adult participants, yielding a total of 180 data points for students and 190 data points for adults. The tallies for students and adults are presented in Tables 1 and 2, respectively.

Table 1: Number of Students' Correct and Incorrect Answers Based on Self-assessments of Lesson Understanding

	Correct answer	Incorrect Answer	Total
Understood lesson	98 (66.2%)	50 (33.8%)	148
Did not understand lesson	12 (37.5%)	20 (62.5)%	32
Total	110	70	180

Table 2: Number of Adults' Correct and Incorrect Answers Based on Self-assessments of Lesson Understanding

	Correct answer	Incorrect Answer	Total
Understood lesson	109 (73.6%)	39 (26.4%)	148
Did not understand lesson	4 (9.6%)	38 (90.4%)	42
Total	113	77	190

 Copyright © IJSSER 2021, All rights reserved

ISSN: 2455-8834

Volume:06, Issue:06 "June 2021"

Overall, students showed an average match between self-assessments of understanding/not understanding and their problem-solving performance 65.6% of the time, while adults showed an average match 77.4% of the time. Initially, this suggests that adults may be more reliable than middle school students in assessing how well they understand or do not understand what they were taught. A comparison of adults' vs. students' data reveals that adults are more reliable in their self-assessment of understanding what they have learned, 66.2% for students vs. 73.6% for adults, $z = 2.2$, $p < .05$. Adults are also more reliable than students in their self-assessment of not understanding what they have learned, 62.5% for students vs. 90.4% for adults, $z = 2.89$, $p < .01$.

The above analysis is comparative between students and adults. However, that analysis does not address the objective reliability of student and adult self-assessments. If the goal is to use self-assessments as a source of input into an algorithm that determines whether the format by which a question is answered or a lesson is taught is understood by an end user, then it is important to understandhow diagnostic a self-assessment of "I understand" or "I don't understand" is. To shed light on this question, additional analyses were performed that considered the two extremes of the spectrum. At one end of the spectrum, a statement of "I understand" or "I don't understand" has no diagnostic value, i.e., it is indistinguishable from a coin flip. At the other end of the spectrum, the self-diagnosis is virtually a perfect indication of understanding or lack thereof. Accordingly, we assume that if a self-assessment has no diagnostic value, its associated problem-solving performance is 50%, i.e., the respondent's likelihood of getting the correct answer to a problem is 50% regardless of whether he or she claims to understand or not understand the topic. Similarly, we assume that if a self-assessment has virtually perfect diagnostic value, the associated problem-solving performance is 99% (expecting 100% performance may be too unrealistic since people are not perfect), i.e., the respondent is 99% likely to get the correct answer to a problem when he or she claims to understand the topic and 99% likely to get the incorrect answer to a problem when he or she claims not to understand the topic.

We, therefore, compared the self-assessment probabilities for students and adults against the 50% and 99% criteria. For students, their self-assessment accuracy was 66.2% when they claimed to understand the topic. This was both significantly higher than the 50% non-diagnostic level, $z=3.94$, $p < .01$, and significantly lower than the virtually perfect 99% level, $z = -40.1$, $p < .001$. This suggests that their self-assessment was somewhat diagnostic of their true understanding, but not sufficiently diagnostic that it could be relied on alone. On the other hand, students' self-assessment accuracy was 62.5% when they claimed not to understand the topic. This was significantly below the virtually perfect 99% level, $z = -20.75$, $p < .001$, and not statistically different from the non-diagnostic 50% level, $z = 1.41$, ns.

 Copyright © IJSSER 2021, All rights reserved

ISSN: 2455-8834

Volume:06, Issue:06 "June 2021"

For adults, their self-assessment accuracy was 73.6% when they claimed to understand the topic. This was both significantly higher than the 50% non-diagnostic level, $z=6.67$, $p < .01$, and significantly lower than the virtually perfect 99% level, $z = -31.06$, $p < .001$. This suggests that their self-assessment was somewhat diagnostic of their true understanding, but not sufficiently diagnostic that it could be relied on alone. On the other hand, adults' self-assessment accuracy was 90.4% when they claimed not to understand the topic. This was significantly below the virtually perfect 99% level, $z = -5.6$, $p < .01$, and significantly higher than the 50% non-diagnostic level, 5.24, $p < .01$. This result suggests that adults' self-assessment of not understanding a concept was somewhat diagnostic of their lack of understanding (and more so than students'), but not sufficiently diagnostic that it could be relied on alone. Still, 90% reliability is reasonably high.

Discussion

The present study is part of an ongoing research program to create educational software that optimizes instruction to the needs of each student. A key feature of this software is the capability for students to ask questions as they learn, much as they could do so with human teachers. As our previous research (Leddo et al., 2021) has shown, how a question is answered greatly affects how well the student learns. As stated in the Introduction of the present paper, our goal is to use machine learning to optimize how to answer questions on an individual student basis. This goal raises the issue of how to determine how well the student understood the answer to the question.

The goal of the present study is to evaluate how reliable a person's self-assessment of understanding of what he or she has been taught is. If self-assessment is highly reliable, then educational software can simply ask, as people often do with one another, if the student understood the material. The results of the present study suggest that simply asking people if they understand something is not sufficiently reliable to be used as a sole measure of understanding. While both middle schoolers' and adults' claims of understanding teaching material was statistically more reliable than a coin flip in predicting if they can solve problems using that material, their performance was also statistically less reliable than nearly perfect in making such predictions. When indicating that they did not understand teaching material, middle schoolers' self-assessments were not statistically different than using a coin flip to predict success at solving problems. Adults were more reliable, however, in indicating that they did not understand teaching material as their claims of not understanding predicted failure to correctly solve problems approximately 90% of the time. While this was statistically much higher than a 50-50 coin flip, it was still statistically significantly below a near perfect 99% accuracy level. Finally, overall, adults were more accurate than middle schoolers in self-assessing both understanding and not understanding educational material. However, a potentially confounding

 Copyright © IJSSER 2021, All rights reserved

ISSN: 2455-8834

Volume:06, Issue:06 "June 2021"

variable for this conclusion could be that adults may have seen the topics before when they were in school, something the middle schoolers had not.

Collectively, these results suggest that there is some diagnostic value in having educational software ask users if they understand what they are taught or answers to questions they ask, except perhaps, counterintuitively, when children say they do not understand something, although this finding merits further investigation. However, the results also suggest that additional data should be combined with self-assessments of understanding in order to more accurately determine whether a person has understood what he or she was taught or the answer to a question that was asked. As mentioned in the Introduction, using Cognitive Structure Analysis offers promise as a highly diagnostic way to assess a person's understanding of a topic. While this is more labor intensive than simply asking a person if he or she understands, it appears that self-assessment is not sufficient and needs to be combined with alternative methods. Currently, most instructional software assesses understanding by giving students quizzes. While this is certainly a necessary part of an overall educational process, it seems like overkill for most question-and-answer interactions as might occur with personal assistants. In fact, it is quite possible that such rigorous testing after an explanation or answer to a question is given may actually deter people from asking questions. It is possible that a combination of asking people if they understood what they were told and then asking them some dialog-based diagnostic questions may be a highly reliable middle ground to insure that they understood what they were taught/told.

Conclusion

The present research suggests that people's self-assessment of what they know and do not know can be inaccurate and therefore may not be sufficient to rely on as the sole basis of assessing whether they understood information that was given to them. This has important implications for those involved in creating personal assistants or educational software and even in everyday conversations when people want to make sure others understand them. Since how questions are answered and information is presented to people can great affect understanding of that information, those involved in personal assistant or educational software will need to rely on a variety of data sources to accurately evaluate how well intended recipients of information understand that information. Research suggeststhat both individual characteristics and information type need to be taken into account in order to optimize understanding. Additional research is needed to delineate what other variables matter.

 Copyright © IJSSER 2021, All rights reserved

ISSN: 2455-8834

Volume:06, Issue:06 "June 2021"

References

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. The MIT Press.

John, A., Shahzadi, G. & Khan, K.I. (2016). Students' Preferred Learning Styles and Academic Performance. Sci.Int.(Lahore),28(4),337-341.

Leddo, J. and Cohen, M.S. (1989). Cognitive structure analysis: A technique for eliciting the content and structure of expert knowledge. Proceedings of 1989 AI Systems in Government Conference. McLean, VA: The MITRE Corporation.

Leddo, J. and Sak, S. (1994) Knowledge Assessment: Diagnosing what students really know. Presented at Society for Technology and Teacher Education.

Leddo, J., Guo, Y., Liang, Y., Joshi, R., Liang, I., Guo, W. and Bailey, S. (2019). Artificial Intelligence and Voice-powered Electronic Textbooks. International Journal of Advanced Educational Research, 4(6), 44-49.

Leddo, J., Chen, T., Menachery, A., Agarwal, J. and Agarwal, T. (2021). Towards Improving Personal Assistants and Educational Software: How Questions Are Answered Affects Learning. International Journal of Social Science and Economic Research, 6(02), 696-705.

Nibett, R.E. and Wilson, T.D. (1977). Telling More Than We Can Know: Verbal Reports on Mental Processes. Psychological Review, 84(3), 231-259.

 Copyright © IJSSER 2021, All rights reserved