# Question Vault: Crowdsourcing Assessment Generation

**Derrick Coetzee**
University of California, Berkeley
Berkeley, CA
dcoetzee@berkeley.edu

**Colorado Reed**
University of California, Berkeley
Berkeley, CA
cjrd@cs.berkeley.edu

**ABSTRACT**
Question Vault is a platform for collaborative construction of high-quality educational assessments such as homework and test questions. Inspired by the online encyclopedia Wikipedia, it invites users to collaborate by each working on different aspects of assessments and refining them over a long period of time. In addition to creating free assets that can be deployed in existing courses, we believe our platform will also benefit learners who contribute to it, and create a corpus for future study. This work describes our detailed design, prototype, related work, and open questions.

**Author Keywords**
Assessment; collaboration; wiki; Wikipedia.

**ACM Classification Keywords**
H.5.3. Group and Organization Interfaces: Computer-supported cooperative work; K.3.1. Computer Uses in Education: Collaborative learning

## MOTIVATION

Designing effective assessments is a vital part of any instructor's role. They provide immediate guidance to the student and instructor on how to refine instruction (*formative assessment*), and determine whether in the end the student learned the material (*summative assessment*). However, the design of high-quality assessments that perform these goals accurately and efficiently is a taxing process requiring successive *refinement* based on student feedback over a long period of time. Assessments may also need to be carefully *selected* from a pool of candidates, based on time constraints.

In light of this, it's unfortunate that many instructors teaching similar courses are compelled to repeat this taxing work. A few achieve high-quality results, but share them with no one; many others fail to achieve a similar level of quality. This stems not from a lack of willingness to share, but from a number of pragmatic barriers to sharing that impose an unacceptable burden. By lowering the barrier to sharing—making unencumbered sharing *automatic* as soon as assessments are created—we anticipate that sharing can be greatly increased.

Final report for CS 294 (Autograding), Fall 2013, University of California, Berkeley.


Another issue we aim to address is that of assessment *completeness*. Bare assessments that include only a question or problem statement are much less valuable than assessments that include detailed and well-referenced explanations of both correct and incorrect solutions, hints for students who are uncertain how to proceed, and detailed rationales for instructors who need help deciding whether and how to incorporate the assessment in their courses. But the domain experts who design assessments often have no time or inclination to develop extensive supplementary materials. A collaborative system can enable domain experts to design good assessments, while others create supplementary materials. Even trivial contributions such as votes, difficulty ratings, and brief comments can in aggregate be invaluable for assessment refinement and selection. In addition to making assessments more useful, in the process of developing these materials, issues with the original assessment may be uncovered.

A third problem is that of assessment *archival*. Valuable assessments which are designed by an individual are frequently lost when that person becomes unavailable. Additionally, many individuals take inadequate precautions to protect assessments from accidental loss. A central database that is managed by archivists and distributed widely can ensure preservation.

Finally, we consider the problem of assessment *currency*. Terminology can become obsolete or ambiguous, new results can render correct solutions incorrect, and content can be rendered objectionable by changes in culture. This can be compared to the problem faced by paper encyclopedias, which, being frozen in time, fail to capture changes in language as well as dramatic political or scientific changes. The online encyclopedia Wikipedia effectively addresses these issues by having volunteers maintain the currency of all articles on an ongoing basis. By recruiting volunteers and empowering them to freely update assessments, similar benefits can be achieved in our setting.

Our solution to all of these problems is Question Vault, a "Wikipedia of assessments": a wiki where a variety of volunteers collaborate to create and maintain high-quality, reusable assessments replete with supplementary materials. In the following sections, we summarize our platform's design, describe details of our initial prototype implementation, discuss related work, and finish by outlining remaining future work.

## DESIGN

### Publicly Accessible

To meet our goals as outlined above, we adopt four main features from Wikipedia:

1. All content is publicly available under a free license.
2. All content can be edited by anyone.
3. All changes to content are tracked and can be easily reversed.
4. Each page has an associated discussion page where collaborators discuss proposed changes.

The third feature (change tracking) makes the second feature (world editable) feasible: although some editors will be malicious and some will make mistakes, any such changes can be reversed by other editors who later review these changes. Making the project world editable is valuable for lowering barriers to casual participation and building an adequate volunteer force for maintenance. The fourth feature (discussion pages) is necessary to organize collaboration and resolve disputes regarding what content should be included.

The first feature (content is publicly available under a free license) makes sharing assessments among instructors simple: they simply copy and use the assessments. There is no need to validate their status, no need to ask permission, and no privileging of accredited instructors over unaccredited tutors. However, this feature poses a challenge for summative assessment, since it means that solutions to problems can also be easily copied by students. There are a number of strategies for addressing this:

- **Large corpus**: If an exam is administered in a controlled environment, students cannot anticipate which of the many relevant questions will be on the exam.
- **Mixing**: Students who receive good scores on public assessments but poor scores on original assessments are likely to be copying solutions.
- **Modification**: Instructors can modify the original question to create similar questions with different solutions.
- **Parameterization**: The platform can automatically generate random instances of a general assessment template. Only the person who generates an instance receives solutions for that instance.
- **Plagiarism detection**: Assessments which ask students to explain their reasoning also facilitate detection of students who copy solutions by comparing style.

#### Progressive Hints

*Progressive hints* are a series of hints designed to help students solve a problem, each giving more information than the last about the solution. The interface reveals them one at a time, and students consume just enough hints to solve the problem on their own. Progressive hints are a common feature in intelligent tutoring systems, and find theoretical justification in Vygotsky's zone of proximal development [8].

By including progressive hints, questions become accessible to a broad range of students of varying competency. Instructors who copy assessments may include some hints to reduce difficulty, or may use hints as guidance when assisting students. Hints are valuable for constructing student models, since they can effectively distinguish students with slightly different levels of competency [8].

#### Tracking Student Progress and Integration

Although our license permits instructors to copy content to another website, if we encourage them to instead direct students to our platform, we can provide value-added services like tracking each student's history, progress, and performance, creating student models describing competency in various skills, and so on. We can use this data to recommend additional assessments or instructional materials to students, or to provide feedback to instructors.

In order to make the option of using our platform directly even more attractive, we can integrate our software with e-learning platforms, achieving the look-and-feel of the actual course website and even report grades back to the course system. We would gather extensive data about our assessments, and students would receive the latest up-to-date, enriched content.

#### Collecting Student Feedback

Assessment refinement and selection requires considering the feedback of students. We automatically gather information from all the students who answer questions directly on the website. This includes tacit feedback (the number of times a particular question has been answered, the number of times each response was chosen) as well as active feedback like upvotes/downvotes, difficulty ratings, and brief comments.

Such data can be collected with little overhead, so that students can focus primarily on completing assessments. By aggregating it and placing it on a public statistics page, contributors can immediately act on it to improve the assessment. By retaining this data, we allow students to review their work, while also facilitating sophisticated future analyses, such as identifying correlations between questions and constructing student models.

#### Question Sets and Index Pages

In a database with a large number of questions, an important goal is to identify the most useful and relevant questions for a certain application. This is supported by several mechanisms:

1. Assessments are given automatic ratings based on student feedback and, where applicable, the current user's student model.
2. Assessments are tagged based on their topic area, difficulty, and relevant courses.
3. Contributors maintain a hierarchy of index pages that summarize what topic areas are covered and where to find relevant questions.

Questions are organized into question sets designed to be taken all at once sequentially. Question sets ensure uninterrupted navigation through a set of questions, and results can be aggregated to provide detailed feedback on competency in specific skills. For instructors, question sets serve as recommendations for questions to be used together.

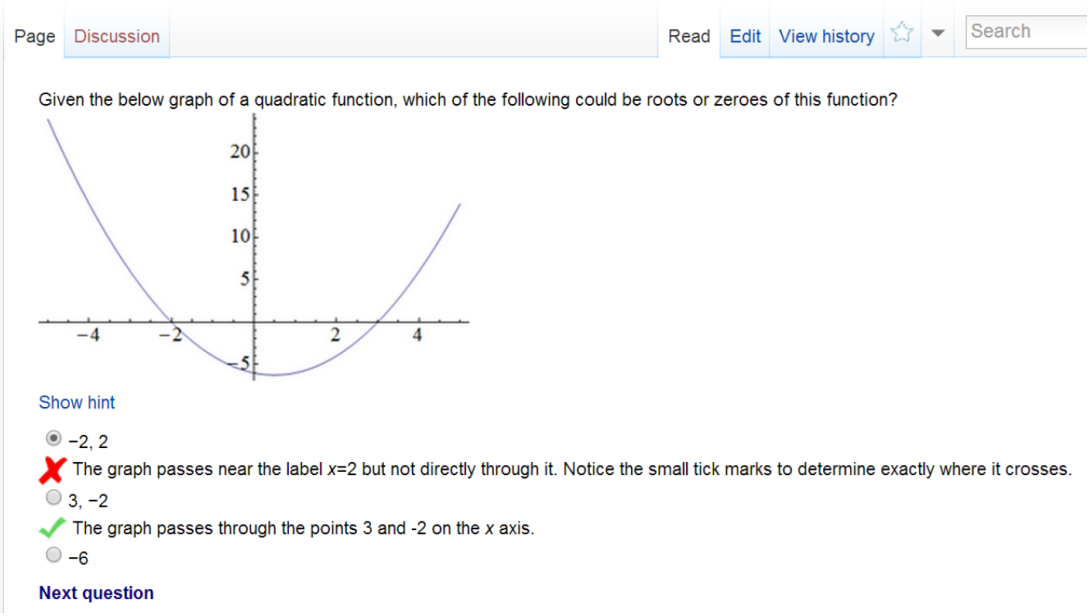


**Figure 1. Screenshot of student answering question. The student has selected one incorrect and one correct answer, and explanations for both are shown. All hints are currently hidden. All content can use any markup supported by the base software, including images and math notation.**

## PROTOTYPE IMPLEMENTATION

To simplify implementation, we initially support only simple multiple-choice questions. Our prototype was built on MediaWiki, [2] the same software used by Wikipedia, and used its extension framework to provide our new functionality. Existing functionality enabled editing, reviewing changes, leaving comments, and discussing proposed changes with collaborators.

Our extension shows the question along with answer choices, and when the student chooses an answer it shows an explanation for why that response is correct or incorrect. The student continues to choose answers until they are satisfied, and may continue to the next question at any time. All hints are initially hidden, and are shown one by one as the student clicks on the "show another hint" link. Figure 1 shows a screenshot of a student answering a question on our website.

Questions are represented as XML blocks with elements for the question text, each response (including explanation), and progressive hints. Editors manipulate this representation directly. This approach allows us to leverage the existing editing and history tools, but it is not as accessible as the form-based editors used by tools like PeerWise and LearningPod; providing such an editor is future work.

Some pages were used to store individual assessments, while others (*index pages*) were used to store manually-curated lists of assessments relevant to particular topic areas. These index pages are organized into a hierarchy and divided into subsections, and are intended to assist students and instructors in locating high-quality relevant content.

Currently, the site has only undergone preliminary development and has not yet recruited contributors. Only limited test content is available, and student progress is not yet tracked.



## RELATED WORK

A Contributing Student Pedagogy (CSP) is a method of teaching where students contribute to the learning of others, with a focus on content contribution [3, 5]. The PeerWise system [4] is a CSP system that has several notable similarities with our design: it asks non-experts (students) to contribute multiple-choice questions, which other students answer, and student feedback is collected and used to highlight the best questions.

However, CSP systems like PeerWise differ in two important ways from our system: 1. their primary goal is *learning through content contribution*, rather than *construction of high-quality content*; 2. CSP systems are typically used in short-term settings, such as a single course offering or a single assignment, whereas Question Vault assessments are intended to endure and be refined over a period of many years. These explain several key design differences: in PeerWise, questions are visible only to other students in the same course, are editable only by the student who created the question, and are discarded after the course completes. In Question Vault, we encourage instructors, students, and enthusiasts to all collaborate to their fullest extent, and avoid duplicating work that has been done before; because learning is secondary, we are not concerned about displacement of students by others.

However, it is possible for the CSP model to be embedded into a Wikipedia-style system like Question Vault, as modelled by the Wikipedia Education Program [6]. In this program, students contribute directly to Wikipedia as a major part of their course. They are mentored by experienced Wikipedia contributors, and their contribution is reviewed by other students in their class according to a rubric; optionally,

it may be reviewed by experts. Both articles and reviews are graded. We believe that if students contributed question sets to Question Vault with similar high stakes and strong guidance, that we can not only produce consistently high-quality assessments, but we can achieve a richer CSP experience where students are asked not just to create content but to evaluate, improve, and value content.

Other sites share our goal of building and sharing high-quality content, but take a different approach. Learningpod [1] is a commercial site that uses a mixture of freely-licensed questions, licensed commercial content, and volunteer contributions to build a public database of practice questions. However, their restrictive license does not permit content to be copied to other websites, and questions are (as in PeerWise) editable only by their author. Because collaboration is restricted, history and discussion pages are unnecessary and omitted. As with PeerWise, many volunteer-provided questions are low-quality or incomplete and there is no way to refine or improve them unless the original author chooses to. Despite these important limitations, the site also provides a model for our work with its appealing and accessible UI and its combination of existing assessments from disparate sources.

## FUTURE WORK

Our design as outlined above is only a skeleton of essential features. A number of practical problems remain to be resolved in order to make the site useful in practice and nurture an active community.

The first is the need to import starting content. A large number of high-quality freely-licensed assessments have been made available as part of programs like OpenStax and the Saylor Foundation, but these need to be converted *en masse* to a format which is interactive and ready-to-edit. Negotiating with and licensing content from instructors and textbook authors can generate even more content. This starting content serves both to draw visitors in search of good questions to the site, and to give volunteers useful tasks to perform (adding supplementary content to each assessment).

The second is the need for clear guidelines regarding how to go about designing assessments. Successful wikis like Wikipedia invariably rely on policies and guidelines to keep contributions in the scope of the project, keep quality high, and defuse disputes. This may require distilling the hard-won lessons of experienced assessment designers in order to promote quality and accessibility [9, 10].

An open design question is how to best combine the strengths of automatic curation of assessments (based on student feedback) and manual curation (index pages). Automatic curation is effective at highlighting assessments that are unexpectedly engaging, or conversely, of identifying problematic questions with subtle problems. Manual curation is more effective at selecting structured question sets that complement one another and follow a logical progression. Side-by-side or mixed lists that combine these strategies may be most effective.

An important step for evaluating our website's UI is to instrument it to record all known user actions. Detailed clickstreams can be used to perform path analysis [7], identify UI issues, and predict user behavior such as abandonment in order to intervene.

Because wikis are inconsistent, some assessments may be incomplete, lacking information about what topics they pertain to, level of difficulty, and so on. Tacit user data can be used to help cope with this sparse data problem: collaborative filtering can be used to identify suitable questions for a user without needing to explicitly know what topics those questions assess. Difficulty can be inferred by examining other students' performance on the same question as well as other questions. Both topic and difficulty can be inferred from correlations between items: highly correlated items tend to be on similar topics and of similar difficulty.

## ACKNOWLEDGMENTS

We thank the course instructors, Armando Fox and John Canny, and other students for useful feedback on our design.

## REFERENCES

1. Learningpod, 2013. `http://www.learningpod.com/`.
2. MediaWiki, 2013. `http://www.mediawiki.org/`.
3. Collis, B., and Moonen, J. *Flexible Learning in a Digital World: Experiences and Expectations*. Open and distance learning series. Kogan Page, 2001.
4. Denny, P., Hamer, J., Luxton-Reilly, A., and Purchase, H. Peerwise: Students sharing their multiple choice questions. In *Proceedings of the Fourth International Workshop on Computing Education Research*, ICER '08, ACM (New York, NY, USA, 2008), 51–58.
5. Hamer, J., Cutts, Q., Jackova, J., Luxton-Reilly, A., McCartney, R., Purchase, H., Riedesel, C., Saeli, M., Sanders, K., and Sheard, J. Contributing student pedagogy. *SIGCSE Bull. 40*, 4 (Nov. 2008), 194–212.
6. Infeld, D. L., and Adams, W. C. Wikipedia as a tool for teaching policy analysis and improving public policy content online. *JPAEJOURNAL OF PUBLIC AFFAIRS EDUCATION VOLUME 19 NUMBER 3* (2013), 445.
7. Montgomery, A. L., Li, S., Srinivasan, K., and Liechty, J. C. Modeling online browsing and path analysis using clickstream data. *Marketing Science 23*, 4 (2004), 579–595.
8. Murray, T., and Arroyo, I. Toward measuring and maintaining the zone of proximal development in adaptive instructional systems. In *Intelligent Tutoring Systems*, S. Cerri, G. Gouardres, and F. Paraguau, Eds., vol. 2363 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2002, 749–758.
9. Oosterhof, A. *Developing and using classroom assessments*. ERIC, 1999.
10. Thompson, S., Thurlow, M., and Malouf, D. B. Creating better tests for everyone through universally designed assessments. *Journal of Applied Testing Technology 6*, 1 (2004), 1–15.