

effectPerformance

Instructional design solutions for
your learning and performance needs



1513 Fairview Avenue
Havertown, PA 19083
Email: info@effectPerformance.com
Voice: 610.449.2060
Fax: 610.449.2061

 www.effectPerformance.com

White Paper

Are your e-learners learning?

A Rapid Prototyping Process and Tool for Test Development

Please cite as follows:

Pretera, G.E. (2004). Are your e-learners learning? A rapid prototyping process and tool for test development. *effectPerformance White Papers*. Retrieved from the effectPerformance, Inc. web site: <http://www.effectPerformance.com/html/library.htm>.

© 2004 effectPerformance, Inc. All Rights Reserved.

Are your e-learners learning?
A rapid prototyping process and tool for test development

Gustavo E. Prester, PhD, CPT
effectPerformance, Inc.

Abstract: How do we know if e-learners are learning unless we implement valid learning assessments? Developing high-quality tests and validating them can be time-consuming, however, and can require expertise that many training professionals lack. Speeding up the test development process and scaffolding the technical aspects of test validation may lead to higher levels of adoption. Applying a rapid prototyping methodology, we have developed a process and tool for accelerating test development.

Introduction

Are your e-learners learning? There is only way to find out... the dreaded test. It is through the assessment that we align learning with performance. If the assessment is derived from performance needs and it is an authentic reproduction of workplace tasks, then the assessment is said to align with performance. Assessments drive the design and development of learning experiences, or at least they should (Dick & Carey, 1990). When learning experiences are designed such that they adequately prepare learners to succeed during assessments, then the learning is said to align with the assessment. The key to performance-based instructional design is to promote alignment among performance, assessment, and learning (see Figure 1). Since the assessment is the critical linchpin between learning and on-the-job performance, it is important to use valid and reliable test instruments.

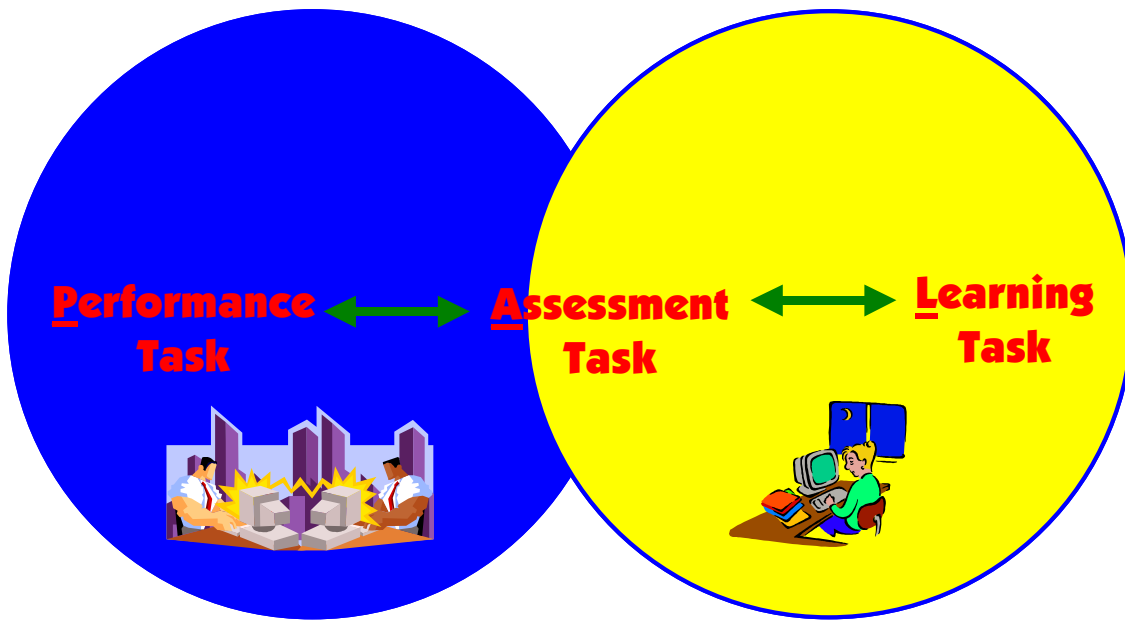


Figure 1. Remember your PALs. Aligning performance, assessment, and learning are critical to developing performance-based instruction

A Case of Aligning Your PALs

While working for a retail company some years back, I conducted a level 3 (transfer) evaluation and found that our trainees were not as technically adept as they needed to be, even after experiencing a week-long store operations training program. Before revising the training program itself, we implemented a level 2 assessment consisting of two components: a knowledge test and a performance test. In order to promote PAL alignment, we began by identifying performance tasks and skills. Next, we developed the tests and piloted them with our trainees. Based on the results, we revised and refined the tests. We repeated this pilot-revise cycle three times before we reached the reliability and validity levels we sought. At that point, we put the finishing touches on the test instruments and implemented the new assessment system.

Our baseline results indicated that while the tests were valid and reliable, learners performed quite poorly in certain areas. The tests helped us pinpoint what those areas

were and so we were able to focus our efforts on those trouble spots. Similarly, learners began doing something they had never done before... they began to tell us what we were doing wrong. They told us where they needed more practice, more examples, more discussion, and clearer instruction. We learned that when trainees perceive the test to be valid and authentic and the skills to be relevant to job performance, they take those test results very seriously. Within a matter of weeks, the trouble spots suggested by our assessment data and the feedback received from our trainees gave us a clear picture of what revisions needed to be made to the training. This is an example of assessments driving the design of learning experiences. It is also a great example of what can happen when we dedicate a short amount of time to developing valid, reliable, and PAL-aligned tests. This paper describes a streamlined approach and performance support tool that enable you to develop, evaluate, and revise tests quickly while ensuring PAL alignment.

A Rapid Prototyping Approach

Rapid prototyping is a methodology that involves taking an idea and turning into a working prototype, testing and revising that prototype until all stakeholders are satisfied, and then building the actual system. Thiagarajan (1999) proposes a rapid prototyping instructional design model in which designers speed up the design process, abbreviate traditional practices, make use of templates and other tools, make better use of stakeholders, and involve end users in the process at key points in the process. Through this type of process, not only does instructional systems design (ISD) consume less time, but time is used more effectively to gain insight into performance needs. In that spirit, I am developing processes and tools to support more streamlined ISD methods that

incorporate more stakeholder (especially, learner) involvement early in the process and greater alignment with work performance. In this paper, I describe an abbreviated process for developing Level 2 assessments (Kirkpatrick, 1998) and a performance support tool that makes validating your test a painless endeavor.

Test development does not take place in a vacuum; it is one small piece in a much broader ISD process. Figure 2 depicts the analysis and design portions of ISD. Note that some steps have dotted borders. These are what I consider optional steps in this rapid prototyping process, to be employed at the designer’s discretion. Note that writing performance objectives are treated as optional. In order to develop the test, the designer needs to identify the assessment criteria. Learning objectives are inherent within these criteria. The test items are developed based on the criteria, so they also inherently contain and convey learning objectives. Writing learning objectives, while it may serve other purposes (e.g., if included in the instruction, they can convey to learners a sense of what to expect), is not necessary for writing good tests, so long as assessment criteria have been properly identified.

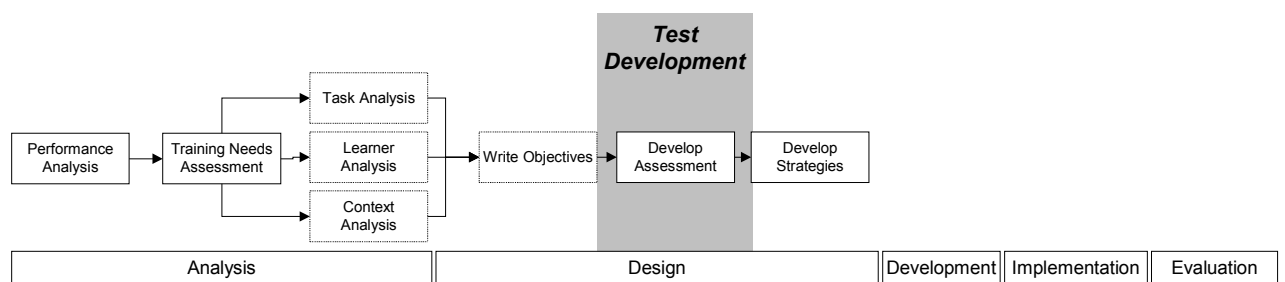


Figure 2. Workflow depiction of the analysis and design steps of the ISD process

Teaching to the Test

Note that in Figure 2, developing the *assessment instrument* comes before developing *strategies*. This is an important distinction, one which this model shares with the Dick and Carey (1990) ISD model. While some may criticize the practice of developing the test before designing the instruction as promoting a “teaching to the test” mindset, I challenge readers to consider that training is not education, is not development, and is not communication. The sole purpose of training is to develop immediately applicable job-related skills. This is the value proposition that training offers to organizations, and we must accept responsibility for this mandate. There is nothing wrong with teaching to the test in the realm of workplace training, so long as the test is a valid and authentic assessment of the skill. If we place more effort into developing job-relevant tests, then the training that is subsequently developed is more likely to share that characteristic. It has been my experience that a good test can drive instruction towards excellence, as in the case I described above. While this topic is hotly debated among educators, one could reasonably argue that the aim of schoolteachers goes beyond, or should go beyond, simply developing their students’ immediately-applicable job-relevant skills. The value proposition of education is much broader and deeper than that of workplace training, and therefore it is not adequate simply to teach to the test. We should, however, be cautious about burdening workplace training with a broader, less feasible mission than that of developing immediately-applicable job-relevant skills.

The Process

This rapid prototyping test development process (Figure 3) consists of four major steps: (1) identify assessment criteria, (2) develop test, (3) pilot test, and (4) revise test. For optimal use, the process should be repeated in order refine the test incrementally. While it is relatively quick and easy, especially with the use of our Item Analysis tool, this process does involve some effort both in terms of developing and revising the test itself and pulling together people to provide input and feedback.

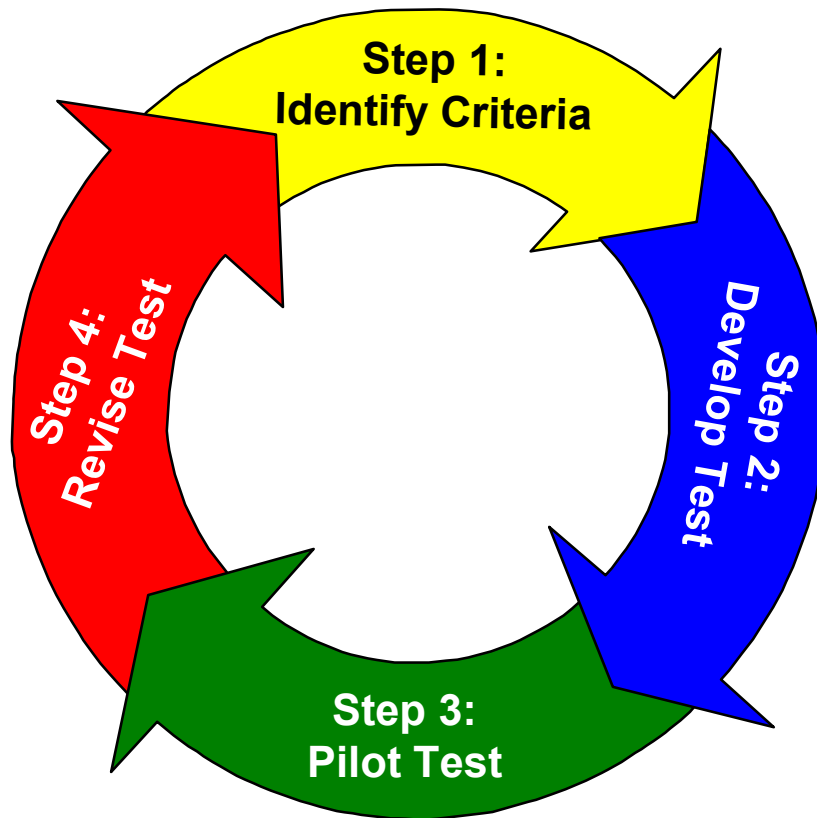


Figure 3. A 4-Step Rapid Prototyping Test Development Process

Step 1: Identify Criteria

The first step is to identify the assessment criteria that will guide your test development. An assessment criterion, much like a learning objective, describes what the worker should be able to do, the condition under which the task should be completed, and the degree of precision required. Fleshing out the assessment criteria for a set of skills and tasks should be a collaborative process. We recommend that you form a review panel consisting exclusively of exemplary workers who have already participated in the training. If the training is new, select exemplary workers or subject-matter experts (SMEs). This group should consist of at least 3 people but no more than 7.

Review the list of skills identified in the training needs assessment (see Prester, 2004) and ask the panel members to describe how they would assess each skill. Here is an example of how you might ask them to do this: “If I was a new worker and I told you that I already knew this job very well, how would you test me to find out if I was telling the truth? What would you ask me about? What would you have me do to prove it?” Take careful notes, as their comments will become the basis for your test items. Continually feed the assessment criteria back to the panel and make sure to ask all members to contribute. This is not the time to weed assessment criteria down, so encourage them to brainstorm without evaluating ideas... no criterion is unreasonable at this stage. Depending on the number of skills, fleshing out the test criteria could require multiple meetings. In the end, you should have at least three to four assessment criteria for each skill.

Step 2: Develop Test

Once the panel provides criteria for each skill, you can develop the test items. There are generally four types of learning assessments (see Figure 4). It is important; critical actually, that the test items reflect the nature of the assessment criteria, so selecting the test format is an important decision.

Test Format Matrix	Performance (Performance, simulation, projects, apprenticeships)	Knowledge (MC, TF, matching, fill-in, short answer, essay, report)
Objective (Correct/Incorrect)	Objective/ Performance	Objective/ Knowledge
Subjective (Rating scales)	Subjective/ Performance	Subjective/ Knowledge

Figure 4. Test Format Matrix depicting the four general types of learning assessments

Knowledge vs. Performance Tests. Some criteria will involve knowledge of terminology, concepts, rules, and techniques or the ability to analyze, trouble-shoot, or problem-solve, which can be tested through objective test item formats such as multiple choice, fill-in-the-blank, matching, sequencing, short answer, and essay. Other criteria may involve performing a task (a psychomotor skill), which may require performance testing of some sort. For example, a multiple-choice question would not be able to tell

you whether or not a cashier can process a transaction correctly, or whether or not a pharmacist can fill a prescription accurately. Conversely, a performance test in which you observe the cashier completing transactions would not necessarily reveal their conceptual understanding of how transactions impact the computerized inventory system. It is absolutely critical that the test items match as closely as possible the nature of the assessment criteria. Note that most performance tests require an in-depth task analysis, which is above and beyond the process outlined in this paper.

Objective vs. Subjective Tests. Some assessments can be measured objectively, while others cannot and so must be measured subjectively. In an objective test, a learner's responses are either right or wrong, depending on how well their responses match up with the established "correct" answer or the "expert" response. Objective test items yield right/wrong or yes/no (binary) kinds of data. Even performance tests can be measured this way, for example, if the judge uses a performance checklist. With subjective test items, the "correctness" of a learner's response varies along a scale, depending on the rating assigned by the judge. In the Olympics, for example, each judge rates the performance of the gymnast and then the ratings are averaged across judges. Even though the score is numeric, it is subjective. In reality, all assessments have elements of both subjectivity and objectivity. Even though objective true-false knowledge test items are either right or wrong, the "correct" answer is sometimes determined on the basis of how reality is subjectively interpreted by the SME or panel of SMEs. Do not get hung up on existential questions... simply remember that not all performance or knowledge tests can be measured objectively. Also, remember that e-learning is generally more conducive to objective assessments than subjective ones.

Writing the Test Items. We recommend that you work with an individual SME while writing test items, or at least have a SME review them before piloting the test. In addition, it is a good practice to have an outsider (someone not involved with the test development and not a member of the target audience) read the test items and directions out-loud to you and explain what they mean (Nitko, 1996). This sanity check ensures that your test directions and test items are written in a way that makes sense to readers. Lastly, have someone proofread your test for grammar and spelling. For additional guidelines on writing specific types of test items (e.g., multiple-choice), please visit the *Toolkit* section of our web site (www.effectperformance.com).

Step 3: Pilot Test

It is easy to write bad test items and extremely difficult to write good test items. Writing good test items is both a science and an art. Unfortunately, designers usually do not know if their test items are good or bad, because their tests do not undergo any form of evaluation. A critical component of rapid prototyping is piloting the product with trainees before fully implementing it. If the test is being developed before the training itself or is being developed concurrently, you will not be able to pilot the test with trainees. An alternative is to pilot the test with two approximately equal-sized groups: a group of typical workers who have not attended training and a group of exemplars or SMEs who already possess the desired skills. The latter could include the members of the panel. If the test has been written reasonably well, you should find that the exemplars and SMEs perform well on the test, while the typical workers who have not yet attended

training perform very poorly. If this is not the case, then right away you know that something major is wrong with the test.

Item analysis refers to a technique for evaluating test items. Specifically, an item analysis answers the question: do the test items distinguish between people who possess the skills and those who do not? Discriminant validity is a measure of a test's ability to distinguish between someone who actually knows the answers and someone who does not know... but guesses correctly on occasion. For example, does your watch know the right time? If your watch were usually off by a several minutes, then your confidence in its ability to tell time would wane. If a test item is measuring well, then odds are that the top performers on the test (those who are skilled) will answer the question correctly, while the low performers on the test (those who are unskilled) will answer the question incorrectly. In other words, the test item can discriminate, or distinguish, between skilled and unskilled test takers. The more accurate the test item, the greater your confidence in its measuring ability. We can capture this relationship mathematically with the variable *Item Discrimination (d)*.

A watch may be very accurate at times, but if it is not consistently accurate, then you would probably want to replace the watch. Similarly, while individual test items may be measuring properly, it is important that the entire set of test items consistently measure properly. This internal consistency of the test is what we call *reliability*. If the test is not reliable, then it is a worthless measurement that tells you little about the learner's actual skill. Reliability can be measured and described mathematically through several formulas, which will be discussed shortly.

The Item Analysis Tool (available online at www.effectperformance.com) was developed to streamline and automate the item analysis process. Simply pull the data from your learning content management system (LCMS) or whatever database is being used to collect the test data and copy it into the Item Analysis tool. The tool automatically calculates the *Item Discrimination* (d) and *Item Difficulty* (p) values for each test item as well as the overall test reliability estimates. All that is left for you to do is to interpret the results and revise the test.

Step 4 Revise Test

Interpreting reliability. In order to revise the test, you need some indicators that give you a sense for how well your test is measuring the skill level of the test takers. The first of these indicators is test reliability. There are three commonly used measures of a test's reliability: KR-20, KR-21, and Cronbach's Alpha. The KR-20 and KR-21 (Kuder & Richardson, 1937) were developed as estimates of the internal reliability of a test. The KR-20 is more commonly used, primarily because it is less sensitive to mean values than KR-21 and so is more stable. Cronbach's Alpha (Cronbach, 1951) is the most conservative estimate of reliability (Nitko, 1996). It is less often used than the KR-20, but it is the most flexible in that it can be used to estimate the internal reliability of tests that involve essays and judge's ratings (where scores can vary along a scale), while KR-20 and KR-21 can only be used to estimate the reliability of dichotomous test items (where the answer is either right or wrong).

Figure 5 shows an example of what the reliability estimates will look like in the Item Analysis Tool. The sample data here consisted of 50 fill-in-the-blank test items

given to 159 trainees. Note that the three estimates are very close to one another. The values will always range from 0 to 1 and the higher the value the more reliable the test. You are generally shooting for a reliability estimate that is above .80. If your test achieves that level of reliability, you will likely only need to make minor changes to your test items. If the values fall far below .80, you may have more substantial revision work to do.

Item Discrimination. What items do you need to revise? Sullivan, Wircenski, and Major (1999) describe a straightforward process for analyzing knowledge-based assessments, in which they emphasize the importance of calculating *Item Discrimination* (d) and using that indicator to identify bad test items. There are two ways to estimate d : (1) calculate the correlation between the scores for the individual test item and the overall test scores (those who do well on the test should do well on this item and vice versa) or (2) sort the test scores in descending order, then for each test item, calculate the average score of the top third and subtract it by the average score of the bottom third (the top scorers on the test are more likely to get this item right than the bottom ones). In the Item Analysis Tool, I use the first method. The value for d_1 will range from -1 to 1 . The closer to 1 , the better the test item is functioning. If an item has a d_1 value of $.8$, there is a very strong correlation between doing well on the test overall and answering this test item correctly. Conversely, a d_1 value of $-.5$ indicates that there is a negative relationship between doing well on the test and answering this item correctly. In other words, people who do not really know the material have a better chance of getting this question right than those who do.

Item Difficulty. The Item Difficulty (p) is simply an indication of how difficult the test item was for the test takers... it is also an estimate of the odds of getting that test item correct for future test takers. The value for p ranges from 0 to 1. The closer to 1 that p is, the easier the test item is, while the closer to 0 that p is, the more difficult. The value for p is calculated by summing the number of correct responses for a given test item and dividing by the total number of test takers. So, for example, a p value of .2 tells you that only 20 test takers for every 100 answered that test item correctly.

Interpreting p and d_1 . Figure 5 shows the values of d_1 and p for each test item. The table shown in the Item Analysis Tool is color coded, so that green values are okay but red values require your attention. To help you visualize the relationship between p and d_1 , the Item Analysis Tool generates a scatter plot of these values (see Figure 6). There are four sections to the plot. The low discrimination area includes any test item that falls below a d_1 value of .1. The ideal range includes test items with p values that range from .55 to .85 and d_1 values of .1 or higher. These parameters are arbitrary but are nonetheless sound guidelines to follow in interpreting the results. Next, we will work through some examples.

Summary Information for....		Fill-in Test (Data1)	
Number of items (k) in the test	50		
Sample size (n)	159		
Average Score (M)	14.40	29%	
Cutoff Score	5.00	10%	
Median (midpoint)	14.00	28%	
Lowest Score	-	0%	
Highest Score	43.00	86%	

Is my test instrument reliable?	
KR20	0.86 YES, RELIABLE
KR21	0.83 YES, RELIABLE
Alpha-1	0.86 YES, RELIABLE

Which items do I need to revise?

Item Number (#)	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Item Difficulty (p)	0.6	0.6	0.7	0.3	0.2	0.1	0.1	0.2	0.3	0.1
Item Discrimination (d1)	0.3	0.4	0.4	0.3	0.5	0.3	0.4	0.5	0.3	0.4
	#11	#12	#13	#14	#15	#16	#17	#18	#19	#20
	0.5	0.6	0.2	0.7	0.6	0.7	0.2	0.3	0.3	0.2
	0.5	0.5	0.3	0.4	0.4	0.4	0.4	0.3	0.3	0.3
	#21	#22	#23	#24	#25	#26	#27	#28	#29	#30
	0.2	0.1	0.3	0.3	0.1	0.3	0.1	0.2	0.1	0.4
	0.4	0.3	0.3	0.4	0.4	0.5	0.4	0.4	0.3	0.4
	#31	#32	#33	#34	#35	#36	#37	#38	#39	#40
	0.6	0.5	0.2	0.0	0.2	0.5	0.1	0.2	0.5	0.2
	0.4	0.3	0.3	0.3	0.4	0.4	0.4	0.4	0.4	0.2
	#41	#42	#43	#44	#45	#46	#47	#48	#49	#50
	0.2	0.2	0.1	0.2	0.1	0.2	0.2	0.1	0.2	0.1
	0.3	0.4	0.3	0.4	0.4	0.3	0.4	0.3	0.4	0.4

Figure 5. Excerpt of Item Discrimination (d1) and Item Difficulty (p) values from the Item Analysis Tool

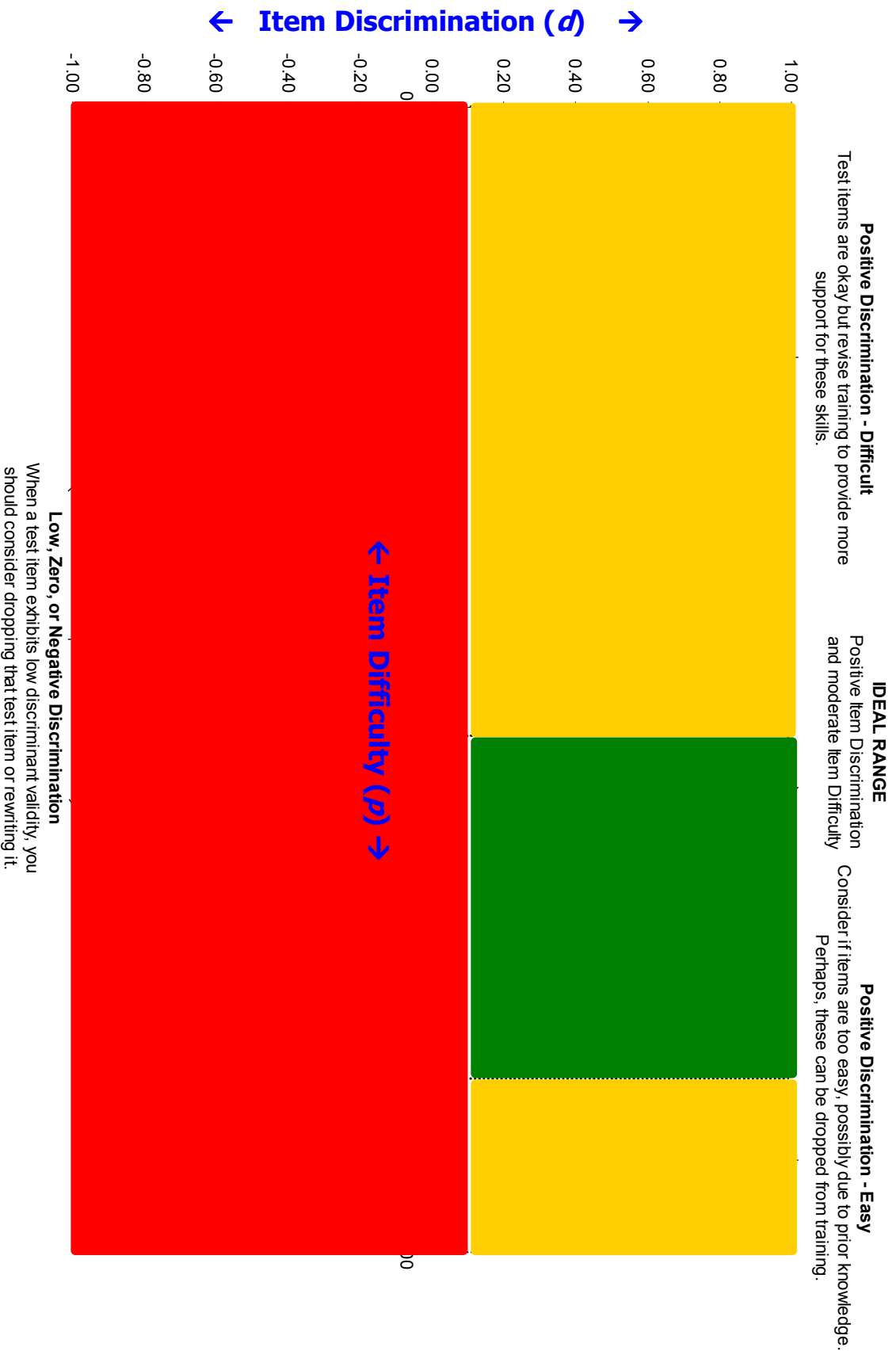


Figure 6. Plot of p and d , excerpted from the Item Analysis Tool

While sometimes test items can be too difficult or too easy, that does not make them bad test items per se. For example, item #5 in Figure 5 has a p of .2 (2 in 10 people got this right) but a d_1 of .5. In this case, the item is valid but the question is very difficult. What do we do? Since the item is valid, the instruction must be insufficient. In this case, it is better to direct more instructional resources (i.e., presentation, coaching, practice, discussion, assessment, and feedback) to the skill and leave the test item as is. As a general rule, if the test item is valid but difficult, revise the training.

What if instead p was .95 and d_1 was .5. The test item is valid but it is too easy. What do you do? Do you re-write the test item to make it more difficult or confusing? Again, the problem is not the test item but the training. This skill may be something that learners already possess through prior knowledge and ability. The correct interpretation may be to remove that skill from the training and then test again to see if learners continue to do well. This suggests that learners have high prior knowledge for this item. Another possible interpretation is that the training is very strong in this area, so the test item is easy as a result of effective presentations, practice, feedback, etc. If this is the case, you may wish to keep the test item but perhaps redirect instructional resources from this skill to one that needs it more, or simply leave things alone. As a general rule then, when the test item is valid but extremely easy, consider the possibility that prior knowledge is high.

What if d_1 is low, zero, or negative? Low discrimination values (below .1), zero values, and negative values are indicators of poor test item performance. With negative discriminators, simply delete or rewrite the questions. With low and zero values, you may be able to salvage the test item by revising the wording. You will need to review

each of these items individually to find out what is misleading or confusing about them. You should go back and ask some of the test takers to walk through the test with you and describe how they interpreted each item. This is a great way to reveal problems and involve workers/trainees in your design and development efforts.

Once you have revised the test, you can implement it but we recommend that you continue to collect and analyze data to ensure that your changes made a difference. Rapid prototyping works best through a series of user input-develop-test-refine cycles. If the revisions are major and the reliability was low to begin with, you should definitely run a second pilot. However, even if only minor changes were made, it is best to repeat the identify criteria-develop-pilot-revise cycle until you can be confident that the test is reliable and valid.

Summary

In this paper, we have described a short and simple process for developing and evaluating a test. The process involves (1) gathering input from stakeholders regarding assessment criteria, (2) converting the assessment criteria into test items to produce a test, (3) piloting the test and running an item analysis with the pilot data, and (4) interpreting the results and revising the test items to make the test more reliable and the questions more effective. A key success factor in any rapid prototyping process is getting stakeholder feedback early and often. By piloting the test and conducting a thorough item analysis, it relieves some of the pressure to get the test right the first time. Instead, designers can move quickly from assessment criteria to test development knowing that mistakes will be caught and fixed through piloting and analysis. Since learning

assessments drive the selection of instructional content and instructional design strategies, reducing the development time, improving the relevance, and raising the quality of learning assessments should lead to more effective training.

References

- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.
- Dick, W., & Carey, L. (1990). *The Systematic Design of Instruction*. Glenview, IL: Scott, Foresman.
- Kirkpatrick, D. (1998). *Evaluative training programs: The four levels* (2nd ed.). New York, NY: Barrett-Kohler.
- Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*, 151-160.
- Nitko, A. J. (1996). *Educational Assessment of Students (2nd Ed)*. Englewood Cliffs, NJ: Prentice-Hall.
- Prester, G. E. (2004). Training needs assessment: Process and tools to help you identify and prioritize training needs quickly. *effectPerformance White Papers*.
- Sullivan, R. L., Wircenski, J. L., & Major, M. J. (1999). Analyzing knowledge-based tests. In D. L. Kirkpatrick (Ed.), *Another Look at Evaluating Training Programs* (pp. 113-118). Alexandria, VA: ASTD.
- Thiagarajan, S. (1999). *Rapid Instructional Design*. Workshops by Thiagi, Inc. Retrieved 11/18/2003, from the World Wide Web: <http://www.thiagi.com/article-rid.html>