

Let's Ask Students About Their Programs, Automatically

Teemu Lehtinen
Aalto University
Espoo, Finland
teemu.t.lehtinen@aalto.fi

André L. Santos
Instituto Universitário de Lisboa (ISCTE-IUL)
Lisboa, Portugal
andre.santos@iscte-iul.pt

Juha Sorva
Aalto University
Espoo, Finland
juha.sorva@aalto.fi

Abstract—Students sometimes produce code that works but that its author does not comprehend. For example, a student may apply a poorly-understood code template, stumble upon a working solution through trial and error, or plagiarize. Similarly, passing an automated functional assessment does not guarantee that the student understands their code. One way to tackle these issues is to probe students' comprehension by asking them questions about their own programs. We propose an approach to automatically generate questions about student-written program code. We moreover propose a use case for such questions in the context of automatic assessment systems: after a student's program passes unit tests, the system poses questions to the student about the code. We suggest that these questions can enhance assessment systems, deepen student learning by acting as self-explanation prompts, and provide a window into students' program comprehension. This discussion paper sets an agenda for future technical development and empirical research on the topic.

Index Terms—Automatic assessment, automatic question generation, program comprehension, programming education, self-explanation

I. INTRODUCTION

Students' solutions to program-writing assignments are typically checked for correctness and, perhaps, other qualities such as style, efficiency, or test coverage. The results of this manual or automatic assessment are often accepted as a partial assurance that the learner has reached some intended learning outcomes. Most automatic assessment systems [1] attempt to provide rapid, constructive feedback on errors, so that students can fix their programs and learn from their mistakes.

Instructors would like students to design programs and write code that the students themselves understand. Reality bites, however. We, like many others, have seen learners search for and copy poorly understood code from textbooks and internet forums. Sometimes students fiddle with example code through trial and error until it happens to work, or are guided by teaching assistants or automatic assessment to the correct solution without reaching understanding. (Automatic assessment may exacerbate bad study habits such as trial-and-error and starting late. [2])

Learners' programs may then feature code that they do not properly comprehend: they cannot trace their program's execution and may have a shaky grasp of the principles behind the code. Plagiarism is another reason for submitting code

without understanding. Similar problems may also stem from legitimate collaboration with student peers.

Whatever the underlying reasons, successful task performance is not the same as successful learning, and a student with a program that works is not the same as a student who understands why and how the program works. Program comprehension must be acknowledged as an educational objective and that objective should be addressed in pedagogy.

Our work addresses a dual issue: 1) for deeper learning, students should reflect on the code that they write and ensure they understand it; 2) the quality of assessment is limited by a sole focus on the resultant program at the expense of understanding.

The second issue is problematic not only for teachers who wish to ensure that students do not get excessive credit for code they do not understand, but also for the learners themselves. For example, automatic assessment that focuses on the end product alone may lull learners into a false sense of understanding and lead them to miss out on important learning outcomes—a sort of *unproductive success* [3].

In order to explore students' understanding and stimulate reflection, teachers can engage students in dialogue about their code and ask questions about it (see, e.g., [4], [5]). However, when there are many students and even more programs, it is unfeasible to maintain such one-on-one human dialogue. This is the case especially in large introductory courses, such as those at Aalto University, where the rapid provision of feedback at scale is only possible because of automated assessment.

II. ARTICLE GOALS AND STRUCTURE

We propose that useful questions about student-written programs can be created automatically. These *questions about learners' code* (QLCs) can then be posed to the same student who wrote the code. The driving question that we investigate is this:

What opportunities and challenges are there in automatically generating questions (QLCs) for students to answer about their own programs?

We explore this question by connecting it to the research literature and through critical argumentation. This is a “discussion paper.” That is, our present purpose is to put forward

the idea of QLCs, to discuss its foundations in computing education research and the learning sciences, and to identify opportunities and challenges in implementing and adopting automatically generated QLCs.

As we explore these questions, we also sketch out some use cases for QLCs. As an example use case, after a learner submits a program for automatic assessment and it passes all or most unit tests (or comparable checks of correctness), the system may ask one or more questions about the program.

The work presented here is theoretical, and does not yet feature a concrete software system to support QLCs or an empirical evaluation thereof. Instead, we outline future research on tool development and empirical evaluations of QLCs.

The rest of this article is structured as follows. Section III introduces the concept of QLCs. Section IV reviews the related work that forms the theoretical backdrop for QLCs, motivates this research, and informs QLC design. Section V discusses, in broad terms, the technical problem of generating QLCs automatically and outlines a plausible process for approaching this problem. Section VI identifies use cases for QLCs and considers some different forms that QLCs may take. Section VII notes several limitations of the QLC idea. Section VIII sets down an agenda for constructive and empirical research on QLCs, and Section IX briefly concludes the article.

III. QUESTIONS ABOUT LEARNERS' CODE

A. Definition

We define QLCs as follows.

- 1) They are questions about program code that a student has written;
- 2) they refer to concrete constructs or patterns in the student's program; and
- 3) they are posed to the student themselves by a computer.

Automatic QLCs have an additional characteristic:

- 4) They are automatically generated from an analysis of the student's code.

An automatic QLC is not a canned, teacher-created question that is specific to a single programming assignment that students work on. Instead, we envision that teachers can affect which types of QLCs are generated for their students and, if they wish, configure question generation differently for different assignments. In this article, we focus on automatic QLCs. Hereafter, when we discuss QLCs, we mean automatic QLCs unless otherwise specified.

Many kinds of QLCs are possible. As an example, Figure 1 presents a solution to a programming assignment; let us imagine this solution is from a learner. The goal of the assignment is to provide practice on program writing in general and recursion in particular. Figure 2 lists a few plausible QLCs for this code.

Even when QLCs are automatically *generated*, their *assessment* may or may not be automatic. Some QLCs have answers that are relatively simple to assess automatically, such as multiple-choice or single-value questions. On the other

Task: Write a recursive function to find the smallest character in a String.

```
static char smallest(String word) {
    return smallestFrom(word, 0);
}

static char smallestFrom(String word, int index) {
    if (index == word.length() - 1) {
        return word.charAt(index);
    }
    else {
        char current = word.charAt(index);
        char rest = smallestFrom(word, index + 1);
        return current < rest ? current : rest;
    }
}
```

Fig. 1. An example task and a solution in Java.

- 1) You wrote two functions. Which of those are recursive?
- 2) What are the parameter names of your function `smallestFrom`?
- 3) How deep does the call stack grow when executing `smallest("ABBA")`?
- 4) When executing `smallestFrom("ACDC", 0)`, which character is assigned to `rest` during the *second* invocation of `smallestFrom`?
- 5) Which of the following best describes the role of your variable `rest`?¹

Fig. 2. Example QLCs for the code in Figure 1.

hand, QLCs may also have open-ended answers that are self-, teacher-, or peer-assessed. Abstract examples of both kinds of questions appear in Table II.

B. Motivation

QLCs may engage learners to reflect on their code and their programming knowledge, thus enhancing learning. They may prompt additional practice on program comprehension, set in the (presumably) familiar context of the learner's own program.

QLCs extend the assessment of student-created programs beyond functionality and style to program comprehension. By looking into the relationship between the learner and their program, QLCs complement other forms of assessment that focus on the program alone.

If collected, learners' answers to QLCs could be valuable to teachers and researchers who wish to explore students' programming knowledge. QLCs might also have uses in discouraging or detecting plagiarism. In the sections that follow, we elaborate on these potential uses and benefits of QLCs.

¹*Roles of variables* are common patterns of variable use [6] that can be taught to novices and identified automatically through static analysis [7]. For example, a *gatherer* variable gradually accumulates a result by combining multiple inputs, and a *most-wanted holder* tracks the highest, smallest, or otherwise most appropriate value among various candidates.

IV. RELATED WORK

The subsections below relate our work on QLCs to four extant threads of research: student learning of code reading, knowledge elaboration through self-explanation, automatic generation of questions for computing education, and automated tutoring systems, respectively.

A. Learning to Read Programs

1) *Student Difficulties*: Even after taking a programming course, learners frequently have fragile program-reading skills, and many struggle to trace or explain code [8], [9]. Moreover, learners are sometimes averse to tracing their code even when it might help [10], [11].

Fuller et al. [12] discuss the learning trajectories of novice programmers in terms of an adaptation of Bloom’s taxonomy. They cite evidence for some students getting stuck at an *Apply/Remember* stage, where the student relies on trial and error to imitate the use of a construct without understanding.

Salac and Franklin [13] showed that, in a primary-school context, students’ use of constructs is not a reliable indicator of understanding those constructs. Similarly, in a higher-education context, Kennedy and Kraemer [14] found that there were students who could produce a program that works but were unsure about their own code and did not grasp the underlying concepts.

QLCs target these difficulties by prompting students to reason about and explain the code they have produced. We propose that QLCs could prompt students’ progress from surface-level imitation to deeper understanding, and partially assess that progress.

2) *Types of Programming Knowledge*: Schulte’s *Block Model* of program comprehension [15], [16] captures how understanding a program requires different kinds of knowledge—knowledge about the program text, about its dynamic behavior, and about the program’s purpose; moreover, knowledge is required at different scales ranging from individual “atoms” to combinations of constructs to the full program. Table I summarises the Block Model.

We suggest that useful QLCs can be generated at various levels of the Block Model, as illustrated by the examples in Table II.

Understanding how code works is not valuable for its own sake only; the skill is required when writing, debugging, and extending programs [16]. Research suggests that ability to trace code in detail and explain what given code accomplishes tends to precede ability to write comparable code oneself [17]. In any case, writing code that you can read is, by definition, predicated on code-reading ability. We hypothesize that QLCs can support the growth of program-comprehension skills and strengthen the relationship between them and program-writing skills.

3) *Pedagogies*: Recognizing the importance of reading code, scholars have explored pedagogies where learners practice reading given code or are explicitly taught strategies for tracing [18]. For example, researchers have examined comprehension-before-writing pedagogies [19], invited learners to predict, investigate, and eventually modify the behavior

of code [20], and asked students to visually simulate given programs [21].

The present work is loosely related, given that our goal is also to promote students’ understanding of existing code. However, our work differs materially in that QLCs do not involve teacher-provided code and are an extension of code-writing practice rather than a wholly distinct learning activity. QLCs are meant to deepen and assess learners’ engagement with the code that they authored, and to do so automatically. We see QLCs as complementary—rather than alternative—to code-reading practice on given code.

Some tracing-based pedagogies emphasize the need for students to trace their own programs. For example, Hertz and Jump [22], who taught students to use a diagrammatic notation for tracing code, noted that their code-writing problems always require students to trace the code they wrote. QLCs share the goal of promoting students’ understanding of their own code, but instead of graphical tracing, we propose to ask students to answer automatically generated questions of various kinds.

B. Self-Explanation

To *self-explain* [23] is to generate explanations for yourself—rather than a teacher or a peer—as you engage in a learning activity such as problem solving, reading some text, or studying an example. It is often but not necessarily internal, i.e., not expressed aloud or in writing. Self-explanations may explain conceptual content, justify the selection of solution steps, or otherwise elaborate on the learning activity. (Self-explanation is thus not a separate learning activity but a cognitive process that accompanies an activity; it may occur before, during, and/or after the activity.)

Successful learners tend to generate better, more principled self-explanations, but many learners fail to do so [24], [25]. Fortunately, teachers and environments can encourage self-explanation, and there is strong evidence (with caveats) for improved learning from *self-explanation prompts* (SEPs) [23], [26]; moreover, learners can be taught to self-explain better [24], [27]. SEPs are typically designed manually, but work on automatic prompt generation is underway [23].

The correlation between successful self-explanation and task performance has been replicated in programming contexts, as has the impact of self-explanation training [25], [27], [28]. Moreover, research suggests that explicit subgoal-labeling promotes better self-explanations of example programs and thereby helps learners see structural similarities between examples and improves transfer [28], [29]. Asking learners to annotate given code with comments [30] or elaborate on their attempts to solve code-construction puzzles [31] are promising programming-specific SEPs.

Our work on QLCs overlaps that on SEPs; by promoting reflection, QLCs might have some of the same benefits. Depending on how QLCs are phrased, contextualized, and assessed, they may be considered as self-explanation prompts, assessment items, or both.

TABLE I
THE BLOCK MODEL (PARAPHRASED FROM [15])

	Text	Execution	Function
Atom	language elements	elements' behavior	elements' purpose
Block	syntactically or semantically related elements	a "block's" behavior	a "block's" purpose; program subgoal
Relational	connections between "blocks"; e.g., method calls	flow between "blocks"; e.g., call sequences	integration of subgoals
Macro	entire program	the program's behavior	the program's purpose

TABLE II
EXAMPLE QUESTIONS FROM QLC TEMPLATES. (SIMPLIFIED FOR BREVITY.)

Template Question	Answer Type	Block Model level [16]
Which of the following are variable names in your function?	multiple choice	atom-text
A loop starts on line [N]. Enter the number of the last line inside this loop.	single value	block-text
Line [N] uses a variable. Enter the line number where that variable is declared.	single value	relational-text
What is assigned to variable [V] on line [N] when executing function [F] on expression [E]?	single value	atom-execution
During the program execution, how many iterations are performed by the loop starting in line [N]?	single value	block-execution
Which of the following best describes the role of your variable [V]?	multiple choice	relational-execution
How deep does the call stack grow when executing function [F]?	single value	relational-execution
Describe the purpose of the condition on line [N].	open-ended	atom-function
Justify your choice of name [V] for the variable declared on line [N]—do you have a better suggestion?	open-ended	block-function
Select the part of your program that is responsible for [X]. (<i>The question could be generated from a subgoal annotated onto a model solution.</i>)	select in code	block-function
Explain, in your own words, the purpose of the loop that begins on line [N], and how that loop helps method [M] accomplish its task.	open-ended	relational-function
Here is a little example program that has some similarities with yours. Select the part of your program that serves a similar purpose as the highlighted code in the example.	select in code	relational-function (across programs)

C. Automatic Question Generation

There is earlier work on automatically generating questions for programming education. For example, Zavala and Mendoza [32] designed exercise templates, each of which could be automatically instantiated into hundreds of specific variants of the same exercise, which vary in specific values and variable names. Thomas et al. [33] automatically generated small programs accompanied by multiple-choice questions that prompted students to trace the generated code. This line of work is primarily motivated by the need to provide many fresh questions for students to practice on, and/or as a precaution for plagiarism. Our approach shares the general notion of automatically generated questions but differs fundamentally in that we do not aim to generate variants of the same exercise but to ask questions about existing code produced by the learner.

Outside of computing education, researchers have applied natural-language processing to text documents with the aim of automatically generating questions for learners [34], [35]. These questions have the potential to work as SEPs during independent study of the texts. The questions are open-ended with no well-defined correct answer; such questions are difficult to assess automatically or at scale, however. Our work has a related goal in automatically generating questions that may prompt self-explanation; we explore this idea in a programming context, generate the questions from learner-produced content, and look into automatic assessability.

D. Automated Tutors for Programming

1) *Intelligent Tutoring Systems*: Intelligent tutoring systems are designed to provide similar automated support as one-to-one human tutoring might. Over the years, many intelligent

tutoring systems have been built for programming education. Some of these systems employ both automatically assessed programming exercises and questions to guide students towards set learning goals. [36]

As an example, the ITEM/IP tutoring system tests students' knowledge by asking questions that require the students to trace given example programs until they arrive to an answer [37]. QLCs expand on that idea by using the code that the student wrote as an input to generating comparable as well as other types of questions about the properties of the program.

2) *Automatic Assessment Systems*: Automated systems for assessing student programs are popular not only to grade student work at scale but also to generate detailed, rapid feedback that aims to improve that work and thereby to tutor the student. Feedback generation requires detailed static and dynamic analysis as well as program transformations [1]. We propose to apply the previously researched program-analysis techniques in QLCs generation.

By providing feedback on programming exercises, automatic assessment systems generally aim to guide students while they are working towards fulfilling the functional requirements of a programming assignment. QLCs, on the other hand, prompt student to explain properties of a program artefact that may already have fulfilled some functional criteria—they are potentially relevant also to students who have submitted a functionally correct solution on the first attempt. We see QLCs as a promising extension to automated assessment systems, as discussed in Section VI below.

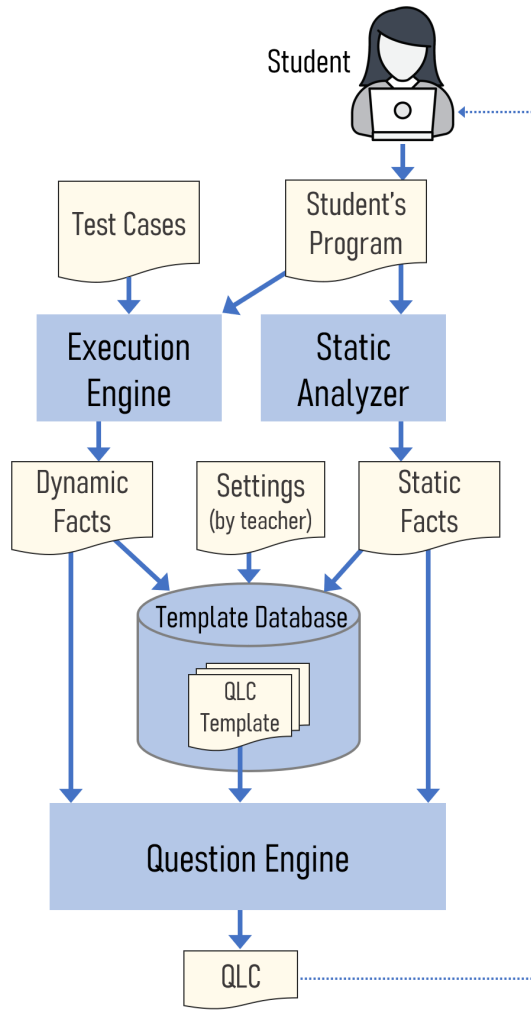


Fig. 3. A process for generating QLCs.

V. GENERATING QLCs AUTOMATICALLY

A. A General Process for Creating QLCs

We propose a general process for generating QLCs, consisting of several steps (see Figure 3).

The student's program goes through a *static analyzer* to extract *static facts* from it, such as which variables there are and which functions are recursive. Static analysis of student programs has been successfully applied for purposes of generating feedback (e.g., [38]) and visualization (e.g., [7]). For QLCs, the static analysis does not need to be complex in order to be useful, as it may essentially consist of checking if certain structural elements are present in the student's program in order to decide if a QLC is applicable to it. More ambitious analyses may be possible, too, of course.

The program also goes through an *execution engine* that collects *dynamic facts* by executing the code; typically, this execution would be based on teacher-provided test cases. Dynamic facts involve variable values, control flow, and call stack history; for instance, the call stack depth at particular function call is a dynamic fact. (One implementation approach

for the execution engine is to use a debugger interface if one is available for the programming language; cf. [39].)

A *template database* stores *QLC templates*; see Table II for examples. In addition to the question texts shown in the table, each QLC template includes a list of requirements that a program must meet for the question to apply. For example, a question might only apply if the program features recursion.

The static and dynamic facts—especially the former—are used to determine which *QLC templates* are applicable. Many QLC templates are likely to match any given program; the selection of templates among the applicable ones can be further adjusted through teacher-governed settings. These settings may be specific to a course or even a particular programming assignment. They specify which kinds of QLCs are preferred and how many QLCs should be generated. Randomization can add further variety to the questions that are presented to different learners or to the same learner at different times.

The *question engine* instantiates each chosen QLC template using the static and dynamic facts, thus producing a QLC (and its correct answer, if possible). The templates are instantiated to match the learner's concrete program or part thereof (such as a function). For example, the QLC template “How deep does the call stack grow from executing function [F]?” might be instantiated by replacing function [F] with `smallest("ABBA")`; similarly, the other templates in Table II would have the parts in square brackets filled using facts collected from the student's program.

B. Other Inputs to QLC Generation

By definition, learners' code is the key input to QLC generation. However, there are other sources of information that might be additionally exploited in order to generate better QLCs.

For example, the results of passed and failed unit tests might be used to tailor QLCs. Students could be asked about differences between their code and a model solution. Teachers might annotate model solutions, e.g., by attaching subgoal labels to sections of code; QLCs could then target those subgoals. (Some low-level, domain-generic subgoals such as “initialize variables before starting loop” might be detected automatically, without a teacher's intervention.) Students might be asked to compare their program to other programs which would also need to be provided as input to the QLC generator.

To personalize question topics and difficulty, a QLC generator would also need data on the learners' prior coursework, including which QLCs the learners have previously answered.

VI. USE CASES FOR QLCs

There are a number of scenarios where QLCs might usefully complement automatic assessment or otherwise integrate into programming education. We outline some of them below.

A. Use Case: QLCs after Unit Tests

Automatic assessment systems collect students' program code and execute it in a safe environment for analysis and feedback. These platforms could be extended with a QLC

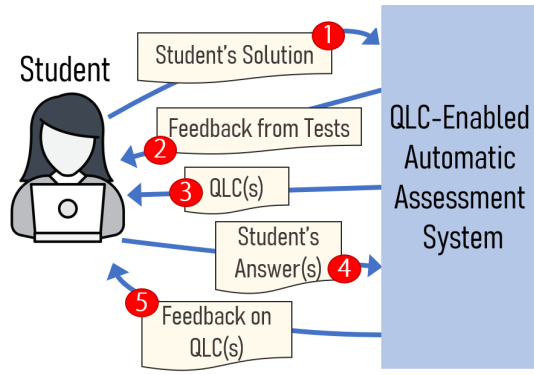


Fig. 4. QLCs integrated with automated assessment. The numbers indicate the timeline of the interaction between a student and the system.

service that takes student submissions as input, generates QLCs, presents them to students, collects students' answers, and provides feedback on those answers.

In an automatic-assessment context, one option is to ask QLCs *after successful testing*—i.e., once the student's program passes functional tests. In this scenario, students are expected to have a good understanding of their own program when the QLCs are presented and the QLCs might serve as additional assessments. This is the use case that we are currently focusing our implementation efforts on.

Figure 4 illustrates the planned workflow between the learner and the system in this scenario. The learner writes a solution for a given exercise served by the system, which holds unit tests for it. After submitting a program for assessment (1), the learner receives the test results as feedback (2). Upon passing the tests, the learner is additionally presented with one or more questions about their solution (3). The learner answers the questions (4) and immediately receives feedback on the answers (5). That feedback, like the questions, is phrased in terms of the student's own code.

B. Alternative Use Cases

Alternatively or additionally, an automatic assessment system could ask QLCs *after failed tests*. This may prompt reflection but could also distract the student inconveniently unless the questions are particularly helpful.

An automatic assessment system might also ask QLCs *before testing*, perhaps as a precondition to displaying test results. That timing might be perceived as intrusive by students, however; moreover, the literature suggests that self-explanation prompts tend to be more useful when the student already knows whether the work they are explaining is correct or incorrect [26], and the same may apply to QLCs.

A different setting for QLCs is the learners' programming environment, such as an IDE. If the students run unit tests in this environment, QLCs might be presented before or after those tests, as described for automatic assessment systems above. Alternatively, students might be given the opportunity to ask for QLCs when they wish to answer them. A QLC-enabled IDE might even notify students when a potentially

interesting QLC is available after a change to the program, although this option would need an exceptionally successful implementation so as to be both useful and unintrusive.

Interactive ebooks with embedded programming assignments are increasingly common (e.g., [40]–[42]). They offer similar integration options for QLCs as automatic assessment systems and IDEs do.

C. Alternative Forms of QLCs

In this paper, most of our discussion revolves around the notion of generating questions about a single program at a time. However, asking questions that link multiple programs also merits consideration. A student could be asked to comment on the commonalities or differences between their own program and another, which other program could have been authored by the teacher, the student themselves, or even another student. An example of such a cross-programmatic question appears at the bottom of Table II. To promote transferable learning of problem-solving schemata, these questions could ask students to identify recurring patterns such as roles of variables [6] or labeled subgoals [28], [29].

We have also focused our discussion on programming assignments that are closed-ended in the sense that they have a well-defined solutions that can be automatically assessed with predefined unit tests. QLCs could be generated from open-ended student code as well. Generating QLCs from open-ended code is simpler if the questions depend on static facts only, but QLCs with a dynamic nature may also be realistic if, for instance, the student has created unit tests.

VII. CHALLENGES AND LIMITATIONS

Assessments and self-explanation prompts direct students' attention to what is being asked and divert attention from other content that may be as important or even more so [26]. Whether QLCs are used as assessment items or SEPs, this presumably applies to them as well. This represents a technical and pedagogical challenge: we must generate questions that are relevant to the appropriate learning objectives. Having the technical ability to ask about a property of a program does not mean that it is appropriate to pose that question to students; whether a question is appropriate varies and depends on context and goals. Moreover, even where a QLC *is* appropriate for a particular situation, we must be alert to the possibility that its presence may de-emphasize other content that is also important but was not asked about.

Whether a QLC is good also depends on the learner's prior knowledge of programming and comprehension of their own program [43], [44]. In cases where students *do* understand the code that they have written, they may find QLCs tiresome. If QLCs are forced on students during problem solving, they may annoy the students. QLCs posed after passing unit tests may be perceived as a nuisance, as students may feel that getting the program to work is alone a sufficient assessment of their ability and that they have already done the hard part; "selling" the idea of QLCs to students appropriately in pedagogy may be key.

To avoid unsuitable and repeated questions, it may be necessary to track learners' history of programming assignments and QLCs. Teacher settings could define policies so that a particular QLC templates stops being used once a student has answered such questions correctly a set number of times.

Research suggests that simple multiple-choice questions are not ideal as self-explanation prompts [23]. Open-ended QLCs can be more varied and richer in content than multiple-choice or single-value questions; on the other hand, open-ended questions are currently difficult or impossible to assess automatically. Progress is expected in automatic grading of short-text answers about programming [45].

Having reviewed the literature on self-explanation, Rittle-Johnson and Loehr [26] note that "the substantial time demands of self-explanation raises the question of when alternative activities would more easily or effectively achieve the same learning outcomes." Similarly, the additional time expenditure from answering QLCs must be evaluated against the questions' possible benefits.

VIII. ONGOING AND FUTURE WORK

As the QLC concept is very broad and flexible, there is a lot of ground to be covered in investigating the technical challenges of QLC generation, different types of QLCs, and different pedagogical uses of QLCs.

A. Technical Development and Evaluations

We are currently working on QLC generation in an automatic assessment system that poses questions to students after unit tests pass, as described in Section VI-A. This software will largely follow the question-generation process outlined above in Figure 3, albeit initially it will have limited teacher configuration and template selection will be based on static facts only.

Using this implementation, we hope to evaluate:

- 1) to what extent students are capable of answering QLCs;
- 2) how students feel about QLCs;
- 3) if students' ability to answer QLCs correlates with other outcome measures; and
- 4) how all of the above depends on the type of QLCs.

Early work along these lines is underway. Moreover, we are conducting a preliminary investigation where we pose *manually created QLCs* to students after a programming assignment [46] (loosely similarly to our main scenario for automatic QLCs). In this pilot study, a system presents students with self-explanation questions that a researcher has tailored for the particular programming assignment and that are thus not chosen automatically or filled in based on facts collected from the student's program. The results from this pilot study have been tentatively promising so far: some students have reported that QLCs helped them to reflect on their program comprehension, while others noted that they had already thought about the QLC topics as they wrote their program. Furthermore, results confirm that many students indeed do have difficulties comprehending their own code and that the ability to explain one's own code correlates positively with

later success in the course. These early findings highlight the potential of QLCs to improve program comprehension and its assessment.

B. Further Opportunities

Beyond our primary QLCs-after-passing-unit-tests scenario, there are many paths for future research, such as the following.

QLCs could be posed to students after failed unit tests, before testing, or even during the program-authoring process, as outlined in Section VI-B above.

Different ways of assessing QLCs could be explored: automatic, peer, self, and teacher assessment.

A possible thread of future work is to correlate students' ability to answer QLCs with their self-explanation ability. This could provide insight into the mechanisms by which QLCs work (or fail to). Research could investigate whether students can be taught and motivated to use QLCs more effectively as self-explanation prompts and whether such teaching impacts on their QLCs responses and learning.

The impact of QLCs on plagiarism could be evaluated. Such research will need to weigh the possible benefits QLCs on plagiarism prevention against any negative consequences of mandatory and frequent QLCs.

Personalization of QLCs through learner modeling is an interesting and challenging area for future exploration. If the QLC generator could estimate each student's knowledge of different concepts, it could pose questions that suit the particular student better. For example, the questions might target possible misconceptions (analogously to [47]). Students' responses to QLCs could also be a source of data *about* misconceptions (analogously to [48]).

Finally, the techniques developed for generating QLCs might find applications in other contexts, such as generating practice questions from programs authored by teachers or peers.

IX. CONCLUSION

We have proposed the idea of automatically generating questions, QLCs, about students' program code for the students themselves to answer. This proposal opens up opportunities in deepening students' program comprehension, complementing automatic assessment, and discouraging plagiarism. We have suggested several scenarios in which QLCs might be used and outlined the software support needed for generating QLCs; work on one such system is underway.

The practical feasibility and effectiveness of QLCs are as yet untested, but this proposal sets down a research agenda for future investigations. We have painted a fairly broad landscape in which computing education researchers may work on QLCs. In ongoing work, we are exploring a sector of that landscape; we hope that others might do similarly.

REFERENCES

- [1] H. Keuning, J. Jeuring, and B. Heeren, "A systematic literature review of automated feedback generation for programming exercises," *ACM Transactions on Computing Education*, vol. 19, no. 1, 2018.

- [2] T. Auvinen, "Harmful Study Habits in Online Learning Environments with Automatic Assessment," in *2015 International Conference on Learning and Teaching in Computing and Engineering*. IEEE, 2015, pp. 50–57.
- [3] M. Kapur, "Examining Productive Failure, Productive Success, Unproductive Failure, and Unproductive Success in Learning," *Educational Psychologist*, vol. 51, no. 2, pp. 289–299, 2016.
- [4] K. Brennan and M. Resnick, "New frameworks for studying and assessing the development of computational thinking," in *Proceedings of the 2012 annual meeting of the American educational research association*, vol. 1. AERA, 2012.
- [5] R. Garcia, K. Falkner, and R. Vivian, "Instructional Framework for CS1 Question Activities," in *The 24th Annual Conference on Innovation and Technology in Computer Science Education*, ser. ITiCSE '19. ACM, 2019, pp. 189–195.
- [6] J. Sajaniemi and M. Kuittinen, "An Experiment on Using Roles of Variables in Teaching Introductory Programming," *Computer Science Education*, vol. 15, no. 1, pp. 59–82, 3 2005.
- [7] A. L. Santos, "Enhancing Visualizations in Pedagogical Debuggers by Leveraging on Code Analysis," in *Proceedings of the 18th Koli Calling International Conference on Computing Education Research*, ser. Koli Calling '18. ACM, 2018, pp. 1–9.
- [8] R. Lister, E. S. Adams, S. Fitzgerald, W. Fone, J. Hamer, M. Lindholm, R. McCartney, J. E. Moström, K. Sanders, O. Seppälä, B. Simon, and L. Thomas, "A multi-national study of reading and tracing skills in novice programmers," *ACM SIGCSE Bulletin*, vol. 36, no. 4, pp. 119–150, 12 2004.
- [9] Simon, "Assignment and sequence: Why some students can't recognise a simple swap," in *Proceedings of the 11th Koli Calling International Conference on Computing Education Research*, ser. Koli Calling '11. ACM, 2011, pp. 10–15.
- [10] K. Cunningham, S. Ke, M. Guzdial, and B. Ericson, "Novice Rationales for Sketching and Tracing, and How They Try to Avoid It," in *The 24th Annual Conference on Innovation and Technology in Computer Science Education*, ser. ITiCSE '19. ACM, 2019, pp. 37–43.
- [11] K. Cunningham, R. A. Bejarano, M. Guzdial, and B. Ericson, "I'm not a computer": How identity informs value and expectancy during a programming activity," in *The 14th International Conference of the Learning Sciences*, ser. ICLS 2020. ISLS, 2020, pp. 705–708.
- [12] U. Fuller, C. G. Johnson, T. Ahoniemi, D. Cukierman, I. Hernán-Losada, J. Jackova, E. Lahtinen, T. L. Lewis, D. M. Thompson, C. Riedesel, and E. Thompson, "Developing a computer science-specific learning taxonomy," *ACM SIGCSE Bulletin*, vol. 39, no. 4, pp. 152–170, 12 2007.
- [13] J. Salac and D. Franklin, "If they build it, will they understand it? Exploring the relationship between student code and performance," in *25th Annual Conference on Innovation and Technology in Computer Science Education*, ser. ITiCSE '20. ACM, 2020, pp. 473–479.
- [14] C. Kennedy and E. T. Kraemer, "Qualitative observations of student reasoning," in *The 24th Annual Conference on Innovation and Technology in Computer Science Education*, ser. ITiCSE '19. ACM, 2019, pp. 224–230.
- [15] C. Schulte, "Block model: An educational model of program comprehension as a tool for a scholarly approach to teaching," in *Proceedings of the Fourth International Workshop on Computing Education Research*, ser. ICER '08. ACM, 2008, pp. 149–160.
- [16] C. Izu, C. Schulte, A. Aggarwal, Q. Cutts, R. Duran, M. Gutica, B. Heinemann, E. Kraemer, V. Lonati, C. Mirolo, and R. Weeda, "Fostering Program Comprehension in Novice Programmers - Learning Activities and Learning Trajectories," in *Proceedings of the Working Group Reports on Innovation and Technology in Computer Science Education*. ACM, 12 2019, pp. 27–52.
- [17] M. Lopez, J. Whalley, P. Robbins, and R. Lister, "Relationships between reading, tracing and writing skills in introductory programming," *Proceedings of the Fourth International Workshop on Computing Education Research*, pp. 101–112, 2008.
- [18] A. Luxton-Reilly, Simon, I. Albluwi, B. A. Becker, M. Giannakos, A. N. Kumar, L. Ott, J. Paterson, M. J. Scott, J. Sheard, and C. Szabo, "Introductory programming: A systematic literature review," in *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, ser. ITiCSE 2018 Companion. ACM, 2018, p. 55–106.
- [19] B. Xie, D. Loksa, G. L. Nelson, M. J. Davidson, D. Dong, H. Kwik, A. H. Tan, L. Hwa, M. Li, and A. J. Ko, "A theory of instruction for introductory programming skills," *Computer Science Education*, vol. 29, no. 2-3, pp. 1–49, 2019.
- [20] S. Sentance, J. Waite, and M. Kallia, "Teaching computer programming with PRIMM: A sociocultural perspective," *Computer Science Education*, vol. 29, no. 2-3, pp. 136–176, 2019.
- [21] J. Sorva, "Visual Program Simulation in Introductory Programming Education," Doctoral dissertation, Department of Computer Science and Engineering, Aalto University, Espoo, Finland, 2012.
- [22] M. Hertz and M. Jump, "Trace-Based Teaching in Early Programming Courses," in *Proceedings of the 44th ACM Technical Symposium on Computer Science Education*, ser. SIGCSE '13. ACM, 2013, pp. 561–566.
- [23] K. Bisra, Q. Liu, J. C. Nesbit, F. Salimi, and P. H. Winne, "Inducing Self-Explanation: A Meta-Analysis," *Educational Psychology Review*, vol. 30, no. 3, pp. 703–725, 2018.
- [24] R. K. Atkinson, S. J. Derry, A. Renkl, and D. Wortham, "Learning from examples: Instructional principles from the worked examples research," *Review of Educational Research*, vol. 70, no. 2, pp. 181–214, 2000.
- [25] K. Kwon and D. H. Jonassen, "The influence of reflective self-explanations on problem-solving performance," *Journal of Educational Computing Research*, vol. 44, no. 3, pp. 247–263, 2011.
- [26] B. Rittle-Johnson and A. M. Loehr, "Eliciting explanations: Constraints on when self-explanation aids learning," *Psychonomic Bulletin & Review*, vol. 24, no. 5, pp. 1501–1510, 2017.
- [27] K. Bielaczyc, P. L. Pirolli, and A. L. Brown, "Training in self-explanation and self-regulation strategies: Investigating the effects of knowledge acquisition activities on problem solving," *Cognition and Instruction*, vol. 13, no. 2, pp. 221–252, 1995.
- [28] L. E. Margulieux, B. B. Morrison, M. Guzdial, and R. Catrambone, "Training learners to self-explain: Designing instructions and examples to improve problem solving," in *Proceedings of International Conference of the Learning Sciences*, ser. ICLS 2016. ISLS, 2016.
- [29] B. B. Morrison, L. E. Margulieux, and A. Decker, "The curious case of loops," *Computer Science Education*, vol. 30, no. 2, pp. 127–154, 2020.
- [30] C. Vieira, A. J. Magana, M. L. Falk, and R. E. Garcia, "Writing in-code comments to self-explain in computational science and engineering education," *ACM Transactions on Computing Education*, vol. 17, no. 4, pp. 1–21, 2017.
- [31] G. V. F. Fabric, A. Mitrovic, and K. Neshatian, "A Comparison of Different Types of Learning Activities in a Mobile Python Tutor," in *The 25th International Conference on Computers in Education*, ser. ICCE 2017. Asia-Pacific Society for Computers in Education, 2017, pp. 604–613.
- [32] L. Zavala and B. Mendoza, "On the Use of Semantic-Based AIG to Automatically Generate Programming Exercises," in *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, ser. SIGCSE '18. ACM, 2018, pp. 14–19.
- [33] A. Thomas, T. Stopera, P. Frank-Bolton, and R. Simha, "Stochastic Tree-Based Generation of Program-Tracing Practice Questions," in *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, ser. SIGCSE '19. ACM, 2019, pp. 91–97.
- [34] D. Lindberg, F. Popowich, J. Nesbit, and P. Winne, "Generating natural language questions to support learning on-line," in *Proceedings of the 14th European Workshop on Natural Language Generation*. Association for Computational Linguistics, 2013, pp. 105–114.
- [35] I. Labutov, S. Basu, and L. Vanderwende, "Deep questions without deep understanding," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, vol. 1. Association for Computational Linguistics, 2015, pp. 889–898.
- [36] T. Crow, A. Luxton-Reilly, and B. Wuensche, "Intelligent Tutoring Systems for Programming Education: A Systematic Review," in *Proceedings of the 20th Australasian Computing Education Conference*, ser. ACE '18. ACM, 2018, p. 53–62.
- [37] P. Brusilovsky, "Intelligent Tutor, Environment and Manual for Introductory Programming," *Educational and Training Technology International*, vol. 29, no. 1, pp. 26–34, feb 1992.
- [38] N. Truong, P. Roe, and P. Bancroft, "Static analysis of students' java programs," in *Proceedings of the Sixth Australasian Conference on Computing Education - Volume 30*, ser. ACE '04. Australian Computer Society, Inc., 2004, p. 317–325.
- [39] M. Striwe and M. Goedicke, "Using run time traces in automated programming tutoring," in *Proceedings of the 16th Annual Conference*

- on *Innovation and Technology in Computer Science*, ser. ITiCSE '11. ACM, 2011, pp. 303–307.
- [40] B. J. Ericson and B. N. Miller, “Runestone: A platform for free, on-line, and interactive ebooks,” in *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, ser. SIGCSE '20. ACM, 2020, pp. 1012–1018.
 - [41] A. Korhonen, T. L. Naps, C. R. Boisvert, P. Crescenzi, V. Karavirta, L. Mannila, B. N. Miller, B. B. Morrison, S. H. Rodger, R. J. Ross, and C. A. Shaffer, “Requirements and design strategies for open source interactive computer science eBooks,” in *ITiCSE 2013 Working Group Reports*, ser. ITiCSE -WGR '13. ACM, 2013, pp. 53–72.
 - [42] T. Sirkiä and J. Sorva, “How do students use program visualizations within an interactive ebook?” in *Proceedings of the Eleventh Annual Conference on International Computing Education Research*, ser. ICER '15. ACM, 2015, pp. 179–188.
 - [43] L. Corno, R. E. Snow *et al.*, “Adapting teaching to individual differences among learners,” in *Handbook of research on teaching*. New York: Macmillan, 1986, vol. 3, pp. 605–629.
 - [44] S. Kalyuga, P. Ayres, P. Chandler, and J. Sweller, “The expertise reversal effect,” *Educational Psychologist*, vol. 38, no. 1, pp. 23–31, 2003.
 - [45] M. Fowler, B. Chen, S. Azad, M. West, and C. Zilles, “Autograding explain in plain english questions using nlp,” in *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, ser. SIGCSE '21. ACM, 2021, p. 1163–1169.
 - [46] T. Lehtinen, A. Lukkarinen, and L. Haaranen, “Students Struggle to Explain Their Own Program Code,” in *Proceedings of the 2021 ACM Conference on Innovation and Technology in Computer Science Education*, ser. ITiCSE '21. ACM, to appear.
 - [47] L. Gusukuma, A. C. Bart, D. Kafura, and J. Ernst, “Misconception-Driven Feedback: Results from an Experimental Study,” in *Proceedings of the 2018 International Computing Education Research Conference*, ser. ICER '18. ACM, 2018, pp. 160–168.
 - [48] T. Sirkiä and J. Sorva, “Exploring programming misconceptions: An analysis of student mistakes in visual program simulation exercises,” in *Proceedings of the 12th Koli Calling International Conference on Computing Education Research*, ser. Koli Calling '12. ACM, 2012, pp. 19–28.