



# What can we assess using multiple choice problems that involve learner generated examples?

Shai Olsher, Michal Yerushalmy

## ► To cite this version:

Shai Olsher, Michal Yerushalmy. What can we assess using multiple choice problems that involve learner generated examples?. CERME 10, Feb 2017, Dublin, Ireland. hal-01942129

**HAL Id: hal-01942129**

**<https://hal.science/hal-01942129>**

Submitted on 2 Dec 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# What can we assess using multiple choice problems that involve learner generated examples?

Shai Olsher<sup>1</sup> and Michal Yerushalmy<sup>2</sup>

University of Haifa, Faculty of Education, Haifa, Israel; [olshers@edu.haifa.ac.il](mailto:olshers@edu.haifa.ac.il),  
[michalyr@edu.haifa.ac.il](mailto:michalyr@edu.haifa.ac.il)

*Multiple choice (MC) items are the natural choice for automated online assessment. Ideally, making a choice should be based on knowledge and reasoning. Nevertheless, studies demonstrate that often various techniques (e.g. guessing) are the common practices. In the last decade technology has been employed to support real-time feedback as formative assessment for teaching and learning. This study examines whether and how learner generated examples, when required as support to the choice made in MC task, could be automatically identified to give insights into learners' understandings. Results show discrepancies between chosen correct statements and their supporting examples. Other automatically assessed characteristics are related to learner's approaches and strategies.*

*Keywords: Reasoning with examples, geometry, multiple choice questions, automatic online assessment, formative assessment.*

## Theoretical background

Multiple choice (MC) tasks are the most well-known tasks when it comes to automatic assessment (Farrell & Rushby, 2016; Sangwin & Kocher, 2016). They are used for testing various topics of study as well as different levels of abilities from basic through high order thinking. They hold several advantages including objectivity of scoring, and availability of more items in each assessment due to short solving times (Farrell & Rushby, 2016). However, MC tasks are often criticized for being biased (Hassmen & Hunt, 1994), for sometimes being measurements of how fast a student could make an educated guess or use elimination techniques, and not necessarily assessing what the MC tasks were designed to assess (Lau, Lau, Hong, & Usop, 2011).

Recent use of technology has made it possible to automatically assess responses by not only using MC tasks (Stacey, & Wiliam, 2012; Sangwin & Kocher, 2016). Immediate presentations of information on tasks performed in a technological environment are used as means to formative assessment: serving as feedback to modify teaching and learning activities (Black & Wiliam, 1998). One of these avenues is by automatically assessing learner generated examples (LGE) in a dynamic geometry environment (DGE) (Leung & Lee, 2013). Example generation tasks may serve as possible means to show conceptions of mathematical objects, or concept images (Vinner, 1983), informing about possible difficulties and inadequacies (Zaskis & Leikin, 2007). Another use of examples is for determining the validity of mathematical statements (Nagari-Haddif & Yerushalmy, 2015), in which systematic analysis of LGE could shed light on the evolving understanding of the status of examples in proving or refuting a mathematical claim (Buchbinder & Zaslavsky, 2009). Combining the accessibility of MC tasks in assessment with the potential reasoning abilities and demonstration of understanding by generating examples in a technological environment has the potential to enhance formative assessment in the mathematics classroom.

## Methodology

This study is part of a long-term project aiming to explore ways in which automatic assessment could give more insight about student understanding in mathematics (Olsher, Yerushalmy, and Chazan, 2016). Specifically, we ask whether the requirement to provide examples to support a chosen answer gives the assessor additional insight into students' understanding in MC tasks. The participants were 32 secondary Israeli geometry teachers, from all over the country who taught different levels and ages ranging from 7<sup>th</sup> to 12<sup>th</sup> grade. The study was conducted as part of a broader national professional development effort to expose teachers to innovations in mathematics education.

## Tasks

The study included three tasks and the context was mathematical similarity. The tasks were designed as interactive diagrams describing a geometrical context, using the STEP platform<sup>1</sup>. The interactive diagrams were constructed using GeoGebra and they enabled the participants to drag a set of elements in the diagram, according to the predefined characteristics determined by the designers of the task. The context was described in the task, and several statements were provided for the participants to consider. The participants were required to select the statements that are correct in regards to the diagram. More than one statement could be correct in relation to the interactive diagram in any single task. The tasks were similar to conventional multiple choice tasks accompanied by an interactive diagram, with one main difference: In these tasks the participants were expected to experiment and manipulate the interactive diagram into a state that fits one or more of the statements. In order to add another layer of reasoning, we asked the participants to attach a screen capture of the applet in a state that exemplified each of the statements they have chosen, thus requiring the participants to add an example instead of just select a statement as in traditional multiple choice tasks.

## Automatic checking of tasks

The STEP platform enabled an automatic analysis of the predefined mathematical properties of the submitted solutions in order to characterize these solutions pedagogically and mathematically (Olsher, Yerushalmy and Chazan, 2016). The tasks in this study were designed so that the system would indicate if the corresponding example fits the criteria in the relevant statement<sup>2</sup>, and enable the teacher to immediately have access to filtered answers accordingly. Yet, it is important to state that at the present time technology cannot determine correctness on its' own for these types of rich tasks. For each task, a well-defined set of conditions should be applied in order to determine the type of feedback affording formative assessment. Meeting the conditions of the checking algorithm does not mean correctness. It just means that this is what was automatically checked, and any

---

<sup>1</sup> Seeing the Entire Picture - STEP – is a formative assessment platform developed at the University of Haifa's Center for Mathematics Education Research and Innovation (MERI). For more detail about this platform, see [www.visustep.com](http://www.visustep.com).

<sup>2</sup> When automatically checked, a margin of accuracy was determined by the teacher in which solutions are considered sufficiently accurate. For example, in this case, parallel lines, or coinciding points.

interpretation about correctness is purely suggestive, and should be carefully examined by both the assessor and the assessee.

## Analysis

Our unit of analysis was the task. The first stage included locating discrepancies between a correct choice and the accompanying interactive example. We checked which participants chose the set of correct statements and compared it with the number of participants to correctly attach examples for all of the statements. The second step included a refinement of the analysis. We counted how many correct statements were chosen per-task (more than one choice could be correct for a single task), and compared it with the number of incorrect examples that do not meet the required answers' conditions. The third stage included the coding of the discrepancies according to pre-set categories (e. g. *familiar mistakes* or *additional reasoning*) in order to study the characteristics which could be subsequently assessed automatically.

## Results

Table 1 shows the distribution of answers (consisting of chosen statements and supporting examples submitted by the participants (n=32) to the 3 different tasks.

Task (number of correct statements)	No. of participants which submitted an answer (n=32)	Sum of correct statements chosen by submitting participants	Sum of correct examples attached to correct chosen statements	No. of participants with all correct statements chosen (n=32)	No. of participants with correct answers and correct supporting examples. (n=32)
1 (3)	30	66	49	13	5
2 (2)	28	40	32	11	7
3 (3)	21	39	27	7	5
Total	N.R*	145	108	N. R.*	N. R.*

\* N. R = Not relevant

**Table 1: Answers submitted, statement choices, and examples provided for 3 tasks**

As can be seen in Table 1, a total of 145 correct statements were chosen and submitted. For 108 of them (74.5%) correct examples were submitted. The remainder (25.5% of the correct choices) were submitted with incorrect or no examples. We now investigate the work related to two statements of task 3 in order to learn the nature of the examples that did not seem to be coherent with the choice of statement. In this task (Figure 1), the topic is the recognition of similar triangles, and ratios between areas of similar triangles. The context of the dynamic figure is presented to the participants in multiple representations: a verbal description in the task description, starting with: "point D is the midpoint ...", a symbolic representation in the digital geometry environment (DGE):  $ED \perp AB$ ,

$AD=DC$ , and a DGE construction: A draggable triangle ABC with point D and E. Measurement tools and numerical feedback are not supported in this task.

In the dynamic figure below, point D is the midpoint of segment AC in triangle ABC. ED is a line segment perpendicular to side AB of triangle ABC. Explore the dynamic figure considering the ratio between the areas of triangles ABC and ADE.

Below are three statements. For each statement for which you have a supporting example, please save the example as a way to show that you think that this statement is true. Leave the statements that you do not think are true as they are.

1  $S(ABC):S(ADE)=4:1$

2  $S(ABC):S(ADE)=2:1$

3  $S(ABC):S(ADE)=k:1 \quad k>10$

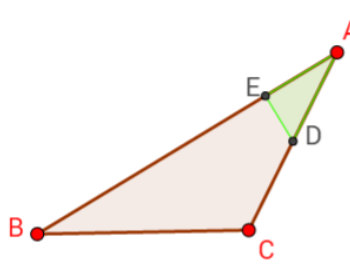
Save

Save

Save

---

$ED \perp AB$   
 $AD = DC$



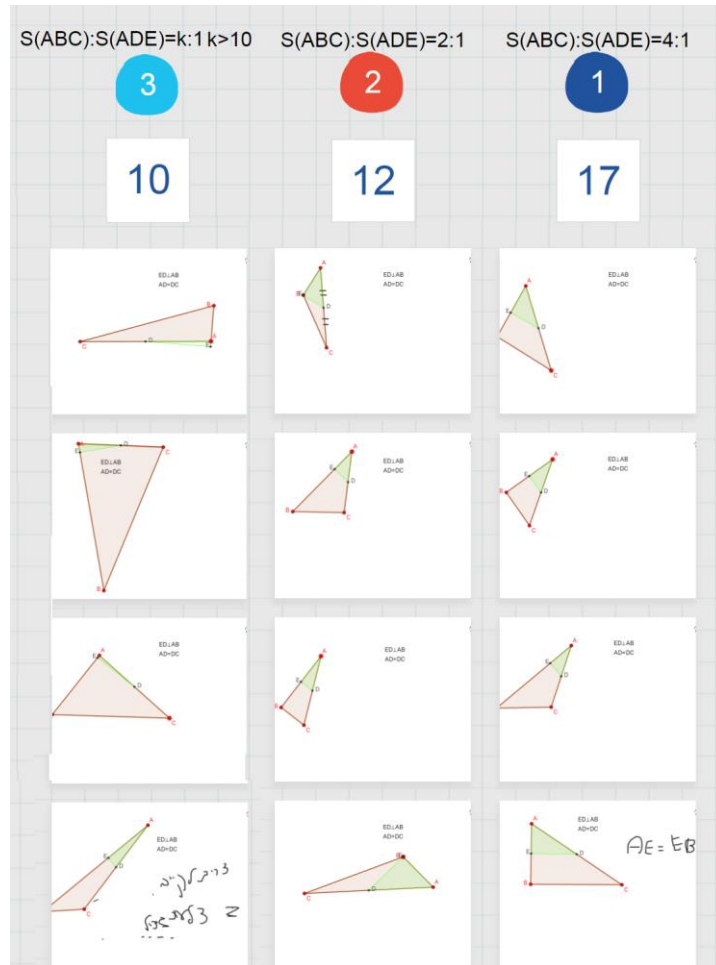
**Figure 1: Multiple choice with supporting example task**

In terms of correctness, the red points in the diagram could be dragged to create an example for any of the three statements in this task, making all the choices potentially correct ones. In order to construct a supporting example for the first statement, the lines ED and BC should be parallel<sup>3</sup>, in order to construct a supporting example for the second statement points E and B need to coincide. A supporting example for the third statement would be any position where  $AB > 5AE$ . But in order to construct such examples (mostly for statements 1 and 2) participants are required to have some understanding regarding similarity and ratios of areas of similar triangles.

Automatic assessment of this task was performed with the STEP platform, which is designed to present the submitted examples in several representations, including a visual representation of all examples attached to each of the statements (Figure 2), while enabling the assessor to automatically

<sup>3</sup> There is also another option to construct this where E is outside ABC and  $AE=AB$ .

filter the results according to mathematical and pedagogical criteria (Olsher, Yerushalmy, and Chazan, 2016), as will be demonstrated for this task.



**Figure 2: A sample of supporting examples for statements presented on the STEP platform**

### **Incorrect examples that do not meet the required answers' conditions**

Analyzing the collection of examples per-statement suggest a finer categorization. Statement 3-1 (the first statement in task 3) stated that the ratio between the area of triangle ABC and the area of triangle ADE is 4:1. Triangles ABC and ADE are similar as ED is perpendicular to AB. In addition, AD has the same length as DC. Thus, any example in which ED is parallel to BC, which means AE has an equal length to EB and vice versa provides a supporting example. There were a total of 17 examples submitted for this statement. 13 of which met the criteria for correctness. In Figure 3 appear the 4 submitted examples that were automatically marked as incorrect, as the automatically calculated ratio between the relevant triangle areas was not approximately 4:1.

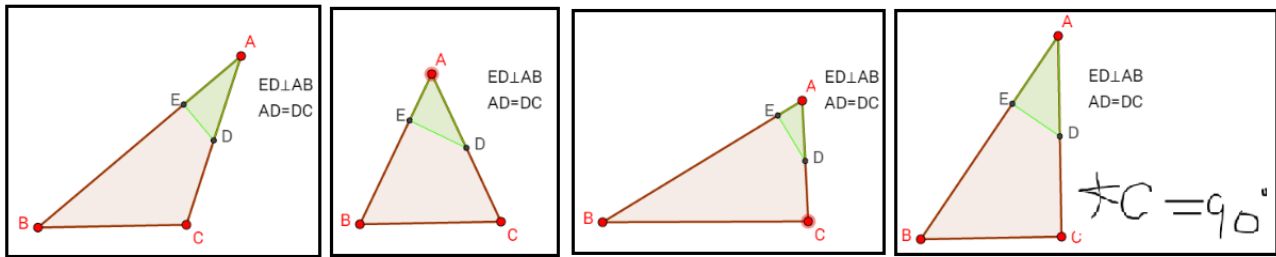


Figure 3: Incorrect submitted examples for task 3-1

The main characteristic that could be automatically assessed with this representation is the possibility that these participants did not address the characteristics relevant for this statement in their submitted examples - ED is not even approximately parallel to BC.

### Incorrect examples in line with familiar student mistakes

Statement 3-2 stated that the ratio between the area of triangle ABC and the area of triangle ADE is 2:1. Triangle ABC and ADE are similar. Thus, any example in which points E and B coincide provides a supporting example. There were a total of 12 examples submitted for this statement. 8 of which met the criteria for correctness. In Figure 4 appear the 4 submitted examples (a, b, c and d from left to right) that were automatically marked as incorrect, as the automatically calculated ratio between the relevant triangle areas was not approximately 2:1.

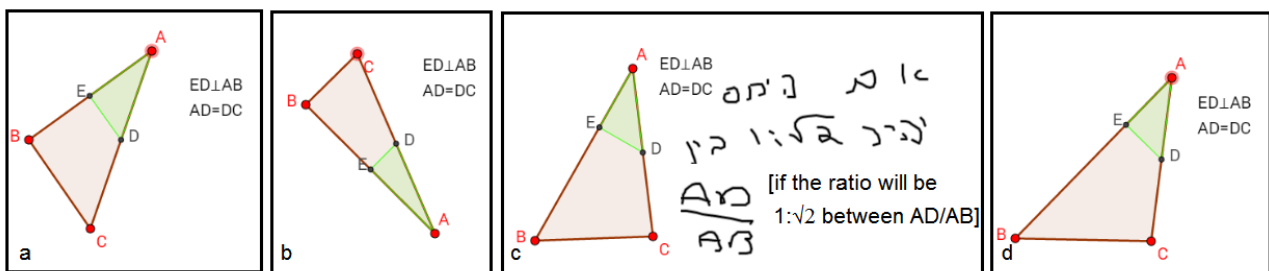


Figure 4: Incorrect submitted examples for task 3-2

In Figure 4, we see two incorrect examples (4a, 4b) that were further automatically categorized as "familiar mistakes". In these examples, the ratio between the lengths of BC and DE is approximately 2:1. These examples are a possible indication of holding the image of "linearity" between ratio of lengths and ratio of areas, a familiar phenomenon from the literature and teacher practice.

### Examples with additional verbal, symbolic or free-hand graphic reasoning

Figure 4c shows an example that includes a correct verbal explanation but without any matching change to the dynamic diagram. One of the functionalities of the STEP platform is a free-hand pen tool which can be used for making annotations and marks, or any other use that the participant might find helpful. The participants were aware that the platform automatically checked their figures, and that text or graphical markings, if submitted, are presented for the teacher to review but not analyzed automatically. The example above is one of 18 submitted examples (across all three tasks in this study) that were accompanied with free-hand markings on the interactive diagram. Apart from verbal explanations, the examples also included annotations in the form of either symbolic writing or in graphical marks on the diagram. There are numerous possible explanations

for such responses. The participant might not have been able to construct the example, but thought about its mathematical properties, and wanted to demonstrate his knowledge. Annotation could also indicate that the participants needed to justify their example in a more robust, mathematical fashion, not fully accepting the diagrammatic example alone as a valid justification for a statement, which is closer to Israeli standard classroom practice.

The example in figure 4d was not automatically categorized beyond its' incorrectness, as it did not fit the predefined filter for a familiar student mistake.

One other aspect of the automatic assessment of MC tasks is the correctness of the entire task (e. g. choosing all of correct statements *and* providing them with correct supporting examples). In this study 13, 11, and 7 participants made a correct choice of statements in the three tasks respectively (Table 1). The number of participants who chose both the relevant statements and also provided a correct corresponding supporting example is lower: 5 (of 13), 7 (11), and 5 (7) (Table 1). This might be because the tasks were not clear enough, not specific about the relevant conditions; or perhaps ill-defined in terms of level of accuracy required. In order to enable efficient formative assessment these analyses are presented to the assessor in various graphic (e. g. Figure 2) and analytic (e. g. Venn diagrams) representations for further investigation.

## Conclusions

This study provides initial information about discrepancies between choosing a correct statement, which could be a result of a guess or good examination tactics (Hassmen & Hunt, 1994; Lau, Lau, Hong, & Usop, 2011; Sangwin & Kocher, 2016), and providing an example to support this statement, which requires the translation of the conditions into a DGE context.

Many of the automatically assessed characteristics of submitted examples were not related to the correctness of the example in supporting the claim, but to other aspects such as student approaches and strategy (e.g. construction of prototypical figures, unexpected solutions). Although, due to space limitations, this report has focused on the limited analysis of discrepancies between chosen *correct* mathematical statements and their supporting examples, it has provided several additional insights into the MC tasks. Some of the solvers did not attend to significant characteristics required to support the answer (e. g. a line that needs to be a mid-section therefore to connect mid-points of two sides of the triangle and to be parallel to the third one), or the fact that they construct an example in line with a familiar student mistake (e. g. the ratio between the areas of similar triangles is the ratio of their sides squared, not the exact same ratio as reflected in the submitted example). These types of phenomena could help teachers better assess the performance level on these types of tasks, in the relevant mathematical topic, providing meaningful real-time analysis in the service of instruction. The automatic analysis and categorization alongside the visual representation methods played a key part in discussions of the results with teachers. This practice is well aligned with what Olsher et al. (2016) claim that the viability of this assessment in the mathematics classroom lies within the ability to automate the assessment process as much as possible, and to provide the teachers with suggestive insights as part of a better picture of their classroom example space and answers.



## Acknowledgment

We gratefully acknowledge funding by the Trump Foundation, Israel Science Foundation (522/13) and I-CORE Program of the Planning and Budgeting Committee.

## References

- Black, P. J., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice*, 5 (1), 7–74.
- Buchbinder, O., & Zaslavsky, O. (2009). A Holistic Approach for Designing Tasks that Capture and Enhance Mathematical Understanding of a Particular Topic: The Case of the Interplay between Examples and Proof. *Task Design in Mathematics Education. Proceedings of ICMI Study 22*, 25.
- Farrell, T., & Rushby, N. (2016). Assessment and learning technologies: An overview. *British Journal of Educational Technology*, 47(1), 106–120. doi:10.1111/bjet.12348
- Hassmen, P., & Hunt, D. P. (1994). Human self-assessment in multiple choice. *Journal of Educational Measurement*, 31, 149–160.
- Lau, P. N. K., Lau, S. H., Hong, K. S., & Usop, H. (2011). Guessing, Partial Knowledge, and Misconceptions in Multiple-Choice Tests. *Educational Technology & Society*, 14 (4), 99–110.
- Leung, A., & Lee, A. M. S. (2013). Students' geometrical perception on a task-based dynamic geometry platform. *Educational Studies in Mathematics*, 82(3), 361-377.
- Nagari-Haddif, G., & Yerushalmy, M. (2015). Digital interactive assessment in mathematics: The case of construction E-tasks. In K. Krainer & N. Vondrova *Proceedings of the Ninth Congress of the European Mathematical Society for Research in Mathematics Education* (pp. 2501-2508). Prague, Czech Republic: Charles University in Prague, Faculty of Education and ERME.
- Olsher, S., Yerushalmy, M., & Chazan, D. (2016). How might the use of technology in formative assessment support changes in mathematics teaching? *For the Learning of Mathematics*, 36(3), 11-18.
- Sangwin, C. J., & Kocher, N. (2016). Automation of mathematics examinations. *Computers & Education* 94, 215–227.
- Stacey, K., & Wiliam, D. (2012). Technology and assessment in mathematics. In M. A. K. Clements, A. Bishop, C. Keitel-Kreidt, J. Kilpatrick, & F. Koon-Shing Leung (Eds.), *Third international handbook of mathematics education* (pp. 721–751). New York, NY: Springer Science & Business Media.
- Vinner, S. (1983). Concept definition, concept image and the notion of function. *International Journal of Mathematical Education in Science and Technology*, 14(3), 293–305.
- Zazkis, R., & Leikin, R. (2007). Generating examples: From pedagogical tool to a research tool. *For the learning of mathematics*. 27(2), 15-21.