ZDM – Mathematics Education
https://doi.org/10.1007/s11858-024-01556-0

ORIGINAL PAPER

Validity of multiple-choice digital formative assessment for assessing students' (mis)conceptions: evidence from a mixed-methods study in algebra

Katrin Klingbeil[1] · Fabian Rösken[1] · Bärbel Barzel[1] · Florian Schacht[1] · Kaye Stacey[2] · Vicki Steinle[2] · Daniel Thurm[3]

Accepted: 2 February 2024

Abstract
Assessing students' (mis)conceptions is a challenging task for teachers as well as for researchers. While individual assessment, for example through interviews, can provide deep insights into students' thinking, this is very time-consuming and therefore not feasible for whole classes or even larger settings. For those settings, automatically evaluated multiple-choice (MC) items could be a solution. However, it is a challenge to design those items and to adapt them for other countries in a way that they adequately reveal students' (mis)conceptions. In this article, we investigate the question whether it is valid to use a German adaption of a multiple-choice test developed in Australia for formative assessment of the *letter-as-object* misconception in Germany. For this, first semi-structured interviews with five German Year 8 students were conducted, and second, 616 students were asked for short written explanations. These data were analysed with regards to the students' (mis) conceptions and compared with their automatic online diagnosis. In general, a high concordance between online SMART test results and students' explanations was observed, confirming that useful diagnoses of student misconceptions can be obtained from such a short well-designed MC test.

Keywords Formative assessment · Online diagnostic · Algebra · Variables · Multiple-choice items

1 Introduction

With the rise of audience response systems and online diagnostic tools, multiple-choice (MC) items are popular as they are efficient to administer (Shin et al., 2019). However, to effectively improve learning, MC items must be of high quality, e.g., comprising of response options that are incorrect yet plausible (Haladyna et al., 2002). Ideally, these distractors contain misconceptions and mental images that are already known from research or result from deep content analyses. Thus, constructing insightful MC items for formative assessment (FA) as well as adapting these for use in other countries calls for an intensive cyclic design process including the gathering of evidence for the FA's validity.

In this article, we present such a process for the multiple-choice online test *Meaning of Letters*, which has initially been developed at the University of Melbourne, Australia, and is now being adapted for German-speaking countries at the University of Duisburg-Essen, Germany. To gather evidence for the validity of the adaption, a mixed-methods study has been conducted.

2 Theoretical background

2.1 Formative assessment with technology

"Formative assessment is considered one of the most effective frameworks to foster learning" (Schütze et al., 2018, p. 697), as it concretises how to realise student focus and adaptivity which has been empirically identified as a key quality dimension for effective teaching (Kunter et al., 2013; Praetorius et al., 2018; Prediger et al., 2022). The FA process "to recognize and respond to student learning in order to enhance that learning, during the learning" (Bell & Cowie, 2001, p. 540), requires an epistemic depth perspective to

✉ Katrin Klingbeil
katrin.klingbeil@uni-due.de

[1] University of Duisburg-Essen, Essen, Germany

[2] University of Melbourne, Melbourne, Australia

[3] University of Siegen, Siegen, Germany

Published online: 15 March 2024

provide tailored support and an optimal monitoring of students' needs and their learning stages. To make this concrete, crucial steps and strategies have been identified, including 1) the clarification of learning goals, 2) the collection and interpretation of information and 3) the use of this information to adapt supportive learning activities (e.g., Ruiz-Primo & Li, 2013). To be able to target specific needs, it is important to precisely specify the goals in Step 1. Narciss (2013, p. 12) highlights "a precise description of competencies" following a thorough task analysis according to the topic-specific cognitive demands as being crucial to realise a reliable FA process throughout the three steps (Shute, 2008).

Although FA can have a positive influence on student learning (Kingston & Nash, 2011), its effective implementation is practiced less frequently than ideal (Bennett, 2011; Yan & Pastore, 2022). The main reason for this could be the time-consuming nature of FA (Bürgermeister & Saalbach, 2018), especially of steps 2 and 3. Collecting adequately accurate information about the learning of all students and an appropriate adaptation is a major daily challenge for teachers. This is where technology-based FA can help. It can provide teachers with quick and reliable information about their students and further information about how to enhance the individual learning. The role of technology in this process is threefold (Cusi et al., accepted). Beside easier *communication* (through, with, or of) technology offers features for *analysing* students' data (overview work progress, solution frequencies, or advanced insights into students' thinking) and for *adapting* suitable material to enhance learning (passive, active, or intelligent). Passive adapting, e.g., means that tasks are offered from which the user can choose. It is important to note that the adapting level corresponds with the level of challenges for the teacher: The lowest demand for teachers is with intelligent adaptivity where material is provided based on a constantly updated student profile (Cusi et al., accepted).

There has been a growing recognition of the need for automated technological assessments that go beyond simply determining correct or incorrect answers in mathematics education, since most existing tests often lack a depiction of students' understanding, thereby limiting the usefulness for teachers regarding FA (Pellegrino & Quellmalz, 2010; Stacey et al., 2018). To address this challenge, SMART tests (Specific Mathematics Assessments that Reveal Thinking, http://www.smartvic.com) have been developed at the University of Melbourne, Australia. SMART tests employ rubrics and analysis techniques that examine not only the accuracy of responses but also response patterns and hence identify potential misconceptions underlying students' answers (advanced analysis) (Steinle et al., 2009). Subsequently, student diagnoses as well as teaching suggestions are provided to the teacher (passive adaptivity). Tests are short and tightly focussed, and their results are intended to be used for teaching within a short time frame. In this, SMART tests are distinct from other digital assessments that aim to cover a broader field, as for example *Pépite* for elementary algebra (Grugeon-Allys et al., 2018). To keep tests short, to make it easy and relatively error-free for students to enter answers (e.g., without formula editors), and to enable a quick automated analysis of student answers, SMART tests predominantly use multiple-choice (MC) items. Despite several known challenges of MC items, such as corrective feedback, working backwards, random guessing (Bridgeman, 1992), or the concern that the reasoning behind made choices is not assessed directly, SMART developers argue that MC items can be designed in a way to reveal students' potential thinking by automated analysis of the pattern in responses that students give to related test items since "in contrast to careless errors, misconceptions […] lead to predictable errors in student work" (Akhtar & Steinle, 2013, p. 36). More specifically, the SMART developers claim that the diagnosis offered by a short test can be sufficiently accurate to provide useful FA information to guide teachers' subsequent teaching. This article investigates this claim in one instance, a SMART test from the field of algebra called *Meaning of Letters*.

2.2 Meanings for algebraic letters

Algebraic conceptions are fundamental for mathematics classroom and there is a long tradition in research addressing this (e.g., Kaput, 1995; Kieran, 2007; Usiskin, 1988). Vergnaud (1996) emphasises the role of symbols for algebraic thinking: "The new thing with algebra for students is that it uses symbols and operations on symbols to calculate certain unknowns, without the need to control at every moment the meaning of the equations" (p. 231).

The SMART test *Meaning of Letters* examines students' most basic understanding of algebraic notation; what meaning do students give to the letters of the alphabet that are used to write algebra. We use the word 'letter' to denote the sign written on the page, to distinguish from the meanings that different students give the sign. Drawing on Serfati's (2005) epistemological approach to mathematical notation, written mathematics consists of signs, with three aspects: materiality (what is seen on the page, whether it is a numeral, letter, operator or other, its shape etc.), syntax (how a sign combines with other signs) and meaning. Meaning is understood by Serfati as that commonly agreed by the community of mathematicians – it does not refer to one person's individual understanding or interpretation – but for our educational work, we also include students' meanings that may be limited, erroneous or otherwise idiosyncratic. Following Drouhard and Teppo (2004) we use the word 'meaning' to refer to the type of mental entity associated by an individual with a sign, and 'understanding' more generally

to characterise how a student relates the sign and its meaning to a larger connected set of relationships.

Küchemann (1981) identified six different meanings for algebraic letters when doing early algebra tasks. At the lowest levels, students assigned an apparently arbitrary numerical value from the outset (e.g., replaced $h+8$ by 108 without evidence) or ignored letters (e.g., replaced $2h+8$ by 10) or considered the letter as standing for a concrete object rather than a number, perhaps an abbreviation for its name. Students at the higher levels considered the letter as standing for a specific unknown number, as a generalised number which could take multiple values, or as variable "representing a range of unspecified values and a systematic relationship is seen to exist between two such sets of values" (Küchemann, 1981, p. 104). For example, a variable understanding of h at least implicitly acknowledges a relationship between values of h and values of $2h+8$. These three higher level meanings of letters are those commonly used by the community of mathematicians (see e.g., Arcavi et al., 2017; Weigand et al., 2022), with different meanings coming to the fore in different contexts. In this article, we use the word 'letter' to indicate only the written symbol, to which different students will give different meanings including as an unknown number or variable. As Küchemann (1981) noted the "blanket use of the term 'variable' in generalised arithmetic is a common practice which has served to obscure both the meaning of the term itself and the very real differences in meaning that can be given to letters" (p. 110).

The empirical data presented here especially identifies the interpretation of letters as objects. Such an interpretation is especially evident in situations in which students deal with relationships between numbers of objects such as pencils or fruit where it is—from a mathematical point of view—"essential to distinguish between the objects themselves and their number" (Küchemann, 1981, p. 106). As will be evident in the discussion below, there are several varieties of the broad *letter-as-object* (*LO*) misconception, but in each case the algebraic letter is thought of not as a number, but as a reference to an object or an abbreviation of its name. For example, given the equation $3f=30$ about the total cost of €30 for some figs with f specified as the number of figs, some students will read it only as an abbreviation of "3 figs cost €30" (imagining 3 figs costing €10 each), rather than as "3 times the number of figs is equal to the number 30" (imagining 10 figs costing €3 each). As with many misconceptions in mathematics, the *LO* misconception is sometimes explicitly taught through inadequate instruction including in textbooks (MacGregor & Stacey, 1997), and it sometimes arises naturally when students read an algebraic sentence through the lens of natural spoken language and writing conventions (MacGregor & Stacey, 1993). The *LO* is a significant misconception especially because it leads students to make mistakes when formulating algebraic equations (e.g., for linear programming) and hence cuts them off from the benefits of being able to use algebra to solve problems.

3 SMART test *Meaning of Letters*

3.1 Development of the test

The reasons for creating a formative assessment test focussing on the *letter-as-object* (*LO*) misconception (Step 1 of FA) have been outlined above (2.2). This section focuses on Step 2, the development of the items, the diagnostic rules, and the design of the report to teachers. As with other SMART tests, the goal was to create a highly focussed short test, giving good information that can help teachers modify their teaching to better meet student needs. Thus, the SMART tests also include Step 3. The creation of the test items and the reports for teachers follows the process of design research, where "development and research take place through continuous cycles of design, enactment, analysis, and redesign" (Design-Based Research Collective, 2003).

The development of the *Meaning of Letters* test (initially called *Letters for numbers or objects?*) by the Australian team began with the algebra education research literature, especially drawing on items used by Küchemann (1981) and a series of local research projects (MacGregor & Stacey, 1993, 1997; Stacey & MacGregor, 2000) that investigated a wide range of elementary algebra items in open-ended (OE) pen-and-paper format as well as clinical interviews. Many items revealed aspects of the *LO* misconception and showed how students' misunderstandings of algebra often related to previous experiences of natural language and writing conventions. The writing convention that initial letters are often used as abbreviations is especially relevant here and needs to be carefully controlled in item development and translation (MacGregor & Stacey, 1997).

After initial trials of items, data analysis (Akhtar & Steinle, 2013) and minor improvements, the *Meaning of Letters* Version 1A consisted of three of the six items shown in Fig. 1: *Doughnuts* drawn from MacGregor and Stacey's work, *Garden* based on Küchemann (1981) and *Wheels* (as closely parallel to *Garden* as possible). A full parallel test (*Meaning of Letters* Version 1B) was also created, and student response data was analysed to see that the items matched very closely.

Akhtar and Steinle (2017) analysed responses from 1433 Australian students in Years 7, 8 and 9 to Version 1A. They found performance improved from Year 7 to 9, but the average facility for the three items reached only 40% for Year 9 students, underlining the importance of helping teachers address the *LO* misconception. Comparing responses to the

Doughnuts	*Garden*	*Biros*
Lucy bought 6 doughnuts for \$12. She wrote the equation $6d = 12$. In Lucy's equation, d stands for: • doughnuts • one doughnut • the cost of one doughnut • dollars	For my garden, I bought r red rose bushes and g white gardenia bushes. The roses cost \$4 each. The gardenias cost \$5 each. Choose the equation that says that the total cost was \$70: • $4r + 5g = 70$ • $10r + 6g = 70$ • $r + g = 70$	Biros are sold in packs of 3. Sam bought p packs and got b biros altogether. Choose the correct equation: • $b + p = 4$ • $p = 3b$ • $p = 3$ • $3p = b$ • $30b = 10p$
Lego	*Wheels*	*Racetrack*
Tina stacked 9 identically sized LEGO bricks on top of each other making a height of 99 mm. She wrote the equation $9y = 99$. In Tina's equation y stands for: • the height of one brick • the LEGO bricks in the tower • one LEGO brick • millimetres	At a bike shop there are b bikes (2 wheels) and t trikes (3 wheels). The equation that says that there is a total of 100 wheels is • $35b + 10t = 100$ • $b + t = 100$ • $2b + 3t = 100$	A car takes 12 minutes to drive round this racetrack. A driver drives r times around the racetrack in m minutes. Choose the correct equation. • $12m = r$ • $12r = m$ • $5r = 60m$ • $r = 12$

Fig. 1 The six items of *Meaning of Letters* (Version 2A)

closely parallel items *Garden* and *Wheels*, they found only 20% gave correct responses to both items, 22% gave one correct and one incorrect response, 52% gave exactly the same incorrect response to both items and 6% gave two different incorrect responses. The observation that a significant proportion (28%) of students do not select *LO* responses consistently (also noted by others, e.g., Warren, 1998) informed the diagnosis into stages given in Table 1. (Note that the diagnostic rule provided in this table is for the 6-item test (Version 2, discussed below) rather than Version 1, the 3-item test.) Stages of understanding for this test indicate how often students selected the *LO* responses. Students with a *LO* misconception are unlikely to be carefully considering the meaning of the letters in an algebraic equation, and so inconsistent behaviour can be expected, prompted by a variety of triggers. The report to teachers additionally flags a subtype of the *LO* misconception: *solution-as-coefficient* (*SAC*). When learning to formulate equations, some students believe they should incorporate a solution into an initial equation, rather than formulate an equation that describes the situation (Stacey & MacGregor, 2000) (example below).

Before the translation of *Meaning of Letters* into German, discussion between the research teams prompted the Australian team to further improve the test, creating Version 2. Since Version 1 was very short, the number of items could be increased from 3 to 6. This enabled a more reliable allocation to the stages (see Table 1) and importantly allowed for the inclusion of items with another algebraic structure (*Biros*, *Racetrack*). This created the 6-item test *Meaning of Letters* (Version 2A) shown in Fig. 1, and a parallel test Version 2B (not shown).

Version 2A consists of three pairs of items with different algebraic structures. *Doughnuts* and *Lego* have one letter. *Biros* and *Racetrack* have two letters with the relationship

Table 1 Rules for online diagnosis and explanations of stages of *Meaning of Letters*

Diagnosis	Rule	Short description
Stage 0	0–1 item correct	*LO* misconception in most items, rarely interpreting algebraic letters as standing for numbers
Stage 1	2–5 items correct	Sometimes algebraic letters correctly interpreted as standing for numbers and sometimes *LO*
Stage 2	6 items correct	Algebraic letters consistently interpreted correctly as standing for numbers, rather than as objects
SAC	At least 1 *SAC* response	Coefficients in the equation interpreted as a/the solution to the problem (from 4 items only)

between them directly stated. *Garden* and *Wheels* have two letters which are not directly related in the text, although constrained by it. Note that the *Lego* item does not use the initial letter of *Lego* or bricks as the algebraic letter, whereas *Doughnuts* uses the initial letter (as do the other four items of the test). The other two pairs of items are constructed to be as close a match as possible. Note that the *Biros* and *Racetrack* items have the same algebraic structure as the famous "Students and Professors" problem (Clement et al., 1981) although the natural language presentation of the relationship is different.

To prepare for the analysis below, Fig. 2 shows likely interpretations of algebra for students selecting each of the five options of the *Biros* item. This item uses the first letter of both involved objects, which might encourage students to make a *LO* interpretation. Additionally, the item stem uses the wording "Sam bought p packs" instead of, for example, the more direct "p is the number of packs Sam bought". Both potential hurdles are intentional as the test aims to inform teachers of any misconceptions that might be present and not to prevent students from making this mistake. All the incorrect responses can arise from the *LO* misconception. The coefficients in the equation of the last option ($30b = 10p$), are numbers (30 and 10) which are a possible solution to the problem. This is an example of the *solution-as-coefficient* (*SAC*) subtype of the *LO* misconception, which is flagged in the report to teachers. Any incorrect response contributes to the *LO* misconception being reported in form of Stage 0 or 1.

In addition to the diagnosis in form of a stage and misconception code, more detailed explanations as well as teaching suggestions are provided for the teacher to choose subsequent learning activities for their students. The suggestions primarily aim at incorporating an increased awareness of the *LO* misconception into the teaching, e.g., by avoiding "fruit salad algebra", paying attention to how students read equations, and emphasising and clearly identifying the meaning of variables. By linking to the Mathematics Developmental Continuum P-10 (a predecessor of current Victorian curriculum resources), further explanations and suggested activities are provided.

3.2 Translation, adaptation and validation

Since understanding variables plays the same fundamental role in algebra education in Australia and Germany, the adaptation of the *Meaning of Letters* test seemed appropriate and meaningful. This was affirmed by a first pilot with two German teachers and their students in which we observed the anticipated (mis)conceptions among German students as well as self-reported teaching habits that could encourage the *LO* misconception (Klingbeil et al., 2022).

The English items were translated into German by the German team (German native speakers with fluent knowledge of the English language) in close cooperation with the Australian team and in compliance with (applicable) ITC guidelines (ITC, 2017). In some cases, the subject contexts were adapted to make them more accessible for German students (e.g., better known plants, German currency). This sometimes led to further necessary changes, e.g., doughnuts became "Enten" (ducks) because an object starting with the same letter as euros (instead of dollars) was required to maintain the intended logic of distractors. To ensure understanding, the German *Wheels* item was changed to be about the number of tyres instead of wheels since the German word for 'wheel' ("Rad") can also be used for 'bike' which could have been confusing for students. Moreover, some of the completion stems were changed into question stems as the German wording seemed rather complicated and not so familiar to students. This should not have an impact on the diagnosis though, as research shows no difference in discrimination between those two item formats (Haladyna et al., 2002).

The explanations of stages and misconceptions and the teaching suggestions were translated considering terms and knowledge German mathematics teachers are likely to be familiar with (judging from pilot interviews and teaching

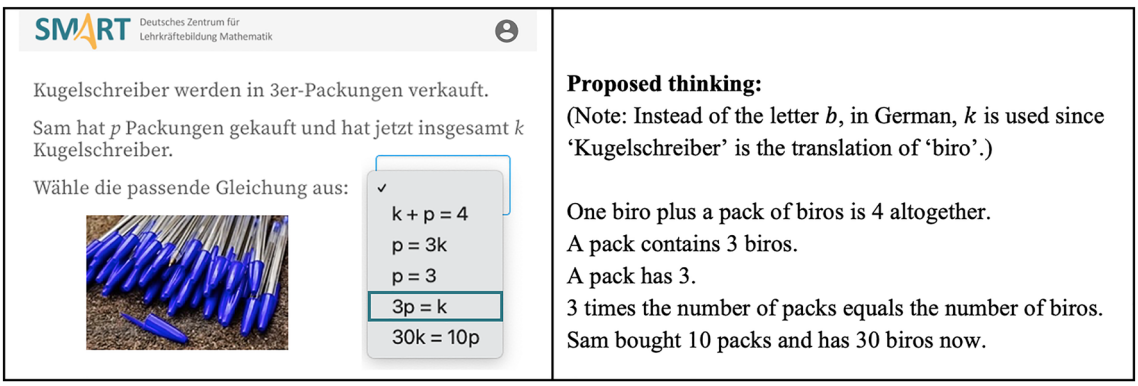

Fig. 2 German *Biros* item with proposed thinking for each option

experience of members of the German team at secondary schools and in in-service teacher training courses). Some passages were elaborated on further and references to *Grundvorstellungen* (Weigand et al., 2022) common in German mathematics didactics were made explicit. Due to design decisions in the newly programmed German online tool, suggestions were slightly restructured. Instead of referring to activities from the Victorian curriculum, tasks based on German textbooks were suggested.

For translated and adapted tests, the ITC guidelines recommend item analysis, reliability analysis and differential item functioning analysis, to investigate the reliability and validity of adapted versions. However, Gikandi and colleagues (2011, p. 2337) made clear

> that it is necessary to reconceptualise and redefine validity and reliability within the context of formative assessment because the typical definitions applied in summative assessment are limited to quantitative conceptualizations, which is not sufficient to establish validity and reliability within the context of formative assessment. … Therefore, a qualitative or mixed methods approach is often required to establish the degree of validity and reliability in formative assessment.

For this, we are following the Validation Framework for Formative Assessment (FA) proposed by Hopster-den Otter et al. (2019). Compared to summative assessment, the authors emphasise the importance of alignment with the teaching and learning process, the need of fine-grained information, and especially the relevance of the use facet of FA. For the validation of a FA, one should "build and evaluate an argument that helps test developers demonstrate that assessment scores are sufficiently useful for their intended purpose." (Hopster-den Otter et al., 2019, p. 719). In the case of the *Meaning of Letters* test, this purpose is a quick assessment whether students currently exhibit any signs of misinterpreting algebraic letters as abbreviations for objects in order to adapt the teaching accordingly at short notice. With this in mind, an interpretation and use argument (IUA) has been developed (see Fig. 3). According to Hopster-den Otter and colleagues (2019, p. 719),

> the IUA for formative assessment consists of inferences regarding a score interpretation as well as inferences regarding a score use. Score-interpretation inferences cover claims about students' performance from the instrument, while score-use inferences involve decisions on this performance and possible consequences in the learning process.

In the following, the proposed inferences are examined in detail and evidence is considered.

Regarding inference (1), it has been shown in Sect. 3.1 that item distractors are constructed to correspond with the *LO* misconception or its subtype *SAC* (e.g., see Fig. 2). For this, items from the literature as well as student responses to open-ended tasks from previous research have been used. Consequently, choosing one or more distractors results in a Stage 0 or 1 diagnosis indicating to the teacher the possibility of *LO* being present. If at least one of the *SAC* distractors is chosen, this will be reported additionally.

To check whether students tend to interpret algebraic letters as abbreviations for objects, it is necessary to use items that offer a context in which this interpretation is easily possible. In this sense, the selected items (see Fig. 1) reflect different situations in which the *LO* misconception can play a role (inference (2)). Items include three different equation types

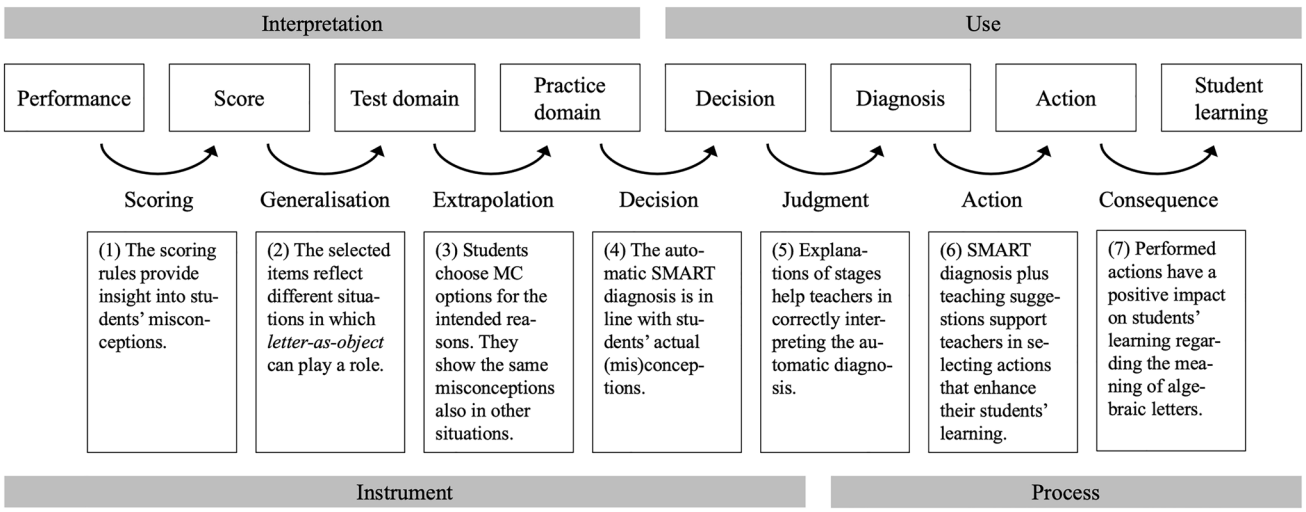

Fig. 3 Interpretation and use argument according to Hopster-den Otter et al. (2019) with proposed inferences for the *Meaning of Letters* SMART test

(see 3.1). Contexts and equations are designed to be accessible for beginning algebra learners and to avoid confounding with higher modelling competencies. Due to the restriction to a small number and the multiple-choice (MC) format, the selected items evidently do not cover all possible tasks or situations in which *LO* could occur; however, this is not a requirement for the purpose of this assessment since it does not aim at broadly generalising to the whole field of algebra.

Inference (3) claims that the items elicit information on students' thinking processes. This includes, first, that students select a response option for the intended reason, and second, that students do not only show their (mis)conception when choosing from MC options, but also in other situations, e.g., when they are formulating equations themselves. This claim will be subject of this article's investigation.

Following Table 1, a diagnostic stage is derived from responses to all six items for every student. To be helpful for the teacher, this stage should be in line with students' actual (mis)conceptions (inference (4)). If this assumption holds for the *Meaning of Letters* test, will be examined in this article.

Especially for FA, the use aspect is important when evaluating its validity. Teacher interviews from a pilot study indicate a confirmation of inferences (5), (6) and (7); however, they cannot be conclusively appraised yet, but will be investigated in the future.

4 Research questions

After having translated and adapted the *Meaning of Letters* SMART test, it is necessary to investigate whether it is valid to use the German version of this test for formative assessment concerning the *letter-as-object* misconception in German Year 7 and 8 classes. For this, we focus on three research questions:

(1) Do students choose response options of the multiple-choice items for the intended reasons? (cf. inference (3))
(2) Do students who choose distractors also show the according misconception in open-ended tasks, e.g., when formulating expressions themselves? (cf. inference (3))
(3) Does the automatic SMART diagnosis adequately reflect the students' thinking with regards to the *letter-as-object* misconception? (cf. inference (4))

5 Methods

5.1 Cognitive interviews

As a first step, we asked a class of Year 8 students to complete the SMART test online (German Version 2A). Based on the automatic online diagnosis, we intended to choose individual students for interviews to cover as wide a range of stages of understanding and misconceptions as possible. However, these choices were limited by the lack of consent forms and absences on the day of the interviews. Hence, we eventually conducted interviews with five students (four on Stage 0, one on Stage 1) one week after the online testing. To answer research question (1), we followed the approach of (cross-cultural) cognitive interviewing according to Willis (2005, 2015). The students were presented with a pen-and-paper test (PP) based on the parallel *Meaning of Letters* Version 2B test and were asked to think aloud and answer semi-structured probing questions by the interviewer. In order to answer research question (2) and to examine possible effects of multiple-choice (MC) items providing corrective feedback or provoking non-prominent misconceptions (following on from comments by Akhtar & Steinle, 2017), the test had been modified slightly: the *Trucks* item (B version of *Doughnuts*) and the *Fruit* item (B version of *Garden*) had been changed into open-ended (OE) tasks by leaving out the response options. The cognitive interviews were audio recorded and transcribed. A data- and concept-driven analysis with regards to the students' (mis)conceptions was conducted by the first author. Subsequently, the analysis was discussed among the German team and then compared with the automatic diagnosis from the online test to answer research question (3). For this article, reported interview data has been translated into English by the German team.

5.2 Online testing with written explanations

To extend the interview results on research question (1) and (3), we added two OE questions to the German Version 2A online test each asking students to give a reason for their choice in the preceding MC item (*Biros* and *Wheels*). The diagnostic rules and teacher reports remained unchanged. As part of our on-going work, the test (here designated as Version 2A*) is being administered to Year 7 and 8 students across Germany. For this article, we drew on data from students who took Version 2A* at the beginning of a teaching unit about variables, algebraic expressions and/or equations. After excluding students who did not receive an automatic online diagnosis (e.g., because they did not submit their answers), we ended up with data from 600 students (228 Year 7, 372 Year 8). To compare students' written explanations to *Biros* with other data from the test (MC responses to *Biros* and the online diagnosis from the full test), two authors from the German research team coded the explanations with regards to underlying misconceptions following a Qualitative Content Analysis (Kuckartz, 2019). Afterwards, code definitions were refined, and deviations discussed until agreement between the two raters was reached (see Table 2 for the developed codes). With the final coding manual, it

was possible to assign exactly one code to each of the 600 written explanations.

6 Results

6.1 Results from cognitive interviews

Before presenting results separately addressing the three research questions, we provide an overview of the cognitive interviews by means of short case summaries (see Table 3).

6.1.1 Misconceptions when choosing multiple-choice responses

When students chose *letter-as-object* (*LO*) intended responses, they predominantly showed this misconception also in their explanations. Sometimes they explicitly stated that the letter was standing for the involved object, sometimes they read the given equation out loud substituting the object names for the letters. This was similar for *solution-as-coefficient* (*SAC*) responses, for example, when choosing the equation $36c = 3p$ (B version of *Biros* item):

> Laura: The p stands for packages, and if you have 3 times 12, then you have 36 and the c stands for coloured pencils. … So, 38 [sic] coloured pencils are 3 packages altogether.

On a few occasions, the interpretation of letters was ambiguous or not consistent, e.g.:

> Lynn: c are the coloured pencils and p are how many packages she has bought? And with p are 3 and 3 times 12 are 36. … So, she has 36 coloured pencils now because she has bought 3 packages.

Here, the student first seemed to interpret c as an abbreviation, though it could be argued that this was supposed to be a short, colloquial way of saying that c stands for the number of coloured pencils. Then she gave a correct interpretation of p, but ended with using the object names and interpreting the coefficients as the solution.

Surprising were Lorena's explanations: Despite choosing correct responses, she would explain her choice with *LO* interpretations using the letters as abbreviations:

> Lorena: $12p$ equals c because that is a 12-pack, and these are coloured pencils.

6.1.2 Misconceptions in open-ended items

Overall, we observed that students showed the *LO* misconception also in open-ended (OE) items when no multiple-choice (MC) options were given. However, in the first interview item, the first intuitive interpretation of the variable t almost unanimously was as the unit symbol for tonnes (metric tons). Only when being asked to explain the meaning of the whole equation, all students changed their mind and gave *LO* responses. Trying to formulate an equation themselves, four of the five students attempted to find a solution in order to use this as coefficients in their equation (*SAC*). Only Lorena managed to skilfully combine the given values and algebraic letters into the correct equation despite not really understanding their meaning (here in the *Fruit* item (B version of *Garden*) in OE format):

> Lorena: Um, it says that an apple costs 2 euros. And a stands for apple and therefore $2a$, so 2. … Because an apple costs 2 euros. And $3k$ because a kiwi costs 3 euros and k stands for kiwi.

Table 2 Examples of the code manual for written explanations to *Biros* (see Fig. 2)

Code	Short description of code	Example of student explanations
LO	Clearly *letter-as-object* (*LO*) (other than *SAC*)	"Because p stands for packs and b for all the biros." ($p = 3b$)
LO-am	Most probably *LO*, but ambiguous; sometimes partly correct	"Because there are three biros in one pack." ($p = 3b$)
SAC	*Solution-as-coefficient* (*SAC*) (subtype of *LO*)	"In 1 pack there are 3 biros. Everything times ten that makes 30 biros in 10 packs." ($30b = 10p$)
CR	Correct reasoning	"The number of packs times three gives the number of biros." ($3p = b$)
CR-am	Correct in principle, but ambiguous	"The three stands for one pack and the b for the total number of biros." ($3p = b$)
NC	Not clear/ambiguous (e.g., contradicting or partly erroneous)	"Since the 3 can be replaced by any number and thus it is shown 3/4/5/…p are equal to 9/12/15/… b" ($3p = b$)
OTH	Other explanation (not LO, SAC, or CR)	"Because the result must be b and this was the only option with b as the result." ($3p = b$)
NME	No meaningful explanation (no reason, guessed, nonsense, or no answer at all)	"Because that's how it is"

Codes are written in Roman while misconceptions in general are italicised

Table 3 Summary of cognitive interview (n = 5)

Student	Online diagnosis	Interview	Short case summary
Lynn	0 *SAC*	Confirms diagnosis	She repeatedly shows the use of variables as abbreviations. However, instead of answering what *t* stands for, she intuitively wants to find a value for *t* when MC options are absent. When choosing equations, she interprets the coefficients as a solution and explains them with a calculation. In the OE item, she struggles to find a solution herself
Laura	0 *SAC*	Confirms diagnosis	She consistently and explicitly shows *LO*. Only once, she manages to choose the correct answer by considering the given values in the task and the coefficients in the MC response options, without proper understanding though ("because of gut feeling"). In two items she shows *SAC*, also when formulating an equation herself
Julian	1 *SAC*	Confirms diagnosis principally (*LO* and *SAC* prevalent)	He consistently shows *LO*, also in the two items asking for the variable's meaning which he had answered correctly in the online test. Thus, the PP test deviates from the online diagnosed stage (0 instead of 1). He shows *SAC* in two items: when choosing and formulating an equation himself; the latter not successfully though
Gabriel	0	Confirms diagnosis partly (*LO* prevalent, but also *SAC*)	While in the online test, he has chosen only *LO* responses, in the interview, he shows *SAC* when formulating an equation himself. When choosing an equation, he also tries to find a solution, but fails. As he does not know the solution, he chooses the option without any coefficients. This response is not counted as *SAC* in the online test, but he seems to interpret (absent) coefficients as indicating (unknown) solutions
Lorena	0 *SAC*	Confirms diagnosis partly (*LO* prevalent, but no *SAC*)	Despite choosing and even formulating only correct equations, she consistently shows *LO* in her explanations. As she only fails the two items asking for the variables' meaning, she could be diagnosed at Stage 1 in the PP test instead of 0 online, but her explanations consistently reveal *LO* which matches Stage 0. She does not show any signs of *SAC* though. Only when directly being asked if a certain *SAC* option would also be correct, she agrees

In addition to using *a* as an abbreviation for apple instead of standing for the number of apples, she does not consider the operation linking the number and the letter at all. Moreover, our interpretation that she is merely combining numbers and letters without complete understanding is affirmed by further *LO e*xplanations in the interview and incorrect answers to the two items asking for the variable's meaning.

6.1.3 Comparing cognitive interviews with online diagnosis

In general, a high concordance between the five interviews with the PP test (using partly modified Version 2B items) and the online diagnosis (Version 2A test, completed 1 week before) was found, but also a few deviations (see Table 3). In all cases, the *LO* misconception identified in the interview was correctly diagnosed by the online test; only the exact stage (0 or 1) sometimes deviated. Since the difference between Stage 0 and 1 only lies in the frequency of chosen *LO* responses, this does not change the diagnosis of the misconception being present. The diagnosis of the *SAC* misconception, however, did not always correspond with the observations in the interview. While one student (Lorena) did not show any signs of *SAC* during the interview despite having chosen one *SAC* response online, another student (Gabriel) did show *SAC* in his explanations, but probably was not able to choose an *SAC* response online because he was experiencing difficulties with calculating a (possible) solution.

6.2 Results from *Biros* written explanations

6.2.1 Comparing *Biros* MC responses with *Biros* written explanations

This section compares the 600 students' responses to the *Biros* item, with the explanations given for their choice (see Fig. 4). Students who chose $p = 3b$ as an answer (n = 341) were very likely to also give a *LO* related explanation (97% of 305 meaningful explanations, LO/LO-am). This is similar for $p = 3$ (n = 31; 95% of 21 meaningful explanations *LO* related). Most students who chose $b + p = 4$ (n = 38) did not give meaningful explanations. Students who chose the intended *SAC* option $30b = 10p$ (n = 78) showed SAC in 92% of 60 meaningful explanations; the remaining five students have been coded as LO-am. The correct option $3p = b$ was

chosen by 105 students. Their explanations are varied: of the 75 meaningful explanations, 32% were coded as (rather) correct (CR/CR-am), 35% as LO, and 28% as ambiguous (LO-am/NC) and 5% as other (OTH).

6.2.2 Comparing online diagnoses with *Biros* written explanations

Comparing the online diagnosis with the coded explanations on *Biros* (see Table 4), two different perspectives can be considered:

1. If a certain misconception is detectable in the student's written explanation, will the online diagnosis report this? (reading Table 4 in columns)
2. If a certain misconception is reported by the online diagnosis, will this misconception also be detectable in the student's written explanation? (reading Table 4 in rows)

Regarding perspective 1, it was found that all but 2 of the 420 students with *LO* related explanations (coded LO, LO-am or SAC) were diagnosed at Stage 0 or 1 (i.e., showing *LO*). From perspective 2, we ask whether students with a *LO* diagnosis (Stage 0 or 1) also gave a *LO* related explanation. Stage 0 students (n = 367) gave predominantly *LO* related (LO/LO-am/SAC) reasons (78%) and no correct explanations. Of Stage 1 students' explanations (n = 222), 59% were *LO* related and a few (6%) were (rather) correct (CR/CR-am). Most other explanations from Stage 0 and 1 students were not meaningful (NME).

From perspective 1, of students with correct explanations (CR, n = 13) 38% were diagnosed at Stage 2, the rest at Stage 1. As would be expected, a correct solution to this item does not ensure correct solutions to the other 5 items. In terms of perspective 2, Stage 2 students (n = 11) wrote as many correct (CR) as ambiguous explanations (CR-am/NC/LO-am); furthermore, one student answer has been coded LO. Including LO-am, 2 of the 11 students who were correct on all items in the test nevertheless gave *LO* related explanations.

Regarding the *SAC* subtype, it was found that all students with an SAC coded explanation (n = 58) were diagnosed at Stage 0 or 1. Additionally, all of these students except for one were flagged as having *SAC* by the SMART system (see Table 4, last row). Taking perspective 2, the picture is different: Only 14% of explanations from *SAC* diagnosed students have been coded as SAC.

In general, it should be noted that only 11 of 600 students answered all 6 items correctly (Stage 2). While this could be expected for Year 7 students (2 of 228 at Stage 2) who are usually encountering formal variables for the very first time at this point, this is more surprising for Year 8 students (9 of 372 at Stage 2) who already should have had opportunities to develop a viable understanding of algebraic letters. The same applies to the very low number of CR/CR-am coded explanations (Year 7: 5/2, Year 8: 8/9). These results underline the importance of drawing teachers' attention to the meanings of algebraic letters.

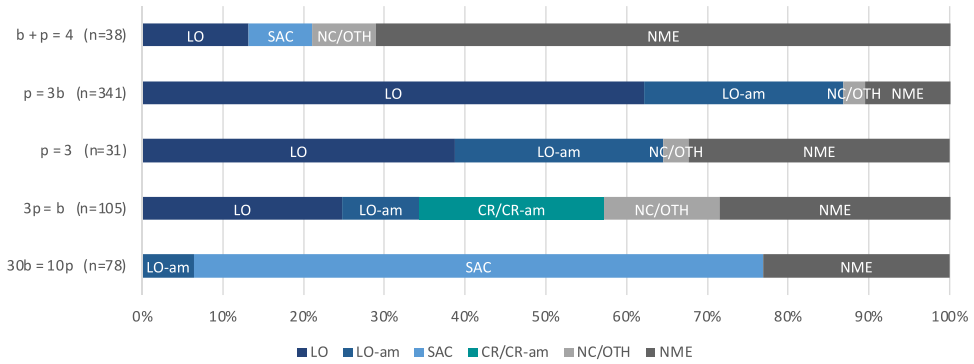

Fig. 4 Proportions of coded written explanations for response options (number of times the option was chosen in brackets; no option chosen: n = 7)

Table 4 Coded written explanations on *Biros* by SMART online diagnosis

Online diagnosis	Coded explanations								
	LO	LO-am	SAC	CR	CR-am	NC	OTH	NME	Total
Stage 0	178	63	45	0	3	5	4	69	367
Stage 1	76	43	13	8	6	12	5	59	222
Stage 2	1	1	0	5	2	2	0	0	11
Total	255	107	58	13	11	19	9	128	600
SAC	*184*	*80*	*57*	*4*	*5*	*6*	*6*	*70*	*412*

7 Discussion and conclusion

7.1 Alignment of chosen response options with students' explanations (RQ 1)

To answer the question whether students choose response options of the multiple-choice (MC) items of the *Meaning of Letters* test for the intended reasons, we first conducted cognitive interviews with five Year 8 students. We found that the chosen response options were predominantly in line with students' explanations especially when they had chosen *letter-as-object* (*LO*) or *solution-as-coefficient* (*SAC*) options. Inconsistencies, insecurities, or indications of guessing were observed only sporadically. Surprisingly, in one case the correct equations were chosen despite showing *LO* in explanations.

These findings are supported by the results from the written explanations to the *Biros* item of 600 Year 7 and 8 students. Students mainly chose response options for the reasons that were initially assumed when designing them. The exceptions are the correct option $3p = b$ and option $b + p = 4$. While the correct option was also chosen for *LO* reasons (interpreted as "One 3-pack contains biros"), only a few of the explanations for $b + p = 4$ showed the *LO* misconception, and most were non-meaningful (NME). In general, a high number of NME explanations (n = 128) was observed.

To evaluate the described deviations, it is important to remember that the SMART diagnosis is not resulting from only one single item, but from six items. Thus, it can be argued that a deviation between one single response and its explanation is not automatically problematic. A student choosing the correct option for the wrong reasons in the *Biros* item would probably still show their *LO* misconception in items explicitly asking for the meaning of the letter (like Lorena in the interview). This emphasises the importance of basing the diagnosis on a combination of various items. For option $b + p = 4$, it could be discussed if this response should not be included for the *LO* diagnosis because only 21% explanations have been coded as LO or SAC whereas the majority was coded NME. However, although it is not clear whether this option has been chosen due to *LO* thinking or as a result of arbitrarily combining given letters and numbers, a correct interpretation of algebraic letters is obviously lacking. Therefore, including this response for a Stage 0 or 1 diagnosis seems reasonable as similar support by the teacher should be helpful. Similar arguments may be applied to the 128 NME explanations (all diagnosed at Stage 0 or 1) in general. While it is unclear if their non-meaningful explanations are a result of laziness, guessing, insecurity, other ways of thinking, problems with formulating one's thoughts, or the *LO* misconception, teachers will be alerted to a lack of understanding by the diagnosis and can investigate underlying reasons more deeply where required.

7.2 Alignment of chosen response options with responses to open-ended tasks (RQ 2)

Considering the question whether students would show the same (mis)conceptions in open-ended (OE) tasks as in MC tasks, the interviews with five Year 8 students show that this was indeed mainly the case. All students also gave *LO* and/or *SAC* answers in the absence of MC response options. However, without the corrective feedback of MC options, one student found a value for the variable instead of explaining the meaning of the variable within the given context; this answer did not provide any diagnostic information for or against a possible *LO* misconception. Regarding the *SAC* subtype, all students who chose *SAC* responses during the interview also attempted to find a solution in order to use this as coefficients when formulating equations themselves (not always successfully though). Inversely, using an *SAC* approach in OE tasks did not necessarily align with choosing an *SAC* response if the student encountered difficulties with calculating a solution.

Overall, we found that the investigated misconceptions were not evoked by response options, but also occurred in their absence. Since the OE items required different activities (e.g., formulating equations instead of choosing them), it can be assumed that the SMART diagnosis can be carefully generalised to related open-ended tasks. However, it cannot be ruled out that the online test, which exclusively consisted of MC items, one week before the interview had any effects on how the OE items were approached.

7.3 Alignment of SMART diagnosis with students' (mis)conceptions (RQ 3)

Concerning the question whether the automatic SMART diagnoses adequately reflect students' thinking with regards to the *LO* misconception, the five cognitive interviews revealed that the automatic online diagnosis of the *LO* misconception matched the student explanations during the interviews very well besides from differences in the exact stage (in one case Stage 0 instead of 1 online). For the subtype *SAC*, deviations between online diagnosis and interview were found in two cases. The analysis of the written explanations to the *Biros* item showed a very high concordance between the SMART diagnosis and students' explanations in terms of perspective 1: students exhibiting the *LO* misconception in their explanation were almost exclusively diagnosed at Stage 0 or 1, and, if applicable, *SAC* additionally was flagged. Students with correct explanations to the

Biros item were appropriately diagnosed at Stage 1 or 2. The results from both the interviews and the written explanations demonstrate that teachers will receive valuable information from the SMART test on students who show signs of *LO* in their explanations. Whether students are diagnosed at Stage 0 or 1 has no influence on the provided teaching suggestions. The rare cases in which students were incorrectly not flagged as having *SAC* seem acceptable in a formative setting since they were still diagnosed as showing *LO* of which *SAC* is a subtype. Thus, teaching targeted at *LO* should also resolve *SAC* issues.

Taking perspective 2, it could be observed that the majority of students being diagnosed at Stage 0 and 1 indeed gave *LO* related explanations. Some Stage 2 students, however, also showed misconceptions or gave unclear reasons, so that teachers could not rely on these students having a completely correct understanding. It is important to note, though, that here only 11 students were diagnosed at Stage 2. Therefore, it is not appropriate to draw generalisations based on this limited sample size. Concerning *SAC* flagged students, a discrepancy to their explanation becomes visible: most of them did not give an SAC reason for their choice of response option in the *Biros* item. This makes clear that it is not possible to draw conclusions from the online diagnosis to a certain explanation for one single item. However, this does not necessarily mean that the diagnosis is inadequate. Since choosing an *SAC* option in just one item is enough to get an *SAC* diagnosis, it is possible that these students simply did not choose the *SAC* option in the *Biros* item but would show this misconception in one of the other items. In fact, we found that only 78 of 412 *SAC* diagnosed students chose the *SAC* response option in the *Biros* item, while 314 of them chose it in *Garden* and 344 in *Wheels*. Thus, the *Biros* item seems to be not as predictive for the *SAC* misconception as the *Garden* or *Wheels* items with their different equation type.

7.4 Limitations

As one limitation, the design of the investigated test needs to be considered. The set of six multiple-choice items makes the test quick to administer but limits the activities the students are required to perform. Also due to its tight thematic focus on the *LO* misconception, no generalisations to broader algebraic competencies can be drawn. This limitation can be also seen as a strength though: the test can provide teachers with very concrete detailed information. This is especially valuable since interpreting algebraic letters correctly is a fundamental requirement for further algebra learning. In terms of the investigation methods, it needs to be considered that, in order to capture students' thinking, interviews and written explanations to one item were used. This can of course only be an approximation and is prone to misinterpretation by researchers. However, we deem this approach sufficient to give an indication if items indeed evoke intended thinking processes. Regarding the students' written explanations, it needs to be taken into account that these were mostly very short and often not unequivocally clear, making it difficult to distinguish between linguistic inaccuracies and manifest misconceptions. Therefore, also "ambiguous" codes were used (LO-am, CR-am, NC; see Table 2). Despite its ambiguity, we consider the LO-am code in our analysis as indicating the *LO* misconception which is corroborated by results from the five cognitive interviews. In general, this observation of ambiguous student explanations further encourages the idea of (also) using MC items to elicit information about students' thinking as students are not always capable of realising and formulating their own struggles and misconceptions.

7.5 Conclusion

Having examined our three research questions in detail, we see our extrapolation (3) and decision (4) inferences, important steps of the validation process (see Sect. 3.2), supported. We therefore conclude that the MC items of the SMART test *Meaning of Letters* indeed can be used to assess the *LO* misconception of German Year 7 and 8 students with a formative objective. The potential of the test lies especially in providing teachers with information on their students' thinking by reporting probably existing misconceptions (perspective 1) as students with LO or SAC explanations were flagged by the SMART system accordingly. For perspective 2, concordance between SMART diagnosis and students' explanations (to one item) was not as high, but still adequate for formative use. Since SMART tests are not about labelling or making high-stake decisions, but about enabling teachers to provide suitable support for learners, observed deviations are considered acceptable. However, to be able to conclusively evaluate the validity with regards to the effects of the test on teachers' judgement, taken actions and students' subsequent learning (inferences (5), (6) and (7)) further investigations will be necessary.

While we have investigated the potential of MC items for formative assessment exemplarily for the *Meaning of Letters* SMART test, we suggest that our findings are paradigmatic: MC items have a high potential for revealing misconceptions if they focus on fine-grained aspects and are developed carefully drawing on previous research and student data. However, a combination of various items seems to be essential to ensure an adequate overall assessment and to capture different aspects of students' thinking. Here, technology can be powerful in terms of automatic analyses across several items.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

Akhtar, Z., & Steinle, V. (2013). Probing students' numerical misconceptions in school algebra. In V. Steinle, L. Ball & C. Bardini (Eds.), *Proceedings of the 36th annual conference of the Mathematics Education Research Group of Australasia* (pp. 36–43). MERGA.

Akhtar, Z., & Steinle, V. (2017). The prevalence of the 'letter as object' misconception in junior secondary students. In A. Downton, S. Livy & J. Hall (Eds.), *Proceedings of the 40th annual conference of the Mathematics Education Research Group of Australasia* (pp. 77–84). MERGA.

Arcavi, A., Drijvers, P., & Stacey, K. (2017). *The learning and teaching of algebra: Ideas, insights, and activities*. Routledge.

Bell, B., & Cowie, B. (2001). The characteristics of formative assessment in science education. *Science Education, 85*(5), 536–553.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5–25. https://doi.org/10.1080/0969594x.2010.513678

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement, 29*(3), 253–271. https://doi.org/10.1111/j.1745-3984.1992.tb00377.x

Bürgermeister, A., & Saalbach, H. (2018). Formatives Assessment: Ein Ansatz zur Förderung individueller Lernprozesse. *Psychologie in Erziehung Und Unterricht, 65*(3), 194–205.

Clement, J., Lockhead, J., & Monk, G. (1981). Translation difficulties in learning mathematics. *American Mathematical Monthly, 88*(4), 286–290.

Cusi, A., Aldon, G., Barzel, B., & Olsher, S. (accepted). Rethinking teachers' formative assessment practices within technology-enhanced classrooms. In B. Pepin, G. Gueudet, & J. Choppin (Eds.), *Handbook of digital resources in mathematics education*. Springer.

Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher, 32*(1), 5–8.

Drouhard, J.-P., & Teppo, A. (2004). Symbols and language. In K. Stacey, H. Chick, & M. Kendal (Eds.), *The future of the teaching and learning of algebra: The 12th ICMI study* (pp. 226–264). Kluwer.

Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. *Computers & Education, 57*(4), 2333–2351. https://doi.org/10.1016/j.compedu.2011.06.004

Grugeon-Allys, B., Chenevotot-Quentin, F., Pilet, J., & Prévit, D. (2018). Online automated assessment and student learning: The PEPITE project in elementary algebra. In *Uses of technology in primary and secondary mathematics education: Tools, topics and trends* (pp. 245–266). https://doi.org/10.1007/978-3-319-76575-4_13

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309–333. https://doi.org/10.1207/S15324818AME1503_5

Hopster-den Otter, D., Wools, S., Eggen, T. J., & Veldkamp, B. P. (2019). A general framework for the validation of embedded formative assessment. *Journal of Educational Measurement, 56*(4), 715–732. https://doi.org/10.1111/jedm.12234

International Test Commission. (2017). The ITC guidelines for translating and adapting tests (Second edition). *International Journal of Testing, 18*(2), 101–134. https://doi.org/10.1080/15305058.2017.1398166

Kaput, J. (1995). A research base supporting long term algebra reform? In D. Owens, M. Reed, & G. Millsaps (Eds.), *Proceedings of the 17th annual meeting of the North American chapter of the international group for the psychology of mathematics education* (Vol. 1, pp. 71–94). ERIC/CSMEE.

Kieran, C. (2007). Learning and teaching algebra at the middle school through college levels. Building meaning for symbols and their manipulation. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 706–762). Information Age Publishing.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30*(4), 28–37.

Klingbeil, K., Rösken, F., Thurm, D., Barzel, B., Schacht, F., Kortenkamp, U., Stacey, K. & Steinle, V. (2022). $SMART_A$—Online-Diagnostic to reveal students' algebraic thinking and enhance teachers' diagnostic competencies. In U.T. Jankvist, R. Elicer, A. Clark-Wilson, H.-G. Weigand, & M. Thomsen (Eds.), *Proceedings of the 15th international conference on technology in mathematics teaching (ICTMT 15)* (pp. 290–297). Aarhus University. https://doi.org/10.7146/aul.452

Küchemann, D. (1981). Algebra. In K. M. Hart, M. L. Brown, D. E. Küchemann, D. Kerslake, G. Ruddock, & M. McCartney (Eds.), *Children's understanding of mathematics: 11–16* (pp. 102–119). John Murray.

Kuckartz, U. (2019). Qualitative text analysis: A systematic approach. In G. Kaiser & N. Presmeg (Eds.), *Compendium for early career researchers in mathematics education* (pp. 181–197). Springer. https://doi.org/10.1007/978-3-030-15636-7_8

Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., & Neubrand, M. (Eds.). (2013). *Cognitive activation in the mathematics classroom and professional competence of teachers: Results from the COACTIV project*. Springer. https://doi.org/10.1007/978-1-4614-5149-5

MacGregor, M., & Stacey, K. (1993). Cognitive models underlying students' formulation of simple linear equations. *Journal for Research in Mathematics Education, 24*(3), 217–232.

MacGregor, M., & Stacey, K. (1997). Students' understanding of algebraic notation: 11–16. *Educational Studies in Mathematics, 33*(1), 1–19.

Narciss, S. (2013). Designing and evaluating tutoring feedback strategies for digital learning environments on the basis of the interactive tutoring feedback model. *Digital Education Review, 23*(1), 7–26.

Pellegrino, J. W., & Quellmalz, E. S. (2010). Perspectives on the integration of technology and assessment. *Journal of Research on Technology in Education, 43*(2), 119–134.

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of three basic dimensions. *ZDM, 50*(3), 407–426. https://doi.org/10.1007/s11858-018-0918-4

Prediger, S., Götze, D., Holzäpfel, L., Rösken-Winter, B., & Selter, C. (2022). Five principles for high-quality mathematics teaching: Combining normative, epistemological, empirical, and pragmatic perspectives for specifying the content of professional development. *Frontiers in Education, 7*(969212), 1–15. https://doi.org/10.3389/feduc.2022.969212

Ruiz-Primo, M. A., & Li, M. (2013). Examining formative feedback in the classroom context: New research perspectives. In J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 215–232). SAGE Publications. https://doi.org/10.4135/9781452218649

Schütze, B., Souvignier, E., & Hasselhorn, M. (2018). Stichwort—Formatives Assessment. *Zeitschrift Für Erziehungswissenschaft, 21*(4), 697–715. https://doi.org/10.1007/s11618-018-0838-7

Serfati, M. (2005). *La révolution symbolique. La constitution de l'écriture symbolique mathématique*. Pétra.

Shin, J., Guo, Q., & Gierl, M. J. (2019). Multiple-choice item distractor development using topic modeling approaches. *Frontiers in Psychology, 10*, 825.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Stacey, K., & MacGregor, M. (2000). Learning the algebraic method of solving problems. *Journal of Mathematical Behavior, 18*(2), 149–167.

Stacey, K., Steinle, V., Price, B., & Gvozdenko, E. (2018). Specific mathematics assessments that reveal thinking: An online tool to build teachers' diagnostic competence and support teaching. In T. Leuders, J. Leuders, K. Philipp, & T. Dörfler (Eds.), *Diagnostic competence of mathematics teachers—Unpacking a complex construct in teacher education and teacher practice* (pp. 241–263). Springer. https://doi.org/10.1007/978-3-319-66327-2_13

Steinle, V., Gvozdenko, E., Price, B., Stacey, K., & Pierce, R. (2009). Investigating students' numerical misconceptions in algebra. In R. Hunter, B. Bicknell & T. Burgess (Eds.), *Proceedings of the 32nd annual conference of the Mathematics Education Research Group of Australasia* (Vol. 2, pp. 491–498). MERGA.

Usiskin, Z. (1988). Conceptions of school algebra and uses of variables. In A. Coxford (Ed.), *The ideas of algebra, K-12* (pp. 8–19). National Council of Teachers of Mathematics.

Vergnaud, G. (1996). Education, the best portion of Piaget's heritage. *Swiss Journal of Psychology, 55*(2/3), 112–118.

Warren, E. (1998). Students' understanding of the concept of a variable. In C. Kanes, M. Goos, & E. Warren (Eds.), *Proceedings of the 21st annual conference of the Mathematics Education Research Group of Australasia* (Vol. 2, pp. 661–668). Brisbane: MERGA.

Weigand, H.-G., Schüler-Meyer, A., & Pinkernell, G. (2022). Didaktik der Algebra. *Springer Spektrum*. https://doi.org/10.1007/978-3-662-64660-1

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. SAGE Publications.

Willis, G. B. (2015). The practice of cross-cultural cognitive interviewing. *Public Opinion Quarterly, 79*(S1), 359–395. https://doi.org/10.1093/poq/nfu092

Yan, Z., & Pastore, S. (2022). Assessing teachers' strategies in formative assessment: The teacher formative assessment practice scale. *Journal of Psychoeducational Assessment, 40*(5), 592–604. https://doi.org/10.1177/07342829221075121