Improving Teacher-Developed Assessments and Items

Adisack Nhouyvanisvong, Ph.D.

October 2015

Naiku

Naiku is a next generation assessment platform, providing teachers with comprehensive assessment tools to help teachers collect data about their students to make informed instruction.

7805 Telegraph Road, Suite 111
Bloomington, MN 55438
Phone: 612-346-2458
E-Mail: info@naiku.net
Web: www.naiku.net

Introduction

Significant research supports the use of frequent formative assessment to aid teaching and learning, particularly when the assessment is well aligned with curriculum and instruction. Yet teachers may not have had appropriate training to adequately develop and evaluate the quality of their assessments and test items.

We assess our students in order to gain visibility into what they know and what they do not know. We seek to gain insight into what they can and cannot do. Proper assessment results inform teacher instruction and student learning.

To get proper assessment results, the assessments and items that we build and create need to be valid and reliable. Consistent and reliable assessments and items allow teachers and students to make correct inferences about teaching and learning.

This paper gives teachers guidelines on how to create reliable and valid assessments and procedures to evaluate and improve the quality of their assessments and items. The following topics are covered:

- Guidelines for Creating a Test Blueprint
- Guidelines for Writing Assessment Items
- Judgmental Item Improvement Procedures
- Empirical Item Improvement Procedures

Guidelines for Creating a Test Blueprint

Before writing new assessment items, it is important to state the obvious upfront. To write quality items, it is imperative to have clear targets. What **learning targets** are to be assessed? Are the learning targets clear to both the teacher and the students? Once the learning targets are understood, the design of the assessment and the choice of items types follow.

Generally, we develop and use assessments because we want teachers to receive feedback about the effectiveness of their teaching. We also want students to receive feedback about the effectiveness of their learning. Assessments, when done properly, close the loop between curriculum and instruction.

So we need to define a clear purpose for the assessment. Do we want to know how our students are learning math in general? Or do we want to know how they are learning numbers and operations? Do we want to know if they understand science? Or do we want to know if they understand the water cycle? When we design assessments, we need to be clear and specific on the purpose of the assessment.

With a clear purpose, we can define the expected outcomes of the assessment. Will the assessment identify which students are proficient or not proficient on specific standards or learning targets? Will the assessment identify student strengths and weaknesses by skill or learning target? Or will the assessment identify or reliably sort our student by ability?

Once we have clearly defined the purpose and expected outcomes from the assessment, we need to identify the standards or learning targets that will be assessed. So how do we know what standards or learning targets to assess or include? If you've looked at the Common Core or your state standards, you know there are many more standards that you can possibly have time to assess.

Since we cannot assess every standard, we must identify the priority standards (Ainsworth, 2015). That is, find and prioritize those standards that students must know and be able to do. These are the key standards. They are clear and can be understood by the teacher and the student. And, most importantly, these are standards that are measurable.

Priority standards receive greater instruction and assessment emphasis. However, we are not eliminating any standards. All standards must be taught. The supporting standards or non-priority standards play a role to help students understand the priority standards. So even though these standards may not be assessed or measured, they must still be taught.

Once you have identified the learning targets to assess, you need to decide on the levels of cognitive complexity or rigor to assess. Most teachers are familiar with Bloom's taxonomy or the Revised Bloom's taxonomy. In this taxonomy, there are 6 levels of cognitive rigor. Another common taxonomy that is used, particular in assessment, is Norm Webb's Depth of Knowledge. In this taxonomy, there are 4 levels. Both are good

taxonomies. My advice is to use the one that you are most familiar with.

Revised Bloom's Taxonomy (2001):

1. Remember
2. Understand
3. Apply
4. Analyze
5. Evaluate
6. Create

Webb's Depth of Knowledge (2002):

1. Recall/Reproduction
2. Skills and Concepts
3. Strategic Thinking and Reasoning
4. Extended Thinking

The last thing to consider in the design of the assessment is the type of items to include on the assessment. Generally, there are two types of items. Items that require students to select an answer are called selected-response items. Items that require students to construct an answer are called constructed-response items.

Common selected-response items are multiple-choice, trule/false, yes/no, and matching items. These items allow you to sample or measure a broader content area or more learning targets. They can be quickly and objectively scored. Because you can use more of these items, they lead to higher reliability and greater efficiency in the assessment.

However, these items often tend to over emphasize lower level recall and reproduction of knowledge. They can be used to assess higher cognitive levels, but it may be harder to write those Items. Also, since students are selecting responses, they are not constructing or writing, which is the main criticism of selected response items.

Constructed-response items, on the other hand, tend to be and can be more cognitively challenging for the students. They are asked to generate a response. From these responses, particularly the wrong responses, teachers get a good source of data and see the common errors or misconceptions that the students have. However, constructed-response items take more time to score and to write. Additionally, because they are subjectively scored, they tend to lead to lower assessment reliability.

The purpose, the outcomes, the learning targets to be assessed, the level of cognitive complexity to use, and the types of items to use make up the test blueprint. The test blueprint is your plan for assessment construction.

Below is an example of a test blueprint. This is a basic blueprint that uses Webb's depth of knowledge and generic "learning targets" (you should replace these with specific Common Core State Standards or your local standards or learning targets). More complex blueprints may include the item type.

Assessment Name:	*Name of the assessment*									
Assessment Purpose:	*Describe the purpose(s) of the assessment*									
Administration:	*Describe how and when the assessment is to be administered*									
Use of Results:	*Describe how the results of the assesment will be used.*									
	Webb (1997) Depth of Knowledge									
	Recall		Use		Strategic		Extended		**Total**	
	#	Points	#	Points	#	Points	#	Points	#	Points
Learning Target 1	3	3	1	1	1	2			5	6
Learning Target 2	2	2	2	2			1	2	5	6
Learning Target 3	1	1	2	2	1	1	1	2	5	6
Learning Target 4			2	2	2	2	1	2	5	6
Total	6	6	7	7	4	5	3	6	20	24

This blueprint allows teachers to see how many items and points there will be on the assessment and what type of coverage there is across the learning targets. The key point is to make sure the assessment covers all the learning targets and is spread across the cognitive levels. We would not want all the questions and points to come from one target or one cognitive level. The more items or points you have per learning target, the more reliable of an estimate of the student's ability you will receive. To have a sufficiently reliable estimate of the score (i.e., the student's ability) by learning target, it is best to include at least 5 points per target.

Guidelines for Writing Assessment Items

In this paper, we consider the four most common item types developed by teachers for their classroom assessments. These item types can be developed by any teacher and used effectively in the classroom.

Multiple-Choice: Consists of an introductory part called a *stem* (either a statement or question) and a set of answer choices, one of which is the answer. This item type is used when there is only one correct answer and several plausible incorrect choices that can help teachers diagnose student misunderstandings.

True/False: Consists of a statement or proposition for students to verify as true or false. These are best used when there is a large body of content to be tested.

Matching: Consists of two lists or phrases where the entries on each list are to be matched. The list of the left contains the *premises*. The list on the right contains the *responses*. Matching items are best used when there are many related thoughts or facts for students to associate to each other.

Constructed-Response: Items that require students to generate their own response, in contrast to selecting a response, are called constructed-response items. These can be simple items such as **fill-in-the-blank** items or more complex such as **extended-response** items. These types of items often require students to demonstrate more in-depth understanding than selected-response items.

General Item Writing Guidelines

Developing assessment items is a relatively easy task. However, developing quality assessment items can be more challenging. The following general guidelines can help improve the quality of all items, regardless of item type. For more detail guidelines with example items, see Chappuis, Stiggins, Chappuis, & Arter (2012) and Popham (2011).

1. **Keep wording simple and focused.** This adage in effective written communication is just as applicable when writing items as when writing expository essays.
2. **Eliminate clues to the correct answer**. Be careful that grammatical clues do not signal the correct answer within an item or across items on a test.
3. **Highlight critical or keywords**. Critical words such as "MOST", "LEAST", "EXCEPT" can be easily overlooked so highlight them for the student.
4. **Review and double-check the scoring key**. As in any written piece of work, it is always good practice to review and double check, especially the scoring key.

Multiple-Choice Guidelines

1. **Make the stem a self-contained problem.** Ask a complete question with all the necessary information to answer it contained in the stem. This aids in clarity and makes it easier for students to read through tersely stated answer choices to select the correct answer.
2. **Make all answer choices plausible.** All answer choices must be plausible with only one correct answer. Incorrect choices must be reasonable so that they can't be ruled out without having knowledge or proficiency in the content being assessed.
3. **Keep length of answer choices similar.** Answer choices that are parallel and of similar length do not cue the correct answer.
4. **"of-the-above**". Never use "all of the above" as an option. If students know A and B are correct, but not C or D, then they can simply choose E ("all of the above"). Including "none of the above" adds difficulty to the item because students can't simply assume and guess that one of the options is correct.

True/False Guidelines

1. **Include only one concept or idea in each statement.** Make the item completely true or false as stated. Don't make it complex, which will confuse the issue and students.
2. **Avoid using negative statements.** Negatives are often harder to understand and comprehend, especially double negatives. Use negative statements sparingly and avoid using double negatives at all cost.

Matching Guidelines

1. **Provide clear directions on how to make the match.** It's important to let students know what the basis of the matching is supposed to be. If responses can be matched to multiple premises, make that clear in the directions.
2. **Use homogeneous lists of premises and responses.** Keep the set of premises and response homogeneous. For example, don't mix events with dates or names.
3. **Keep responses short and brief.** Premises should be longer. Thus, the responses should be short and brief and parallel in their construction.

4. **Use more responses than premises.** When there are more responses than premises, this prevents students from arriving at an answer through a process of elimination.

Constructed-Response Guidelines

1. **Use Direct Questions.** For most items, it is better to use direct questions rather than incomplete statements. This is especially true for constructed-response items with young students. Students are less likely to be confused by direct questions and they lead the teacher to avoid ambiguity in the item.
2. **Encourage Concise Response.** Responses to short-answer items should be *short*. Although constructed-response items are open-ended, they should be written so that they encourage short concise answer responses.
3. **Keep at the End.** When constructing these items as fill-in-the-blank items, place the blank or blanks at the end of the incomplete statement. Blanks at the beginning are more likely to confuse the students.
4. **Limit the Blanks.** Use only one or two blanks at most. The more blanks there are, the more likely it is to confuse the students.
5. **Use Rubrics.** For extended-response items that require more complex answers and higher-order level of thinking from students than a single word (or number) answer for fill-in-the-blank item, the desired student response can be as long as a paragraph or as extended as an essay. This item type requires more teacher time to score as the teacher must score it manually. To make scoring easier and fairer to students, use a rubric or scoring guide when judging student work. This makes the evaluation criteria clearer to both the teacher and more importantly, to the student.

Judgmental Item Improvement Procedures

As with any form of written communication, it is wise to create drafts and then review and edit them as necessary before making them final. The same adage holds true for writing assessment items. Reviewing and editing items is a judgment-based procedure, whether they be your own or that of others. There are three sources of test-improvement judgments that you should consider: (1) yourself, (2) your colleagues, and (3) your students.

Judging Your Own Items

It is often best (and sometimes the only option) to review and try to improve your own items. Popham (2011) provides the following guidelines to make the review of assessment items systematic and strategic.

1. **Adherence to guidelines**. When reviewing your own items, be sure to be familiar with general item-writing guidelines. Use them to find and fix any violations of item writing principles.

2. **Contribution to score-based inference**. The reason we assess is to make valid score-based inferences of student knowledge and ability. As you review each item, consider whether the item does in fact contribute to the kind of inference you desire to make about your students.

3. **Accuracy of content**. Sometimes, previously accurate content is contradicted by more recent content or findings (e.g., Pluto). Be sure to see if the content is still accurate. And above all else, always make sure the key is correct.

4. **Absence of content gap**. Review your test as a whole to ensure that important content is not overlooked. This is where your test blueprint will come in handy. Identify the priority standards and make sure your assessment content adequately covers those standards.

5. **Fairness**. Assessment items should be free of bias. They should not favor one group nor discriminate against another. Be attentive to any potential bias so that you can eliminate them as much as you possibly can.

Collegial Judgments

Often teachers work in teams or in professional learning communities (PLCs). This school environment provides a great opportunity for teachers to enlist the help and judgment from those colleagues whom they trust. When you ask another teacher to review your items, be sure to provide them with guidelines and

criteria for review. You can provide them with a summary of the five criteria above. If you have item-writing guidelines that you followed when you developed the items, provide those resources to them too.

Teachers are busy professionals. It is often hard to find time to review your own items, let alone find and ask a fellow teacher to review them for you. If your school is part of a large district, there may be a district staff member who specializes in assessment. This person would be a great resource to have review your assessment develop procedures and the quality of your items.

Student Judgments

Another source of good judgment data can come from the students themselves. Students have a lot of experience taking tests and reading through a lot of items. They are often happy to provide you with information to help you improve your assessments.

When asking for student feedback, be sure to ask them after they have completed the assessment. Students should not be asked to take the test and review it at the same time. Doing both simultaneously may lead to poor test performance and poor feedback.

After students have completed the test, give them an opportunity to provide you with feedback on the assessment as a whole and on each specific item. Gaining information as to whether they thought the directions were clear and if any of the items were confusing is valuable information to inform your assessment creation.

Giving students the opportunity to provide feedback to you can not only help you improve your assessments and items, it can also help you further engage your students and build a critical and trusting teacher-student relationship.

Empirical Item Improvement Procedures

You've written your assessment items adhering to item writing guidelines and best practices. You've reviewed them and had a colleague review them before you gave the assessment to your students. And perhaps, students also provided feedback after they completed the test. Now that your students have taken the test, this provides another rich source of valuable information that you can use to improve your assessments and items.

These empirical methods are based on numbers and statistics. But they need not be daunting. If you deliver your tests with online assessment software, all the numbers will most likely be computed for you. All you need to do is know a little about these statistics and what they can tell you.

Reliability Indices

Reliability refers to the expected consistency of test scores. As shown in the formula below, the reliability coefficient expresses the consistency of test scores as the ratio of true score variance to total score variance (true score variance plus error variance). If all test score variance were true, the index would equal 1.0. Conversely, the index will be 0.0 if none of the test score variance was true. Clearly, a larger coefficient is better as it indicates the test scores are influenced less by random sources of error. Generally speaking, reliabilities go up with an increase in test length and population heterogeneity and go down with shorter tests and more homogeneous populations.

$$\rho_X^2 = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

Although a number of reliability indices exist, a frequently reported index for achievement tests is Coefficient Alpha, which indicates the internal consistency over the responses to a set of items measuring an underlying trait. From this perspective, Alpha can be thought of as the correlation between scores if the students could be tested twice with the same instrument without the second testing being affected by the first. It can also be conceptualized as the extent to which an exchangeable set of items from the same domain would result in similar ordering of students.

For large-scale educational assessments, reliability estimates above .80 are common. A reliability coefficient of 0.50 would suggest that there is as much error variance as true-score variance in the test scores.

For classroom assessments, where the number items on a test may be shorter and the number of students

taking a test may be fewer than on large-scale assessments, it is still important to consider and look at the reliability estimate of the test when appropriate, such as when giving midterms or final assessments. When important judgments or inferences are to be made from the test scores, it is important to make sure that those test scores are sufficiently reliable. For classroom assessments, it is desirable to have a test reliability greater than 0.70.

Difficulty Indices

At the most general level, an item's difficulty is indicated by its mean score in some specified group.

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

In the mean score formula above, the individual item scores (x_i) are summed and then divided by the total number of students (n). The mean score is the proportion correct for the item. This is also known as the p-value. In theory, p-values can range from 0.0 to 1.0 on the proportion-correct scale. For example, if an item has a p-value of 0.92, it means 92 percent of the students answered the item correctly. Additionally, this value might also suggest that: 1) the item was relatively easy, and/or 2) the students who attempted the item were relatively high achievers.

For selected-response items such as multiple-choice or true/false items, it is important to view the p-value in relationship to the chance probability of getting the answer correct. For example, if there are four response options on a multiple-choice item, by chance alone, a student would be expected to answer the item correctly a quarter of the time (p-value = 0.25). For true/false items, the chance of answering that item correctly is 0.50.

Discrimination Indices

Discrimination is an important consideration in assessment theory. The use of more discriminating items on a test is associated with more reliable test scores. At the most general level, item discrimination indicates an item's ability to differentiate between high and low achievers. It is expected that students with high ability (i.e., those who perform well on the assessment overall) would be more likely to answer any given item correctly, while students with low ability (i.e., those who perform poorly on the assessment overall) would be more likely to answer the same item incorrectly.

Most often, Pearson's product-moment correlation coefficient between item scores and test scores is used to indicate discrimination. This is also known as the item-total correlation. The correlation coefficient can range from -1.0 to +1.0. A high item-total correlation indicates that high-scoring students tend to get the item right while low-score students tend to answer the item incorrectly; this indicates a good discriminating item. Good test items are expected to yield positive values (e.g., greater than .30). Very good items typically have discrimination values greater than 0.40. Items with low discrimination values (< 0.20) and particularly those with negative values should be reviewed and revised if possible. It not, they should be discarded and not used on future assessments.

Distractor Analysis

In addition to using the statistics provided above, it is often necessary and informative to look deeper at how the items perform. A distractor analysis gives us this look into how the students actually performed on the item. This analysis identifies the number (or percent) of students who picked each option on a selected-response item. It also identifies the number of students who provided an incorrect response to constructed-response items.

For example, high quality multiple-choice items should have at least some students picking each of the response options. Remember, all distractors should be plausible, thus there should be students who pick that distractor. These distractors can tell us what misinformation or miscalculations the students make. Items with distractors where no student selected that option should be reviewed and revised if possible.

Putting It Into Practice

Writing quality assessment items and reviewing and evaluating assessment and item performance need not be daunting and tedious tasks. They can be made much easier with the proper tools. Good Next Generation Assessment platforms such as Naiku provide teachers with these tools. Below, the item writing and item evaluation processes are illustrated with Naiku.

Item Writing

Item creation forms should be simple and intuitive. It should provide all the necessary elements at the forefront. Less often used features should be hidden and made easily available when needed.

Below is the form for writing multiple-choice items in Naiku. The stem is provided at the top. This is where the question or statement goes. The standard four-option choices are provided next. More options can be added or options can be reduced with either a click of the + or – buttons. Math formulas, formatted text, rationale statements, and standards alignment can be added by clicking on the appropriate buttons.

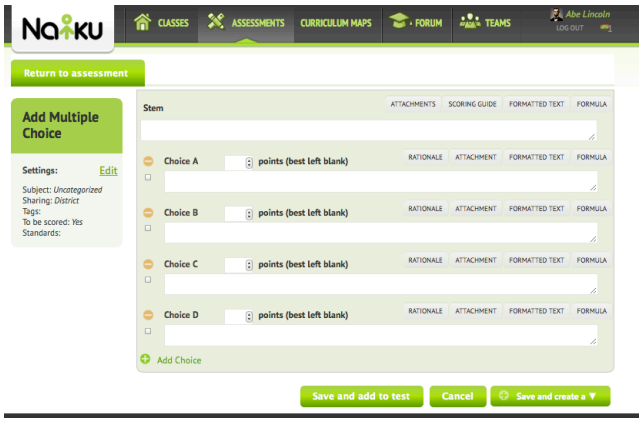

Figure 1. Multiple-Choice Item Creation in Naiku.

Forms for other item types should be equally easy and intuitive to use.

Item Evaluation

Equally important to providing a good tool for creating items, an assessment platform should provide tools to review and evaluate items. Below, I highlight some of the features of Naiku that allow teachers to review and evaluate item.

Teachers can review and provide comments to their own items. In addition, other teachers can review and provide comments to the creator of the item. There are multiple ways to provide feedback for your colleagues. The first and simplest method is to vote the item up/down to quickly indicate your level of satisfaction with the item. The second and more useful method is to leave and provide constructive feedback in the comment box. It is important to tell the writer how the item can be improved, and if so, what changes need to be made.

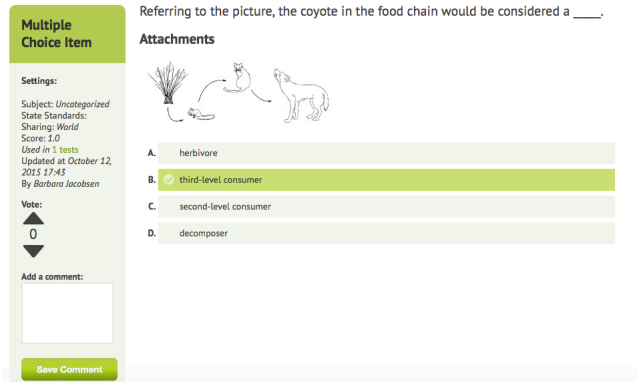

Figure 2. Peer Item Review and Comment in Naiku.

In addition to review by peer teachers, Naiku encourages students to provide feedback to their teachers. This is accomplished during the reflection stage, after students have completed the test. During this process, students can inform their teacher whether they thought the item was confusing, had more than one correct answer, or had no correct answer.

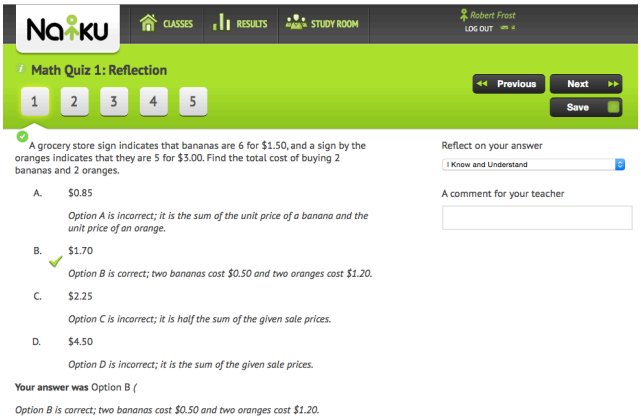

Figure 3. Student Feedback to the Teacher in Naiku.

After all the students complete the test, you can close and score the assessment. This will generate both a test result and item analysis report to help you review and evaluate the quality of the assessment and the items.

The test analysis report is shown below. Note the reliability estimate in the Assessment Statistics box. This statistic is the Coefficient Alpha, which provides a

good estimate of the reliability of the assessment. Also note the other descriptive statistics such as the minimum score, maximum score, mean score and standard deviation, which give a good picture of how your students performed on the assessment. At the bottom of the report, note the scores are presented by standard (or learning target). This helps teachers identify their students' strengths and areas in need of improvement and perhaps extra instruction.

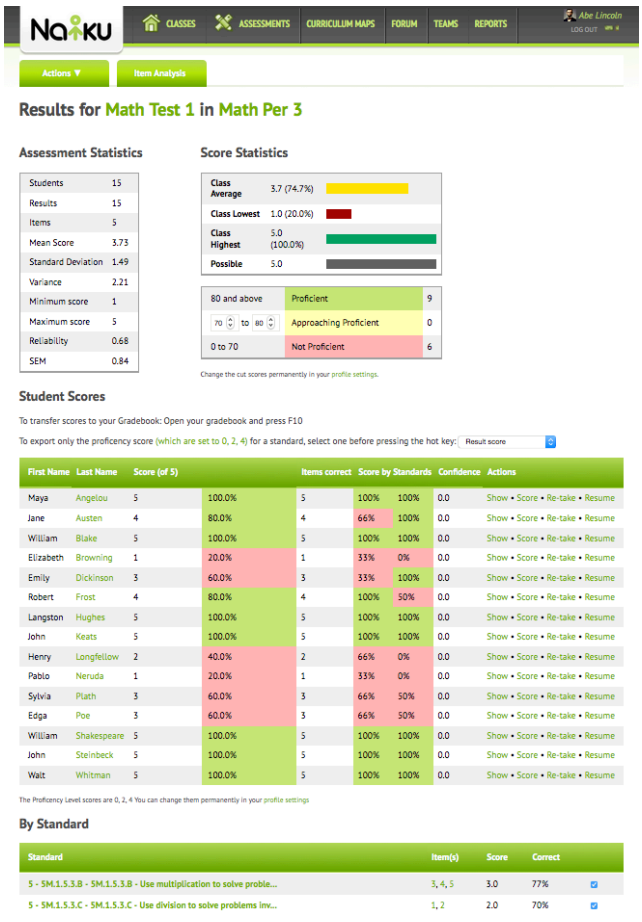

Figure 4. Test Analysis Report Showing Reliability Index in Naiku.

The item statistics and distractor analysis are provided in the Item Analysis report. For each item, the estimate of the difficulty of the item (p-value) and the item's discrimination power (pbis) are provided. These two statistics provide teachers with a lot of information about the quality of their items. In addition, the frequency and percent of students selecting each response option are provided. This information helps teachers revise and improve the items. In addition, this item is useful for special education teachers should they need to make modifications to the item (e.g., removal of the distractors).

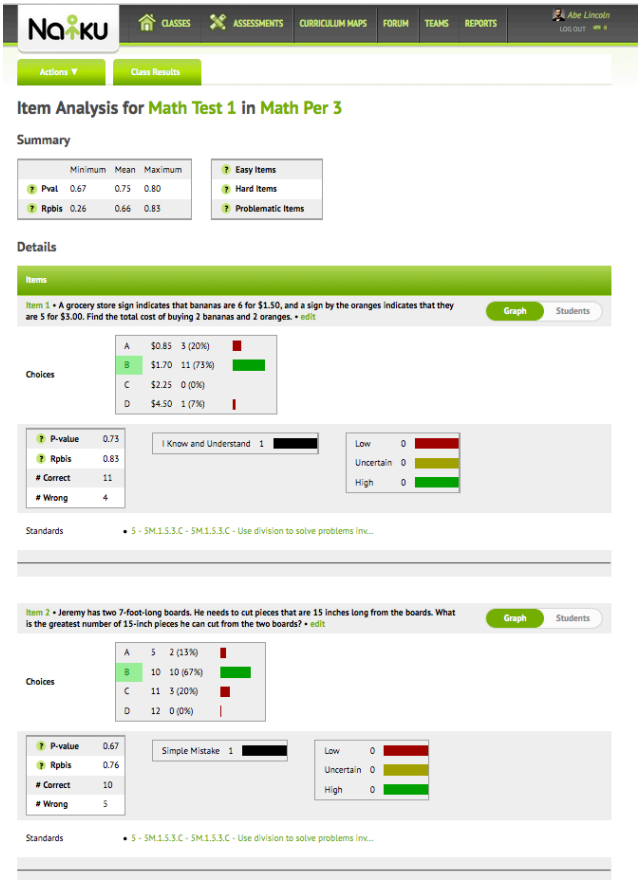

Figure 5. Item Analysis Report Showing Item Difficulty, Discrimination, and Distractor Analysis in Naiku.

At Naiku, we believe that teachers can create good and effective assessments and items. Teachers should familiarize themselves with the best practices and guidelines in assessment and item development. Judgmental approaches can be used to help teachers review and improve their items. Empirical approaches, utilizing item difficulty, item discrimination, and distractor analysis can also be used to evaluate the quality of the items. Although these statistics are best used with norm-referenced assessments, they can still be informative for a teacher's classroom criterion-referenced assessment. However, it is advised that you do not rely too heavily on the statistics, particularly when the number of students who took the assessment

is small and the number of items on the test is also small. Teachers can use a tool like Naiku to help them develop, review, evaluate, and revise their assessments and items.

References

Ainworth, L. (2015). Common Formative Assesssments 2.0. How teacher teams intentionally align standards, instruction, and assessment. Thousand Oaks, CA: Corwin:.

Anderson, L.W. (Ed.), Krathwohl, D.R. (Ed.), Airasian, P.W., Cruikshank, K.A., Mayer, R.E., Pintrich, P.R., Raths, J., & Wittrock, M.C. (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives (Complete edition). New York: Longman.

Chappuis, J., Stiggins, R., Chappuis, S. & Arter, J (2012). *Classroom Assessment for Student Learning.* 2nd ed. Upper Saddle River, NJ: Pearson Education.

Popham, W. J. (2011). *Classroom Assessment. What Teachers Need to Know.* 6th ed. Boston, MA: Pearson Education.

Webb, N. (March 28, 2002) "Depth-of-Knowledge Levels for four content areas," unpublished paper.

About the Author

Dr. Adisack Nhouyvanisvong is an expert in educational assessment, including computer-based and adaptive testing. He has created and ensured the psychometric integrity of large-scale educational assessments for states and national organizations, taught at the University of Minnesota, and is an adjunct professor at St. Mary's University where he has taught educators and graduate students in educational assessment practice and instructional strategies. He has been published in peer-reviewed journals, regularly speaks at education conferences and is currently President of Naiku.