Validity of problem posing as assessment

Rebecca Kuijpers

6435696

Freudenthal Institute, Utrecht University

RP30: Research Project

Dr. Rogier Bos

June 27, 2023

Utrecht

Abstract

This article aims to explore the construct validity of problem posing as an assessment method. The article proposes a framework of factors contributing to the validity of problem posing as assessment. For convergent validity these factors are the complexity of the problem, solvability of the problem and the coverage of the learning goals. For discriminant validity there is only one factor, namely whether the problem covers at least one learning goal. Further, this article explores whether it makes a difference for the validity whether the problem posing task is structured or unstructured. We have analyzed 86 problems posed by 21 secondary education students as part of an advanced mathematics course. The results showed that all previously mentioned factors contribute to the validity of problem posing as an assessment and that taking the students' answer key into account in assessment contributed to the convergent validity as well. Therefore, the answer key is added to the proposed framework of factors contributing to the validity of problem posing. Furthermore, the problems posed from structured tasks had more diversity in complexity of the problems, allowing the teacher to distinguish the mathematical abilities of students better. Therefore, the structured tasks are more convergent valid than the unstructured tasks. The main implication of the results is that problems posed from structured tasks that are rated on complexity, solvability, students' answer keys and coverage of the learning goals are a valid way of assessment.

Validity of problem posing as assessment

In the last decades, teaching has shifted from teacher-centered to student-centered forms of teaching. The attention to the individual needs of every student has increased, which results in changes in education (Weimer, 2002; Wright, 2011). As a part of this development, student-centered assessment has received more attention (see for example Stiggins, 1994; Pedersen & Williams, 2004). While Stiggins remarks that traditional tests can be used in a student-centered classroom, teachers need to take care in determining how their assessment best displays all of the students' capabilities. He names multiple possible types of assessment, namely essay assessment, performance assessment and personal communication assessment. These types of assessments could benefit students that have trouble with the traditional assessment, but they mostly seem to give the teacher a lot of work in their already overcrowded schedule (Collinson & Fedoruk Cook, 2001; Hargreaves, 1992). Not only is the teachers' workload a problem, in subjects like mathematics it is difficult to not assess traditionally as it is the most commonly used method with which problem solving is evaluated. Therefore, it is desirable to investigate assessment methods that are student-centered, are manageable for the teacher and are suitable for mathematics.

In mathematics assessment there are multiple, conflicting interests. On the one hand, it is important that a student can showcase their mathematical abilities. A part of these abilities is mathematical creativity (Krathwohl, 2002), which is particularly difficult to show in a traditional assessment consisting of a set of problems with a predetermined answer key. Traditional assessment normally consists of a set of problems that need to be solved to check whether learning goals have been reached. Occasionally traditional assessments offer some room for creativity, but usually the problems aim at a pre-set correct answer.

On the other hand, teachers want to assess students' mathematical abilities as accurately as possible, with as little work for themselves as possible. Both interests need to be considered when searching for a student-centered assessment method.

Such a method of assessment could be the activity of problem posing. This activity, where students pose problems based on a given context, is a good way to enhance and test students' problem-solving skills (Suarsana et al., 2019; Arikan & Ünal, 2015). It combines both the interests of students and teachers. It allows students to showcase their mathematical abilities, especially their creativity. Students are not limited to the problem their teacher determined for them, but instead get to showcase their strengths and creativity in crafting the problems themselves. Problem posing is known to affect students' attitude and mathematics achievement positively. At the same time, the problems stay within a given context that the teacher designed. This gives the teacher an opportunity to steer the direction in which students design their problems. Because all problems are based on the same context and each problem has limited written text, correcting the problems will require relatively little effort from the teacher. Therefore, problem posing might be the solution to the needs of the student-centered assessment method.

To use problem posing as an assessment method, however, there needs to be evidence supporting its quality. An important part of the quality is the validity of the assessment. If an assessment is valid, it measures what it is supposed to (Kane, 2006). Validity is regarded as one of the most important characteristics of an assessment (Borsboom, Mellenbergh & Van Heerden, 2004; Heale & Twycross, 2015). A foundation on the use of problem posing as assessment has been laid by Kwek (2015) and Mishra & Iyer (2015), amongst others. We believe it to be a very important addition to existing literature to consider the validity of problem posing as an assessment method. The aim of this research is therefore to explore the validity of problem posing as an assessment method.

Theoretical Framework

Assessment

Assessment is a classroom practice that covers a wide range of activities, that should all satisfy the following five principles (Pegg, 2003):

1. The quality of students' understanding and learning should be determined.

2. Assessment should provide material to describe students' knowledge and skills.
3. Assessment should be aligned with the learning goals.
4. Assessment should be aligned to the teaching process and even influence the teaching process.
5. There should be a theoretical framework for an assessment practice.

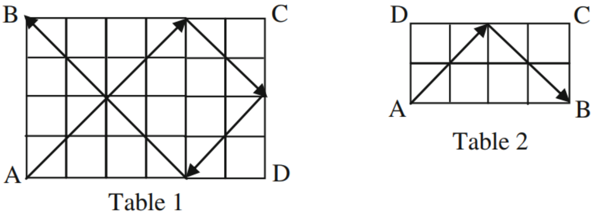
The first two points give an overview of the goal of assessment. The third and fourth point construct what we call constructive alignment (Biggs, 1996). The fifth point calls for a theoretical underpinning of an assessment method, which is the aim of this research regarding problem posing.

Problem Posing

As the name suggests, problem posing is an activity where students pose problems based on a given task. Table 1 provides examples of such tasks. While formulated differently, each task asks students to pose problems, with or without an initially given problem. Benefits of problem posing include a heightened level of student reasoning and reflection, more engagement, and improved problem-solving skills (Cunningham, 2004; Rosli, Capraro & Capraro, 2014). It is an activity that elicits evidence of students' understanding (William & Thompson, 2007).

In 1996, Stoyanova and Ellerton started to differentiate problem-posing situations as free, semi-structured and structured. In a free situation, a student is asked to pose problems starting from a given, naturalistic, or constructed situation without restrictions. In a semi-structured situation, a student is given an open situation to explore by using mathematical knowledge and previously learned skills. In structured situations, students are given a problem, after which they are invited to pose more problems on the same situation. Baumanns and Rott (2020) found that it is very difficult to distinguish between free and semi-structured situations, thus suggesting combining these types of situations as unstructured situations. In Table 1 the first task is a structured task, while the second is an unstructured task. When using problem posing as assessment, it might be important if structured

Table 1. Examples of problem-posing tasks

1	Consider the sequence $1, 2, 3, 4, 5, \ldots, \mathrm{N}$. If $\mathrm{N} = 200$, how many digits have been used? Other questions? (Stoyanova, 1999, p.34)
2	Imagine billiard ball tables like the ones shown below. Suppose a ball is shot at a 45 angle from the lower left corner (A) of the table. When the ball hits a side of the table, it bounces off at a 45 angle. In Table 1, the ball travels on a 4 6 table and ends up in pocket B, after 3 hits on the sides. In Table 2, the ball travels on a 2 4 table and ends up in pocket B, after 1 hit on the side. In each of the figures shown below, the ball hits the sides several times and then eventually lands in a corner pocket.  Table 1 Table 2 Based on the given situation, pose as many interesting mathematical problems as you can (Kontorovich, Koichu, Leikin & Berman, 2012, p.154)

or unstructured problems are used as they could give entirely different results from the problems posed.

Problem posing aligns with Pegg's principles. It provides material to describe students' knowledge and skills. We know posed problems can be used to determine students' understanding and learning. Whether problem posing aligns with constructive alignment is not yet entirely clear. While it is straightforward to implement problem posing activities in the classroom, it has yet to be determined if the posed problems align with the learning goals. Some researchers have addressed problem posing as an assessment practice from a theoretical point of view as well as a practical (see for example Kwek, 2015; Mishra & Iyer, 2015; Silver & Cai, 2005). We would like to extend these theoretical views to include the validity of problem posing as an assessment method.

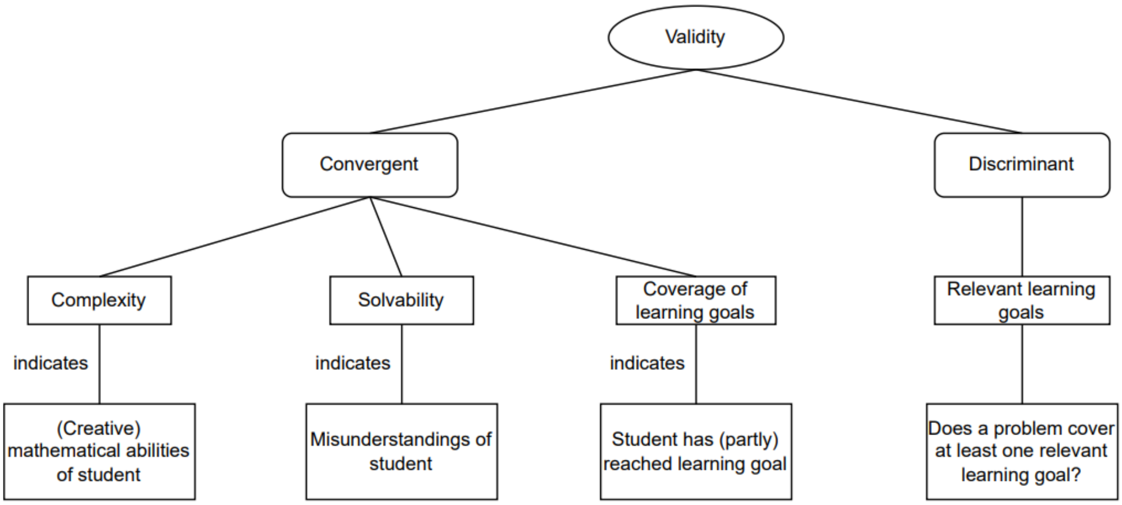
Validity

Kane (2006) defines validation "as the process of evaluating the plausibility of proposed interpretations and uses" and validity "as the extent to which the evidence supports or refutes the proposed interpretations and uses" (p. 17). While several subcategories of validity were defined in the Testing Standards in 1966, the category called construct validity became the one dominantly researched. A construct is a theoretical concept defined by its context (Colliver, Conlee & Verhulst, 2012). In the case of mathematics assessment, the construct is mathematical ability. Construct validity then is composed of numerous factors which determine whether the test measures the construct (Messick, 1992). Sjøberg and Bergerson (2022) note that two factors are especially important and combine several other factors, namely convergent and discriminant validity. Convergent validity "investigates the extent to which the results yielded by two instruments that measure similar concepts converge" (p. 1375) and discriminant validity "investigates the extent to which a set of indicators represents one specific concept only and no other" (p. 1375).

Validity of problem posing as assessment

We now know that evaluating the validity of problem posing as an assessment is important, so an approach to assessing this validity is needed. We propose a model where convergent validity consists of the factors complexity, solvability and coverage of the learning goals and discriminant validity consists of the factor relevant learning goals (see Figure 1). For this assessment to be construct valid, students' mathematical abilities and learning should be made visible through

Figure 1. *Proposed framework for validity of problem posing as assessment*


Validity
Convergent
Discriminant
Complexity
Solvability
Coverage of learning goals
Relevant learning goals
indicates
indicates
indicates
(Creative) mathematical abilities of student
Misunderstandings of student
Student has (partly) reached learning goal
Does a problem cover at least one relevant learning goal?

problem posing. If problem posing is convergent valid, the results should align with traditional assessment. Traditional assessment is often based on a taxonomy like Bloom's (Krathwohl, 2002). Problem posing could be used to assess the conceptual levels of bloom, namely analyze, evaluate, and create. It is less fit for the procedural levels, as it is difficult to check procedural skills when students pose problems rather than solve them. To measure the conceptual levels of Bloom, one needs to check the complexity of the problems and whether they are solvable. Complexity indicates the level of creativity and mathematical abilities a student possesses regarding a learning goal, which are both important parts of traditional assessment as well. Solvability can indicate students' misunderstandings, like mistakes would in traditional assessment. Further, in traditional assessment it is important that all learning goals are addressed. Checking whether all learning goals are covered by a problem therefore contributes to the validity.

Discriminant validity of problem posing as an assessment method becomes visible when looking at whether a problem covers one or more of the learning goals. If the problem does not cover any learning goal belonging to the lesson, the problem is on a different mathematics subject. Because the assessment then measures an unrelated construct, it would not be discriminant valid. Of course, there is a difference if a student does not cover any learning goal, and thus is on a different mathematics subject, and if a student connects a learning goal with a different mathematics subject. In the last case, discriminant validity would be supported, because at least one learning goal is covered and the connection with a different mathematics subject would show mathematical creativity and abilities. This means the last case would even support the convergent validity.

Research Question

To summarize, problem posing is an activity that can be used to assess students' mathematical abilities, creativity, and misunderstandings. It can determine the quality of students' understanding and learning and the problems are material that describe students' knowledge and skills. Problem posing therefore aligns with the first principles of Pegg. However, before using it in a classroom setting as an assessment method, it needs to have a theoretical foundation according to the fifth and last of Pegg's principles. We aim to add to this foundation theoretically as well as practically by researching the following research question:

How can problem posing be used as a construct valid assessment?

This is a broad question that needs further specification, for which three subquestions have been formulated. We have seen that important factors of problem posing as assessment are the complexity and solvability of the problem, together with the coverage of the learning goals. To assess the validity of problem posing as an assessment we will need to research how these factors hold up against traditional assessment. We have also seen that construct validity can be divided into convergent and divergent validity. Therefore, the first two subquestions are:

How do complexity, solvability and coverage of the learning goals contribute to the convergent validity of problem posing as assessment?

How does the coverage of the learning goals influence the discriminant validity of problem posing as assessment?

Finally, we know that problem posing tasks can be divided into structured and unstructured tasks. Because the type and formulation of the task could impact the validity, we would like to know what the differences are in validity between structured and unstructured tasks. Therefore, the third subquestion is:

What are the differences in construct validity between structured and unstructured problem posing tasks as assessment?

Methods

Setting

The study was performed in the context of Wiskunde D Online, a Dutch nationwide online program on advanced mathematics for high achieving secondary school students. As part of this program, students have access to knowledge clips and an online mathematics textbook. Furthermore, they participate in a weekly onsite class for one hour with a mathematics teacher and are invited to hand in their answers to a weekly set of tasks on which they receive feedback from mathematics students from several universities across the Netherlands. These hand-in tasks are not mandatory. This setting ensured that we had access to a mixed group of students, with their only similarity being that they are generally high achieving at mathematics.

Participants

In total 275 students aged 15 to 16 are enrolled in the Wiskunde D Online program. Because the hand-in tasks are not mandatory, not all students work on the chapter on normal distributions when the data collection takes place. Further, not all students will be willing to participate in the research. Thus, we do not expect all 275 students to participate. Because of the sample size in earlier classroom studies on problem posing (see for example Kwek, 2015 and Stoyanova, 1997), as this study is similar in data collection to those, we would like 20 to 30 students to participate.

Materials

To facilitate data collection, the four hand in exercises from the chapter were replaced by problem posing tasks. The tasks existed of a context and a prompt. For both structured and unstructured tasks, the context was identical, but the prompt differed, as can be seen in Table 2[1].

[1] While there were initially four different contexts set up for this research, only two of them were used in the analyzed problems as will be explained in the Results section. These are also the contexts shown in Table 2.

Table 2. Examples of problem posing exercises

<table>
<tr><th rowspan="2">Context</th><th colspan="2">Prompt</th></tr>
<tr><th>Structured</th><th>Unstructured</th></tr>
<tr><td>A random variable X has the following distribution:<table><tr><td>X</td><td>0</td><td>1</td><td>2</td></tr><tr><td>$P(X = x)$</td><td>$\frac{1}{2} - \frac{p}{2}$</td><td>p</td><td>$\frac{1}{2} - \frac{p}{2}$</td></tr></table>Where $0 \leq p \leq 1$.</td><td>A Compute the standard deviation in terms of p.
B Pose two more problems on this distribution. Also make the answer key.</td><td>Pose three problems on this distribution. Also make the answer key.</td></tr>
<tr><td>A factory produces blue, green, and red soaps. The weight of a soap is normally distributed, where $\mu = 100g$ and $\sigma = 3g$ for blue soaps, $\mu = 120g$ and $\sigma = 4g$ for red soaps and $\mu = 80g$ and $\sigma = 3g$ for green soaps. The volume of a soap is normally distributed, where $\mu = 0{,}2L$ and $\sigma = 0{,}002L$ for blue soaps, $\mu = 0{,}25L$ and $\sigma = 0{,}003L$ for red soaps and $\mu = 0{,}18L$ and $\sigma = 0{,}003$ for green soaps. The factory sells blue soaps for €1,-, red soaps for €1,50 and green soaps for €0,85. The amount of sold soaps per day is normally distributed, where $\mu = 40$ and $\sigma = 3$ for blue soaps, $\mu = 35$ and $\sigma = 2{,}5$ for red soaps and $\mu = 40$ and $\sigma = 2{,}5$ for green soaps.</td><td>A Compute the probability that the volume of a blue soap is less than 0,24L or more than 0,26L.
B Pose two more problems on this distribution. Also make the answer key.</td><td>Pose three problems on this distribution. Also make the answer key.</td></tr>
</table>

Students were asked to do the task with either the structured or unstructured prompt, based on whether their birthday was an even or odd date.

Data analysis

The gathered data was analyzed in three ways, as discussed in the theoretical framework. First, the problems were rated on complexity, then on solvability and after that on whether the learning goals were covered by the problems.

Complexity

The basis for determining the complexity of problems is a table of Kwek (2015) (see Table 3) which we adapted slightly. In this table, a problem is of low complexity when the answer is based on a specified procedure. However, a problem is of moderate complexity when the solution requires multiple steps. Because some specified procedures require multiple solution steps, it was unclear whether these problems should be rated as low or moderate complexity. That is why it was added that low complexity problems are routine problems, while moderate and high complexity problems are non-routine problems. A routine problem is a problem that is known to the solver. They have seen a similar problem multiple times and know how to solve it procedurally. A non-routine problem is not necessarily a difficult problem, but it needs to have a twist which means that a solver does not directly know how to solve it (Baumann & Rott, 2021). In this case, a routine problem is a problem

Table 3. Complexity of a posed problem

	Low complexity	**Moderate complexity**	**High complexity**
Description	This category relies heavily on the recall and recognition of previously-learned concepts. Item typically specify what the solver is to do, which is often to carry out some procedure that can be performed mechanically. It leaves little room for creative solutions. The following are some, but not all, of the demands that items in the low-complexity category might make:	Items in the moderate-complexity category involve more flexibility of thinking and choice among alternatives than do those in the low-complexity category. They require responses that may go beyond the conventional approach, or require multiple steps. The solver is expected to decide what to do, using informal methods of reasoning and problem-solving strategies. The following illustrate some of the demands that items of moderate complexity might make:	High-complexity items make heavy demands on solver, who must engage in more abstract reasoning, planning, analysis, judgement, and creative thought. A satisfactory response to the item requires that the solver think in an abstract and sophisticated way. The following illustrate some of the demands that items of high complexity might make:
Cognitive demand	• Recall or recognize a fact, term, or property • Compute a sum, difference, product, or quotient • Perform a specified procedure, **i.e., the problem is a routine problem** • Solve a one-step word problem • Retrieve information from a graph, table, or figure	• Represent a situation mathematically in more than one way • Provide a justification for steps in a solution process • Interpret a visual representation • Solve a multiple-step problem **where the problem is a non-routine problem** • Extend a pattern • Retrieve information from a graph, table, or figure and use it to solve a problem • Interpret a simple argument	•Describe how different representations can be used to solve the problem • Perform a procedure having multiple steps and multiple decision points • Generalize a pattern • Solve a problem in more than one way • Explain and justify a solution to a problem • Describe, compare, and contrast solution methods • Analyze the assumptions made in solution • Provide a mathematical justification

that the students have practiced extensively, such as computing the standard deviation, even though this procedure has quite a few solution steps. As Baumann and Rott (2021, p.34) said: “Integrating a polynomial function of degree 53 can be a tedious computational activity for a mathematician; however, since he or she knows the method of how to integrate a polynomial, this activity is a routine problem for him or her.” The procedure of integrating a polynomial function of degree 53 is one with many steps, but because it is a routine procedure, it would be classified as a low complexity problem. The additions of routine and non-routine problems are added in bold in Table 3.

Solvability

When assessing the solvability of a problem, two things need to be considered. Firstly, the problem situation description has to contain enough information to allow a student to solve the problem. Usually, this information can be found in the given context; sometimes the poser provides additional information. In any case, it should be clear what the problem requires the solver to do, and there should be enough information to solve the problem. Secondly, the problem statement, including additional context, needs to be mathematically correct. For example, a probability should always be a value greater than or equal to zero and smaller than or equal to one. When either in the problem, or in the solution, a value outside of these boundaries is found, the problem is not mathematically correct.

Learning goals

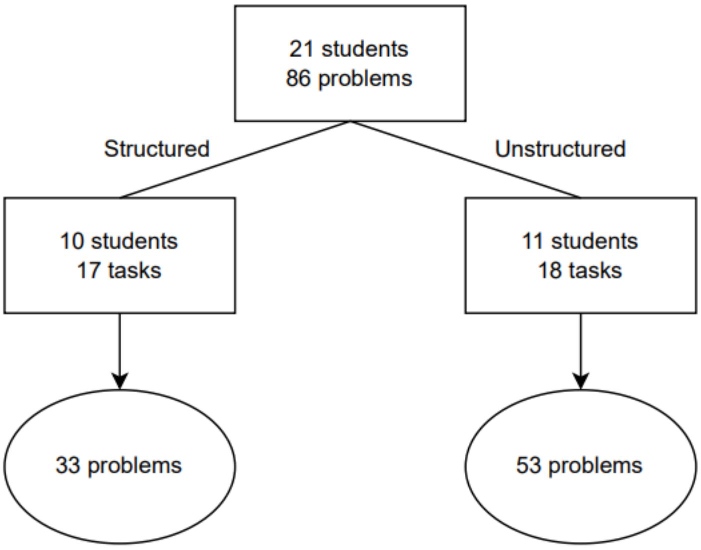
The last, and the least complicated of the rating criteria assesses whether the problem is tied to the learning goals. This is assessed by holding each problem against a list of learning goals that can be found in Appendix A.

Results

In total 275 students of age 15 to 16 are enrolled in the Wiskunde D Online program. Of these 275 students, 21 students participated in the hand in tasks. The tasks with structured prompts were handed in by 9 students, 11 students handed in the unstructured tasks. One student handed a mix of structured and unstructured tasks. Because most of these students' tasks were structured, he was put in the structured group overall. The students were supposed to individually make four separate problem posing tasks, but only seven students handed in the third exercise, of which only two were structured, and only three students handed in the fourth, of which one was structured. Because of the limited number of responses and the skewed distribution of structured and unstructured tasks of the third and fourth task, these tasks were not considered for the data analysis. Because not all students handed in all tasks, there are a total of 17 structured tasks, which gave 33 problems, and 18 unstructured tasks, which gave 53 problems to analyze. An overview can be found in Figure 2.

For assessing interrater agreement, 30 of 90 problems were coded by the author's supervisor, referred to as Rater 2. The codes on solvability and coverage of the learning problems were almost identical and all differences were caused by one of the raters overlooking a learning goal. Before discussion there was a Cohen's kappa (Wongpakaran, Wongpakaran, Wedding & Gwet, 2013) of 0,66 on the coding of complexity. This was mainly due to confusion about whether a route multiple-step-problem had a low or moderate complexity. The raters decided on adding routine and

Figure 2. Number of students and posed problems separated by structured and unstructured tasks



non-routine to the complexity matrix, as discussed in the Method section. This decision gave clarity on most differences, leading to a Cohen's kappa of 0,96.

Examples of coded problems

The students posed many different problems, three of which we will discuss here. The first problem was constructed as a response to the first context and asks the solver to compute the standard deviation when $p = 1/2$. The problem was coded as having a low complexity, because computing the standard deviation is a routine procedure for these students. The problem covers a learning goal, namely computing the standard deviation from a probability distribution. Furthermore, the problem is solvable, because when $p = 1/2$ is filled in in the probability distribution, the probabilities are respectively $1/4$, $1/2$ and $1/4$, as can be seen in Table 4. These chances are all values between zero and one and add up to 1, which means the probability distribution is correct. Furthermore, $0 \leq p \leq 1$, as specified in the context.

The second problem was constructed as a response to the second context and asks for the probability that the gain of the factory is more than €36, just from the green soap. It was coded as having a moderate complexity, because it is not a routine problem. While the computation is a standard one, the students first have to realize that the answer lies in the number of soaps sold, not in the price of the soap. They have to think one step further than the obvious solution. The learning goal that this problem covers is that the student can compute probabilities using the normal distribution, where the average value, standard deviation and boundaries are given. The problem is clearly formulated and solvable.

The third problem was also formulated to the second context and consists of two questions. The first question asked which color soap has a higher probability of being sold less than 39 times per day. The second question elaborated on the first and asked what the differences are in

Table 4. Probability distribution in a posed problem

X	0	1	2
$P(X = x)$	$1/4$	$1/2$	$1/4$

those probabilities per color. This problem was coded as having a moderate complexity as well, because the problem asks the solver to not only compute probabilities, which is a routine skill and therefore rated as low complexity, but also compare these probabilities, which is a skill learned in an earlier grade. The problem covers the same learning goal as the second problem.

The third problem is an unsolvable problem, because the student does not specify whether the difference between probabilities should be computed in percentage points, or relatively. While it is unusual to rate the complexity of unsolvable problems (see for example Silver & Cai, 1996; Ngah, Ismail, Tasir & Said, 2015), we chose to do so. This is because with nearly every, if not all, problems, the intention of author of the problem was clear as was the solution structure, even if it ended up being unsolvable. This was supported by the answer keys the students made for each problem. In the case of the third problem, the answer key shows the student intended percentage points and not the relative difference. The unsolvability was due to poor formulation. This case of poor formulation is more common and will be discussed in the learning goals section.

In the next paragraphs, the results on complexity, solvability and coverage of the learning goals will be elaborated on. Each part will start with a statistical analysis of the difference between structured and unstructured tasks. Then a qualitative analysis of the problems will follow, where not only the differences between structured and unstructured tasks will be considered, but also more general remarks about the factors of the validity will be made.

Complexity

In order to analyze the differences in complexity between structured and unstructured problem posing exercises, a Mann-Whitney U test was applied on the coded problems. This gave $M - W = 641, p = 0.048$, which means that there is a significant difference in complexity between problems posed from structured and unstructured prompts. Table 5 shows that the structured

Table 5. Relative frequencies of complexity for structured and unstructured exercises

	Complexity			Total
	Low	Moderate	High	
Structured	48.5%	36.4%	15.2%	100%
Unstructured	68.0%	30.0%	4.0%	100%

exercises had a relatively low percentage of low-complexity-problems and a relatively high percentage of high-complexity problems. It supports that problems posed from structured exercises are generally of a higher complexity than those posed from unstructured exercises.

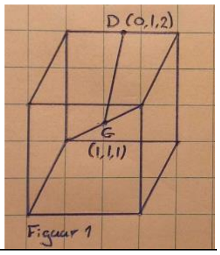
For the problems formulated to the first context it was striking that many of the problems posed from the unstructured task were steps in the calculation of computing the standard deviation. Table 6 shows an example of such a task. To compute the quadratic deviation, the expectation value of X is needed and in order to compute the variance, the quadratic deviation is needed. The standard deviation is computed by taking the square root of the variance, meaning that students who did the structured task performed all these steps already, before even posing a problem. Routine problems such as those in Table 6 were very rarely posed from a structured exercise.

Another notable remark concerns the complexity of the problems. The problems that were highly complex often combined earlier knowledge with the material the students had just learned. In table 7, the first example combines computations with the normal distribution and combinatorics and the second example combines the standard deviation and vector algebra. These problems ask for a deeper analysis and more creative thinking in order to be solved. However, other tasks that were supposed to be highly complex, such as comparing different solution strategies or representations, did not occur at all.

Table 6. An example of an unstructured task with routine problems

	Problems
1	Compute the expectation value of X
2	Compute the quadratic deviation
3	Compute the variance

Table 7. High complexity problems often combined earlier knowledge with the covered material

	Problems	
1	A box contains 20 soaps, of which 6 are blue, 7 are red and 7 are green. What is the probability of grabbing a red soap which weighs more than 125 grams two times in a row? (without replacing)	
2	Compute the standard deviation when $p = \frac{1}{3}$ using vectors and figure 1	

Solvability

To analyze the differences in solvability between structured and unstructured problem posing tasks a chi-square test was applied. The outcome was $p = 0.002$, which means that there is a significant difference in solvability between problems posed from structured and unstructured prompts. In Table 8, we can see that not even seventy percent of problems posed from structured exercises were solvable, and thirty percent were unsolvable. Therefore, the students that completed the structured tasks posed more unsolvable problems than those that completed the unstructured tasks.

The unsolvable problems could be divided into two groups, namely those problems that brought a misconception or misunderstanding of the material to light and those that were poorly formulated. The unsolvable problems that brought misconceptions forward are exemplified by problem 1 in Table 9. The student expanded on the given distribution from the first task. However, from this expansion it is clear that the student does not have a proper understanding of probability distributions. As discrete distributions are closed, the row from $H(x = x)$ should add up to 1, leading to $h = -3/2$.

Table 8. Relative frequency of solvability for structured and unstructured exercises

	Solvability		Total
	Solvable	Unsolvable	
Structured	69,7%	30,3%	100%
Unstructured	94,3%	5,7%	100%

Table 9. Two types of unsolvable problems posed by students

<table>
<tr><th></th><th>Problems</th></tr>
<tr><td>1</td><td>Stochastic variable has the following distribution
<table>
<tr><th>X</th><th>0</th><th>1</th><th>2</th><th>3</th></tr>
<tr><td>H(x = x)</td><td>1-h</td><td>h</td><td>h + 1</td><td>h + 2</td></tr>
</table>
Question = compute $x_i - E(x)^2$ for each x.</td></tr>
<tr><td>2</td><td>Compute the revenue of green soaps with a probability of 0,115069670222.</td></tr>
</table>

As all probabilities should be a value between 0 and 1, a probability of $-3/2$ is mathematically impossible and therefore the problem is unsolvable.

Whether problems were unsolvable due to poor formulation was also clear from answer key students composed. An example of such a problem is problem 2 in Table 9. The problem is to compute the revenue with a certain probability. The problem however, does not say what the probability applies to. It could be the probability that a soap is sold to a customer, or the probability that the green soap has a certain weight. Because the probability is not specified, the problem is unsolvable. However, the student that posed the problem made an answer key in which he used correct techniques to solve the problem. He knew what the probability referred to, he just did not write it down and thus he created an unsolvable problem because of poor formulation.

Learning goals

In order to analyze the differences in mathematical topic between structured and unstructured problem posing exercises, a chi-square test was applied. This resulted in $p = 0.315$, which means there is no significant difference between the number of posed problems that covered the learning goals in structured and unstructured problem posing exercises. The relative frequency of learning goal coverage is shown in Table 10.

Table 10. Relative frequency of learning goal coverage for structured and unstructured exercises

	Was at least one learning goal covered?		Total
	Yes	No	
Structured	97,0%	3,0%	100%
Unstructured	90,7%	9,3%	100%

Table 11. Problems posed by students that do (not) cover any learning goal

	Problems
1	Compute the probability that the weight of a red soap is more than 83 grams.
2	Compute how much money the factory makes on average per day.

Problems that were on different material than the lesson were mostly very simple questions, like problem 2 from Table 11. This indicates that these students did not want to put in effort in the task or did not reach the learning goals at all.

All learning goals, except one, were met in at least one of the problems. The learning goal that was not met stated that the student should be able to compute the standard deviation from a binomial distribution. The students did not get the opportunity to showcase this skill, because the context did not allow for it.

Conclusion and Discussion

First, the three factors that make up the case for validity, namely complexity, solvability, and coverage of the learning goals, will be discussed. For each factor, first its contribution to the validity will be discussed, after which more specifically will be zoomed into structured and unstructured tasks. This will allow us to answer the subquestions and finally the research question. To finish, we will discuss some limitations of the research.

Complexity

In the results section it became clear that students posed problems in all three categories of complexity. As was mentioned in the theoretical framework, the complexity of the problems tells us about the creativity with which a student can use their mathematical knowledge and abilities. A problem categorized as moderate or high complexity shows that a student not only has procedural understanding, but also conceptual. The problems categorized as low complexity from Table 6 only showed procedural understanding. The student had seen these problems many times before, they were routine problems with a step-by-step solution path. On the other hand, the problems from

Table 7 showed a conceptual understanding, where the concepts could even be connected to earlier learned knowledge and even a different field of mathematics. Looking at Blooms taxonomy, it is, indeed, true that the higher the complexity of the problem, the more they belong in the upper levels and the lower the complexity, the lower they are in the levels of Bloom. Of course, caution is important. When a student poses a problem, we never know if the student is using all of their mathematical abilities in the posing process, or if they do not want to put in effort and thus pose routine problems. The students that pose moderate or high complexity problems show promise of higher mathematical abilities, but we cannot say that students who pose problems with lower complexity do not have these abilities. We can only say that they don't show these abilities at that given time.

In the results section it became clear that when given a structured task, students pose more complex problems than when given an unstructured task. Several of the low complexity problems posed from the unstructured version of the first task, were a part of the process to compute the standard deviation. These problems did not appear in the problems posed from the structured task, as those students already had to compute the standard deviation. This suggests that the problem given in a structured task influences the complexity of the posed problems. Because the students with structured exercises were initially challenged more, it was more difficult for them to ask the obvious problems. Further research could investigate whether the starting problem of a structured task matters for the problems posed from the task.

Solvability

In the results section, it also became clear that students posed solvable, as well as unsolvable problems. That means that solvability, just like complexity, is a characteristic of a posed problem that allows a teacher to assess whether the student has reached a learning goal. In fact, in the case of an unsolvable problem, the teacher can signal specific misunderstandings, making the solvability a powerful indicator. It is an important part of determining whether the student has reached a learning goal.

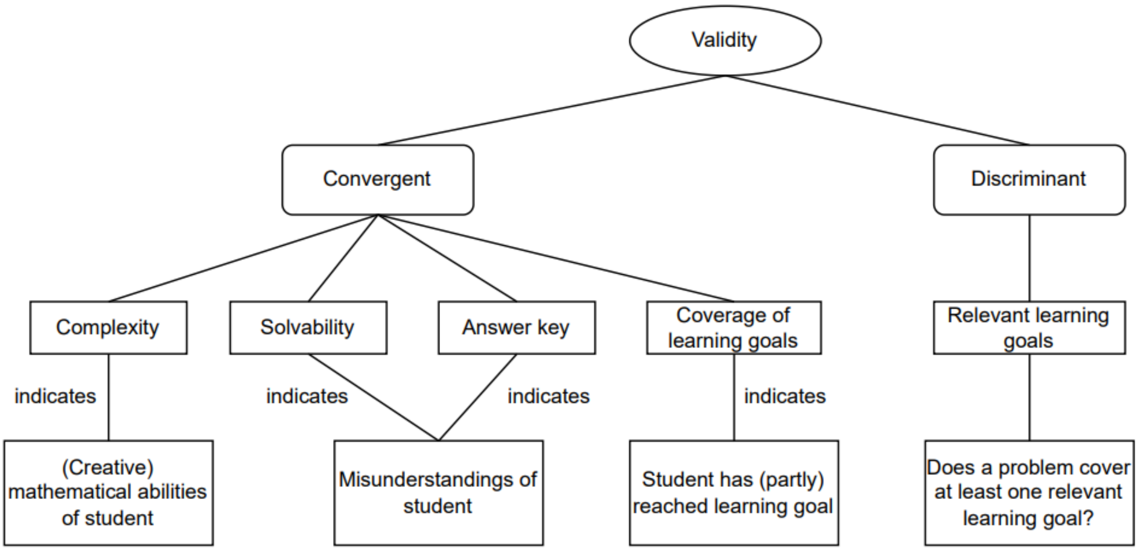
When taking in mind that problems posed from structured tasks have a higher complexity more often, it is not surprising that problems posed from structured tasks were more often unsolvable than those posed from unstructured tasks. As problems get more complex, they are more prone to contain mistakes. Routine problems, such as those in Table 7, are problems that the students have seen and solved multiple times. They could take a textbook problem as an example to formulate their own problem. Furthermore, there are no conditions that have to be mentioned in the problems, because they are so straightforward. In the more complex, non-routine problems on the other hand, there is almost always an extra condition or piece of context contained in the problem, creating possibilities for the problem to become unsolvable.

Several problems were unsolvable because of poor formulation, not poor understanding. This ill formulation is neither a new, nor an incidental problem. Amongst others, Ngah, Ismail, Tasir and Said (2015) and Cai and Silver (1996) also noticed this issue and recommended further research on the topic. In the case of this research, it became clear from the answer key of the problem that students did in fact have a proper understanding of the material. That means that in order to test whether and in what form the learning goals are met, the answer keys made by students are important. It allows the teacher to differentiate between students that have a misunderstanding and those that were sloppy in their formulation. That means that students creating an answer key makes for a higher validity of problem posing as assessment. Therefore, the answer keys have been added in the validity graph from Figure 1, see Figure 3.

Learning goals

All learning goals, except for one, were covered in multiple problems, advocating for the convergent validity with respect to coverage of the learning goals. As mentioned in the results section, this was mainly because the context did not allow students to pose a problem that covered

Figure 3. Evaluated proposed framework of the validity of problem posing as an assessment


Validity
Convergent
Discriminant
Complexity
Solvability
Answer key
Coverage of learning goals
Relevant learning goals
indicates
indicates
indicates
indicates
(Creative) mathematical abilities of student
Misunderstandings of student
Student has (partly) reached learning goal
Does a problem cover at least one relevant learning goal?

this learning goal. This emphasizes the importance of a carefully chosen context in the problem posing task. We recommend that teachers who design problem posing tasks are especially alert on giving students the opportunity through the context to pose a problem on every learning goal.

Lastly, almost all problems that were posed covered at least one learning goal. So, while this is an important part of assessing problem posing, this study did not found it to be a complicating factor. Further research could determine whether the students that posed problems that did not cover a learning goal did so because they did not want to put effort into the task, because they did not understand the material at all, or because of another unknown reason.

Conclusion

In conclusion, the complexity contributes to the convergent validity of problem posing as assessment because it shows the level of conceptual and procedural understanding and creative mathematics abilities, just as a traditional test would. The solvability contributes to the convergent validity, because it highlights misunderstandings a students might have. Together with the answer key it forms an even more powerful instrument, because then it can be determined whether a student really has a misunderstanding, or just poorly formulated the problem. Finally, just as in traditional assessment, all learning goals are covered, contributing to the convergent validity of problem posing as an assessment. We therefore conclude that the factors solvability, complexity, and coverage of the learning goals all contribute to the convergent validity of problem posing as an assessment. Further, we added the answer key to our proposed framework of validity, as shown in Figure 3.

As nearly all problems covered at least one learning goal, the assessment did not measure other mathematical fields, making it discriminant valid. However, we do need to note that some students mentioned they were taken aback because they had never formulated a problem before and some had formulation issues. As formulating a problem was not the construct that was meant to be assessed, the discriminant validity is compromised. However, Suarsana, Lestari and Mertasari (2019) have shown that students need to get used to the activity of problem posing. With sufficient practice, the students will get more comfortable with posing and this will not be an objection anymore. Students will then be able to solely focus on displaying their mathematical abilities. We have worked around the formulation issue by assessing the complexity, regardless of the formulation. We only remarked the poor formulation, but it did not effect the students grade.

There can be made more distinguishment between structured problem posing exercises than between unstructured exercises, both in complexity and solvability. Problems posed from structured exercises simply provide more information about the students' comprehension of the learning goal. Mishra and Iyer (2015) concluded that aligned to traditional assessment, results of assessment with problem posing exercises were similar for novelty learners, but not for advanced learners, based on a study with unstructured problem posing exercises. Further research would have to be done on whether that is also true for structured exercises. Because the structured task allowed for more distinction between students, it could be possible that the problem posing results are more similar to results from traditional assessment for advanced learners.

In conclusion, problem posing is convergent valid and discriminant valid, with some remarks and attention points, as mentioned above. This means that problem posing can be used as a valid assessment, using complexity, solvability, answer key and coverage of the learning goals as assessment factors for a structured task.

Limitations

This research had several limitations due to the circumstances of the Wiskunde D Online program. Because the program is an online, nationwide program, we did not know anything about the circumstances and behavior of the participating students. Because most students are from different schools, their previous knowledge of statistics differed and because there was no contact between the students and the researcher, their knowledge could not be measured beforehand. It is also unknown what is said and done in their live hour of teacher time and if they worked independently or collaboratively on the hand in exercises. For future studies it would be beneficial to be performed in a more controlled classroom setting.

Validity and reliability

The research was carried out carefully. The results aligned with other research, for example in the observation that unsolvability was often caused by poor formulation rather than misunderstanding. For the reliability one remark needs to be made about the participants, namely that they all voluntarily take part in the Wiskunde D Online course, meaning they are usually uncharacteristically motivated and highly achieving in mathematics. In other words, the participants likely do not represent an average mathematics classroom. This could impact the reliability of the research and it would be valuable to conduct this research again in a more traditional classroom setting.

References

American Psychological Association (1966). Standards for educational and psychological tests and manuals. Washington, DC.

Arikan, E. E., & Ünal, H. (2015). Investigation of problem-solving and problem-posing abilities of seventh-grade students. Educational Sciences: Theory & Practice. https://doi.org/10.12738/estp.2015.5.2678

Baumanns, L., & Rott, B. (2020). Rethinking problem-posing situations: A review. Investigations in Mathematics Learning, 13(2), 59–76. https://doi.org/10.1080/19477503.2020.1841501

Baumanns, L., & Rott, B. (2021). Developing a framework for characterising problem-posing activities: a review. Research in Mathematics Education, 24(1), 28–50. https://doi.org/10.1080/14794802.2021.1897036

Biggs, J. B. (1996). Enhancing teaching through constructive alignment. Higher Education, 32(3), 347–364. https://doi.org/10.1007/bf00138871

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. Psychological Review, 111(4), 1061–1071. https://doi.org/10.1037/0033-295x.111.4.1061

Collinson, V., & Cook, T. F. (2001). "I don't have enough time" - Teachers' interpretations of time as a key to learning and school change. Journal of Educational Administration, 39(3), 266–281. https://doi.org/10.1108/09578230110392884

Colliver, J. A., Conlee, M., & Verhulst, S. J. (2012). From test validity to construct validity . . . and back? Medical Education, 46(4), 366–371. https://doi.org/10.1111/j.1365-2923.2011.04194.x

Cunningham, R. K. (2004). Problem posing: An opportunity for increasing student responsibility. Mathematics and Computer Education, 38(1), 83-89. https://eric.ed.gov/?id=EJ720448

Hargreaves, A. (1992). Time and teachers' work: An analysis of the intensification thesis. Teachers College Record, 94(1), 87–108. https://eric.ed.gov/?id=EJ456574

Heale, R., & Twycross, A. (2015). Validity and reliability in quantitative studies. Evidence-Based Nursing, 18(3), 66–67. https://doi.org/10.1136/eb-2015-102129

Kane, M. J. (2006). Content-related validity evidence in test development. In Routledge eBooks. https://doi.org/10.4324/9780203874776.ch7

Kontorovich, I., Koichu, B., Leikin, R., & Berman, A. (2012). An exploratory framework for handling the complexity of mathematical problem-posing in small groups. The Journal of Mathematical Behavior, 31(1), 149–161. https://doi.org/10.1016/j.jmathb.2011.11.002

Krathwohl, D. R. (2002). A revision of Blooms taxonomy: An overview. Theory Into Practice, 41(4), 212–264. https://www.sid.ir/En/Journal/ViewPaper.aspx?ID=323367

Kwek, M. L. (2015). Using problem posing as a formative assessment tool. In Mathematical problem posing (pp. 273–292). https://doi.org/10.1007/978-1-4614-6258-3_13

Messick, S. (1992). The interplay of evidence and consequences in the validation of performance assessments. *ETS Research Report Series*, *1992*, 42(1). https://doi.org/10.1002/j.2333-8504.1992.tb01470.x

Mishra, S., & Iyer, S. (2015). An exploration of problem posing-based activities as an assessment tool and as an instructional strategy. Research and Practice in Technology Enhanced Learning, 10(1). https://doi.org/10.1007/s41039-015-0006-0

Ngah, N., Ismail, Z., Tasir, Z., & Said, M. F. M. (2016). Students’ ability in free, semi-structured and structured problem posing situations. Advanced Science Letters, 22(12), 4205–4208. https://doi.org/10.1166/asl.2016.8106

Pedersen, S., & Williams, D. (2004). A comparison of assessment practices and their effects on learning and motivation in a student-centered learning environment. Journal of Educational Multimedia and Hypermedia, 13(3), 283–306. https://www.learntechlib.org/p/11283/article_11283.pdf

Pegg, J. (2003). Assessment in mathematics: A developmental approach. In Mathematical cognition. IAP.

Rosli, R., Capraro, M. M., & Capraro, R. M. (2014). The effects of problem posing on student mathematical learning: A meta-analysis. International Education Studies, 7(13). https://doi.org/10.5539/ies.v7n13p227

Silver, E. A., & Cai, J. (1996). An analysis of arithmetic problem posing by middle school students. Journal for Research in Mathematics Education, 27(5), 521. https://doi.org/10.2307/749846

Silver, E. A., & Cai, J. (2005). Assessing students’ mathematical problem posing. Teaching Children Mathematics, 12(3), 129–135. https://doi.org/10.5951/tcm.12.3.0129

Sjøberg, D. I., & Bergersen, G. R. (2022). Construct validity in software engineering. IEEE Transactions on Software Engineering, 49(3), 1374–1396. https://doi.org/10.1109/tse.2022.3176725

Stiggins, R. J. (1994b). Student-centered classroom assessment. https://files.nwesd.org/depts/eadmin/Admin_Website/CIT-CL/LiteratureReference/JournalArticles/Student-Centered-Classroom-Assessment_Stiggins.pdf

Stoyanova, E., & Ellerton, N. F. (1996). A framework for research into students' problem posing in school mathematics. In P. C. Clarkson (Ed.), Technology in mathematics education (pp. 518–525). Melbourne, Mathematics Education Research Group of Australasia

Stoyanova, E. (1997). *Extending and exploring students' problem solving via problem posing*[Doctoral dissertation, Edith Cowen University]. Research Online Institutional Repository. https://ro.ecu.edu.au/theses/885/

Stoyanova, E. (1999). Extending students' problem solving via problem posing. The Australian Mathematics Teacher, 55(3), 29–35. https://eric.ed.gov/?id=EJ610477

Suarsana, I. M., Lestari, I. A. P. D., & Mertasari, N. M. S. (2019). The effect of online problem posing on students' problem-solving ability in mathematics. International Journal of Instruction, 12(1), 809–820. https://doi.org/10.29333/iji.2019.12152a

Weimer, M. (2013). Learner-centered teaching: five key changes to practice. John Wiley & Sons.

Wiliam, D., & Thompson, M. (2007). Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), The future of assessment: Shaping teaching and learning (pp. 53–82). Mahwah, NJ: Erlbaum.

Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K. L. (2013). A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC Medical Research Methodology, 13(1). https://doi.org/10.1186/1471-2288-13-61

Wright, G. (2011). Student-centered learning in higher education. The International Journal of Teaching and Learning in Higher Education, 23(1), 92–97. http://files.eric.ed.gov/fulltext/EJ938583.pdf

Appendix A: Learning goals

I can compute the expectation value.

I can compute the variance.

I can compute the standard deviation from a probability distribution.

I can compute the standard deviation from a binomial distribution.

I can compute the variance from multiple stochasts.

I can compute the standard deviation from multiple stochasts.

I can describe the normal distribution.

I can describe the meaning of the surface under the bell curve.

I can describe the meaning of the standard deviation in relation to the normal distribution.

I can compute probability in a normal distribution, given the mean, standard deviation, and boundaries.