Implementation of Standards Setting for a Algebra II Benchmark Exam Using Cognitive Diagnosis

Robert Henson

The University of North Carolina at Greensboro

ERM Mini-Conference, Spring 2008

Outline

- Discuss diagnostic models and the need for standard setting.
- Talk about standard setting.
- Talk about the general procedure.
- Briefly discuss the results.

Diagnostic Models

- As was previously discussed diagnostic models can provide a mastery profile, $\boldsymbol{\alpha}$.
- The probability of a correct response is determine based on the set of attributes that a person has or has not mastered and the Q-matrix.
- The models vary in their complexity and how they define the probability of a correct response.
  - I will briefly discuss the DINA, Reduced RUM, and the LCDM

The Q-Matrix

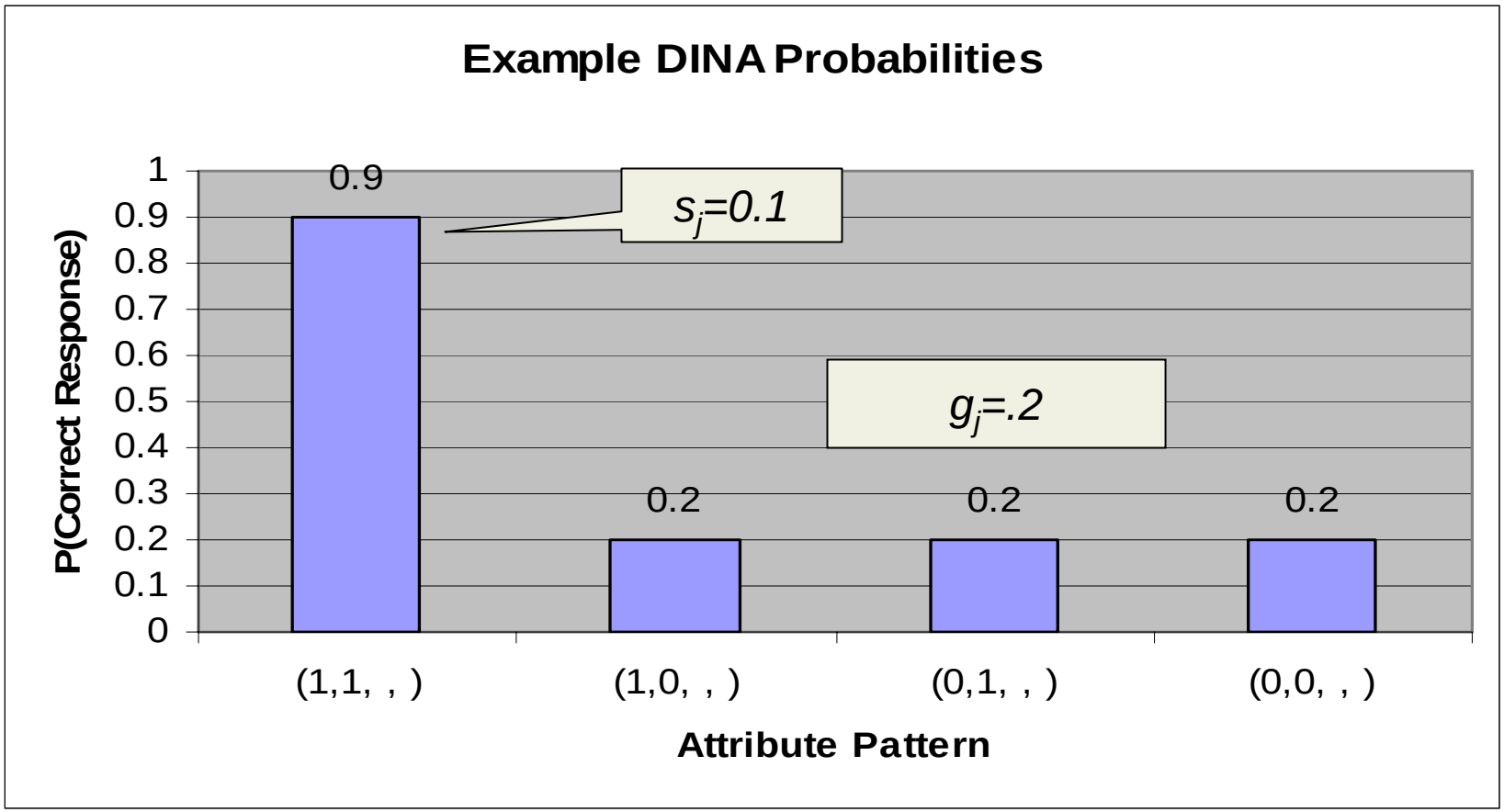
- To aid in our discussion, we assume that we take an item from a test intended to measure basic math.

	Add	Sub	Mult	Div
2+3-1	1	1	0	0

The DINA Model

- One of the simpler models
- Divides people into two groups (ξ_{ij})
  - Have mastered all required attributes
  - Have not mastered all required attributes
- Defines probability for those who have mastered all required attributes ($1\text{-}s_j$).
- Defines probability for these who have not mastered all required attributes (g_j).

The DINA Model



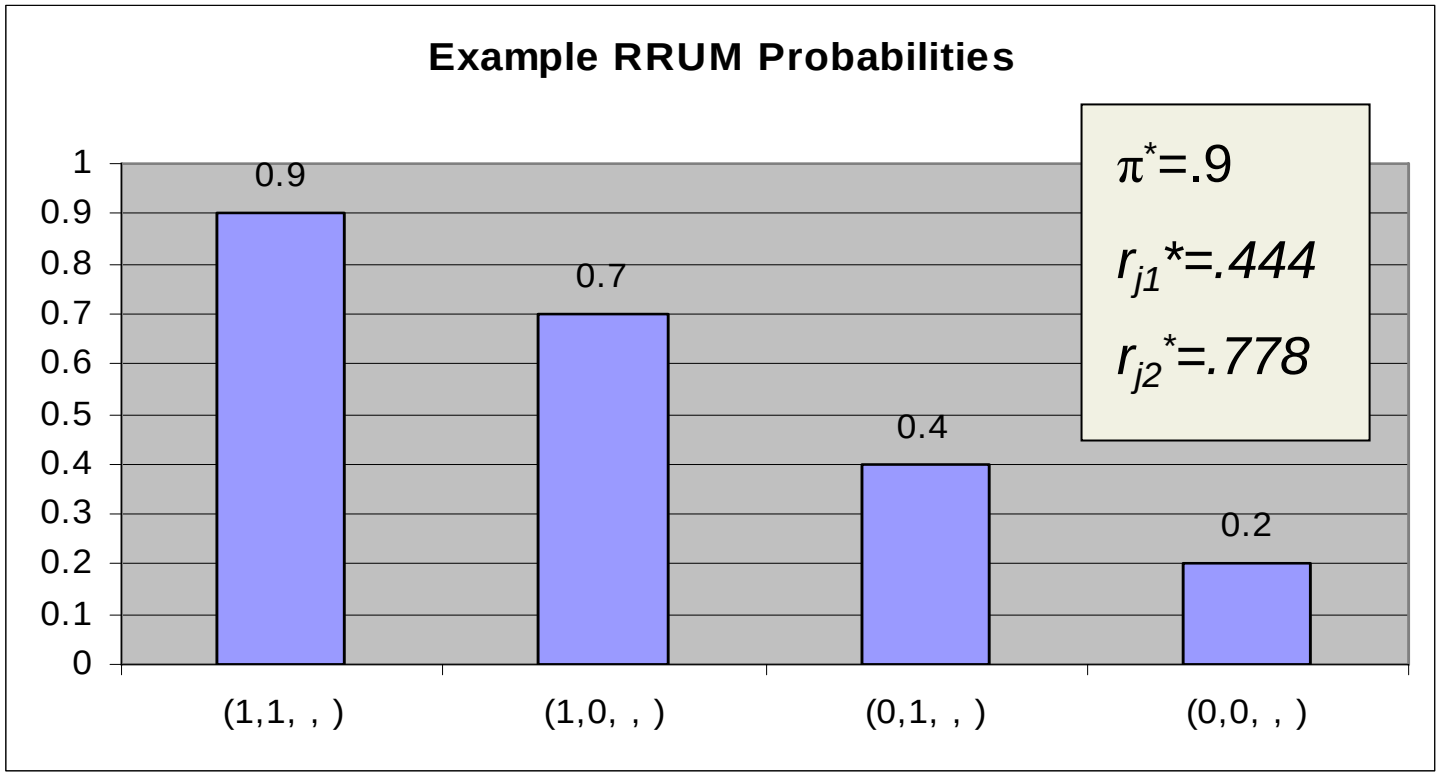
Reduced Reparameterized Unified Model

- As an alternative, the Reduced Reparameterized Unified Model (R-RUM) allows each attribute to contribute differently to the probability of a correct response

- Uses π^* to indicate the probability of a correct response if all attributes are mastered.

- The value r_{jk}^* is used define the penalty of not mastering the k^{th} attribute

Reduced Reparameterized Unified Model



This implies that knowing addition is more important than knowing subtraction for the problem 2+3-1=?

The LCDM

- In this particular case we use the Log-linear Cognitive Diagnosis Model (LCDM, Henson, Templin, and Willse, 2007).

- The LCDM is a special case of a log-linear model with latent classes (Hagenaars, 1993) and thus is also a special case of the General Diagnostic Model (GDM, von Davier, 2005).

- The LCDM defines the Logit of the probability of a correct response as a linear function of the attributes that have been mastered.

The LCDM

- The Logit of a probability is used because the items are dichotomous (correct or incorrect).
  - A linear model is not reasonable for probabilities because of the bounds, so we transform the probability.

$$Logit = \ln\left(\frac{P(X_{ij}=1)}{1-P(X_{ij}=1)}\right)$$

- The Logit ranges from a negative infinity to a positive infinity when using probabilities and so can be modeled as a linear function.

The LCDM

- Given our item, we model the logit of the probability of a correct response as a function of mastery or nonmastery of the two attributes.

- Specifically,

$$\ln\left(\frac{P(X_{ij}=1)}{1-P(X_{ij}=1)}\right)=\lambda_{add}\alpha_{add}+\lambda_{sub}\alpha_{sub}+\lambda_{add*sub}\alpha_{add}\alpha_{sub}-\eta$$

Diagnostic Models

- The LCDM is a general model that allows all possible combinations of the required attributes to have a different probability of a correct response.
- However, the solution may or may not be consistent with other criteria.
- Specifically, examinees are classified based on their responses in a way that best divides examinees.
- But, this classification may not be optimal based on an alternative set of standards, such as the performance on End of Course exams

Standard Setting

- Here we discuss a pilot study designed to set the standards for an Algebra II benchmark exam.
- The purpose is to provide a general procedure for standard setting when using cognitive diagnosis models based on a modified Angoff procedure.
- Also, to provide a general description of possible concerns and limitations that must be considered.

Standard Setting

- Instead, we should set the probability of a correct response for each possible set of required attributes.
- In getting these probabilities the standard is set for all possible combinations of mastery.
- Thus we define how a student will be classified in to each possible category.

Procedure

- Given that we are interested in defining standards for an Algebra II Benchmark Exam, we discuss the general procedure.
  - Provide a description of the method used to request the information needed.
  - Give an example of the form.
  - Discuss the follow-up process of validating the standards.

Standard Setting Procedure

- A total of four Algebra II teachers are recruited.
  - Varied in type of students taught (e.g. advanced students versus students performing at an expected level) and experience.
- Given the test previously explained and a Q-matrix that was previously determined for each item, teachers are asked a set of questions based on the probabilities to be set by the LCDM.
- Questions were asked in a way that did not require knowledge of diagnostic models.

Standard Setting Procedure

- An example item in the benchmark test was:

2. If one factor of $f(x) = 12x^2 - 14x - 6$ is $(2x - 3)$ what is the other factor of f(x) if the polynomial is factored completely.
   a. $(6x - 2)$
   b. $(6x + 2)$
   c. $(6x + 3)$
   d. $(6x - 3)$

- To correctly answer this item students must be able to:
  - Operate with algebraic expressions (polynomial, rational, complex fractions) to solve problems.
  - Use quadratic functions and inequalities to model and solve problems; justify results.

Standard Setting Procedure

- Teachers are asked to:

> …think of a student of yours, or imagine a student, that has mastered a certain set of skills. By "mastered" we mean a student who would not need to spend any additional time learning this material. We will then ask you, "If this person were to take a test with 100 items similar to a particular item, how many items would they get right?"

- Notice that by doing this we are specifically defining *mastery*.
- In addition, we have reframed the question based on a 100 item test.
  - Framing the task in this way, makes it easier to think about.

Example Response for Item 1

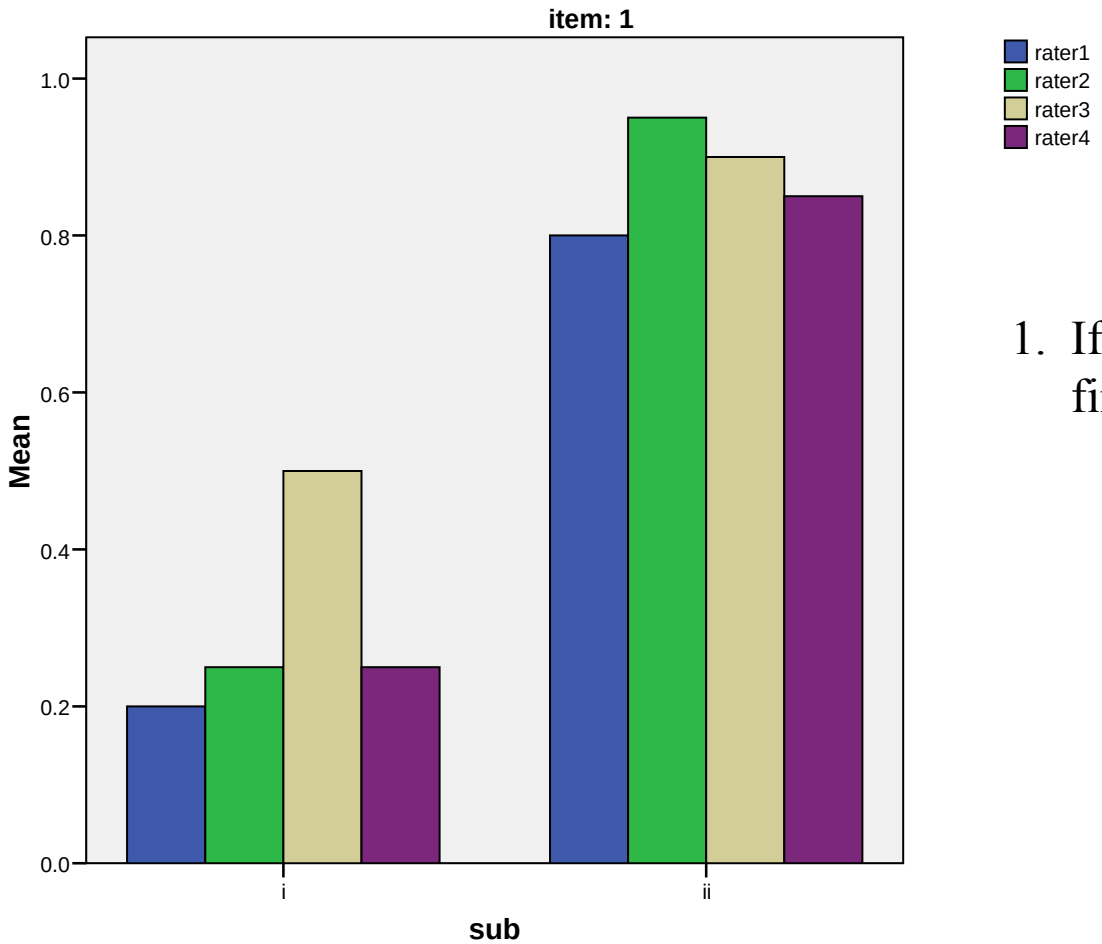
1. If $f(x) = x^2 + 2$ and $g(x) = x - 3$ find $f \circ g(x)$.
   a. $x^2 - 6x + 11$
   b. $x^2 + 11$
   c. $x^2 + x - 1$
   d. $x^3 - 3x^2 + 2x - 6$

Imagine a test that contains a total of 100 items just like this test (assume they are not identical)

1i. __25__ Think of one of your students that has NOT MASTERED Objective 2.01. How many of these items on the 100 item test should they get right?

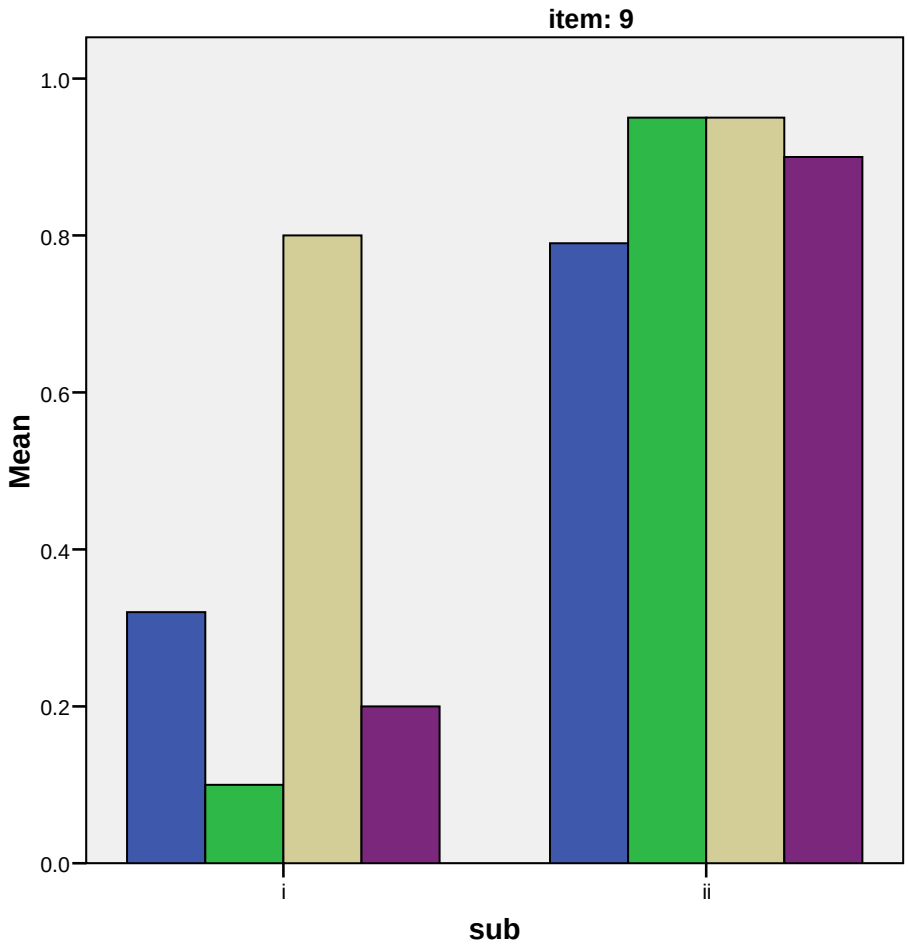
1ii. __85__ Think of one of your students that has MASTERED Objective 2.01. How many of these items on the 100 item test should they get right?

Example Responses (Item 1)



1. If $f(x) = x^2 + 2$ and $g(x) = x - 3$ find .
   a. $x^2 - 6x + 11$
   b. $x^2 + 11$
   c. $x^2 + x - 1$
   d. $x^3 - 3x^2 + 2x - 6$

Example Responses (Item 9)



9. Solve $3(ax + b) = 2(cx - d)$ for x.

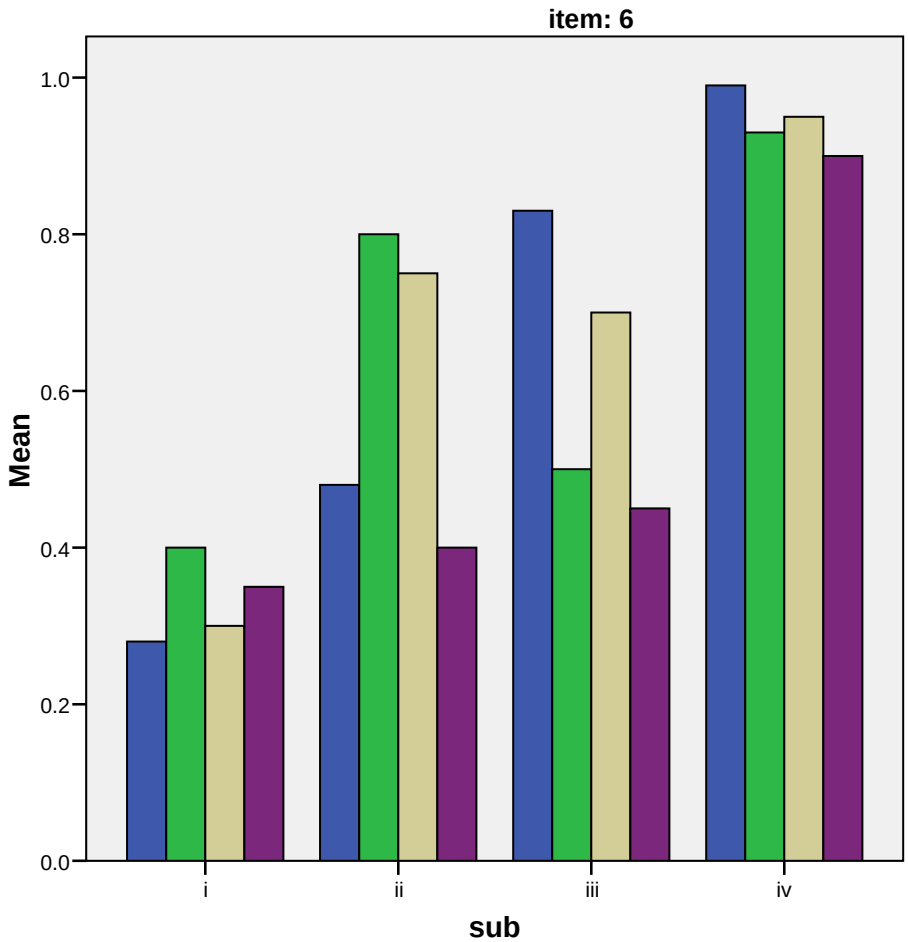
a. $x = \frac{b-d}{a-c}$

b. $x = \frac{bd}{ac}$

c. $x = \frac{-3b-2d}{3a-2c}$

d. $x = \frac{2b-2d-3ax}{2c}$

Example Responses (Item 6)



6. Determine which of the following graphs does not represent Y as a linear function of X.

Analyses

- Based on the teachers responses, the average probability of a correct response is calculated.

- These averages are used to compute model parameters.
  - Specifically, if we know the probabilities (based on the teachers' responses) then we can compute the logit.
  - Also, the questions where phrased in a way that specified the attribute pattern $$\ln\left(\frac{P(X_{ij}=1)}{1-P(X_{ij}=1)}\right)=\lambda_{add}\alpha_{add}+\lambda_{sub}\alpha_{sub}+\lambda_{add*sub}\alpha_{add}\alpha_{sub}-\eta$$
  - Therefore we can directly compute the parameters of the model

Analyses

- Using these fixed parameters as truth, MCMC estimation is used to obtain estimates of the posterior probability that each skill has been mastered.
- An example could be:

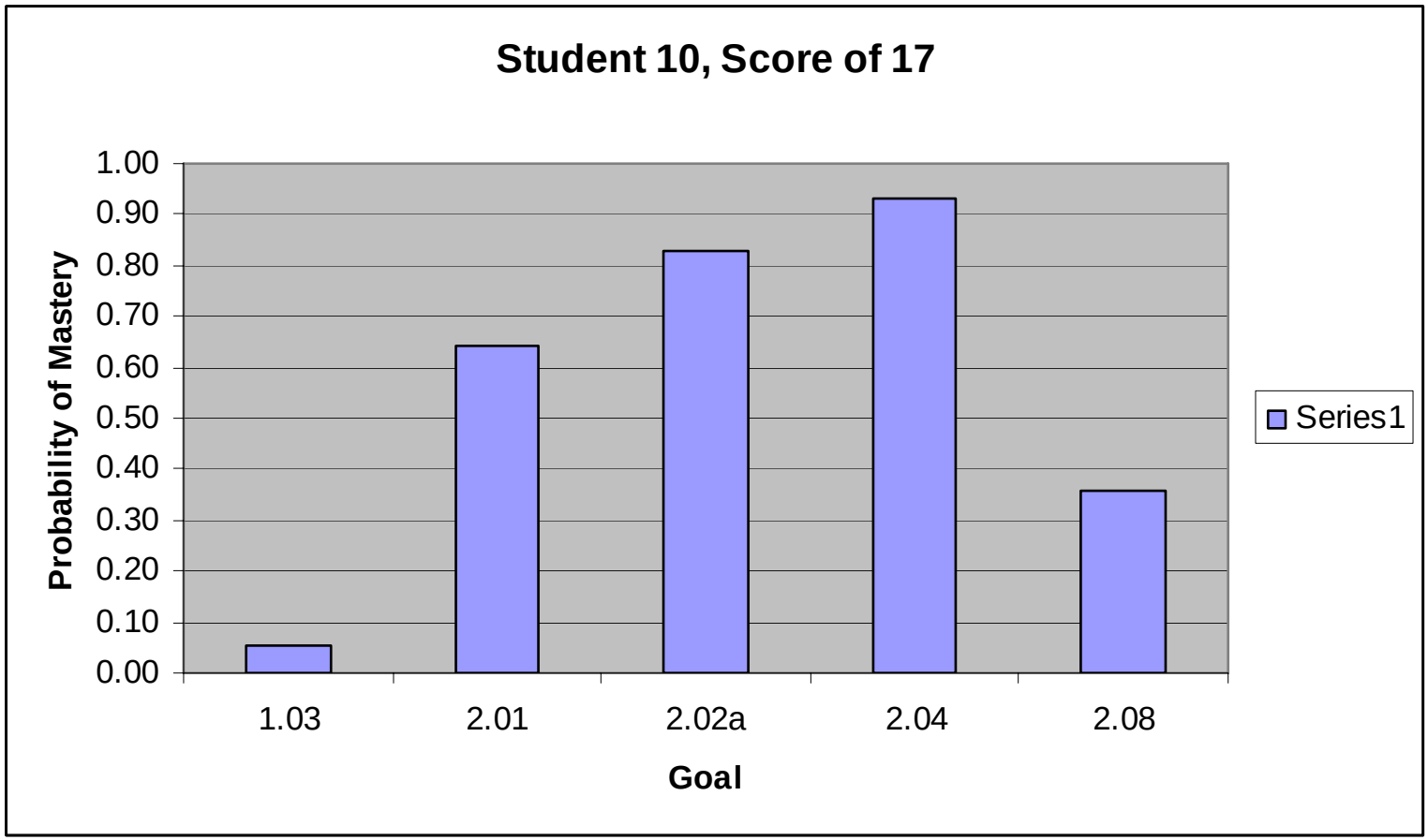
Student ID	1.03	2.01	2.02	2.04	2.08
24	0.25	0.87	0.99	0.44	0.05

- These can be converted to student mastery or nonmastery based on the mostly likely value.
  - Greater than 0.50 equals a master.
  - Less than 0.50 equals a nonmaster.

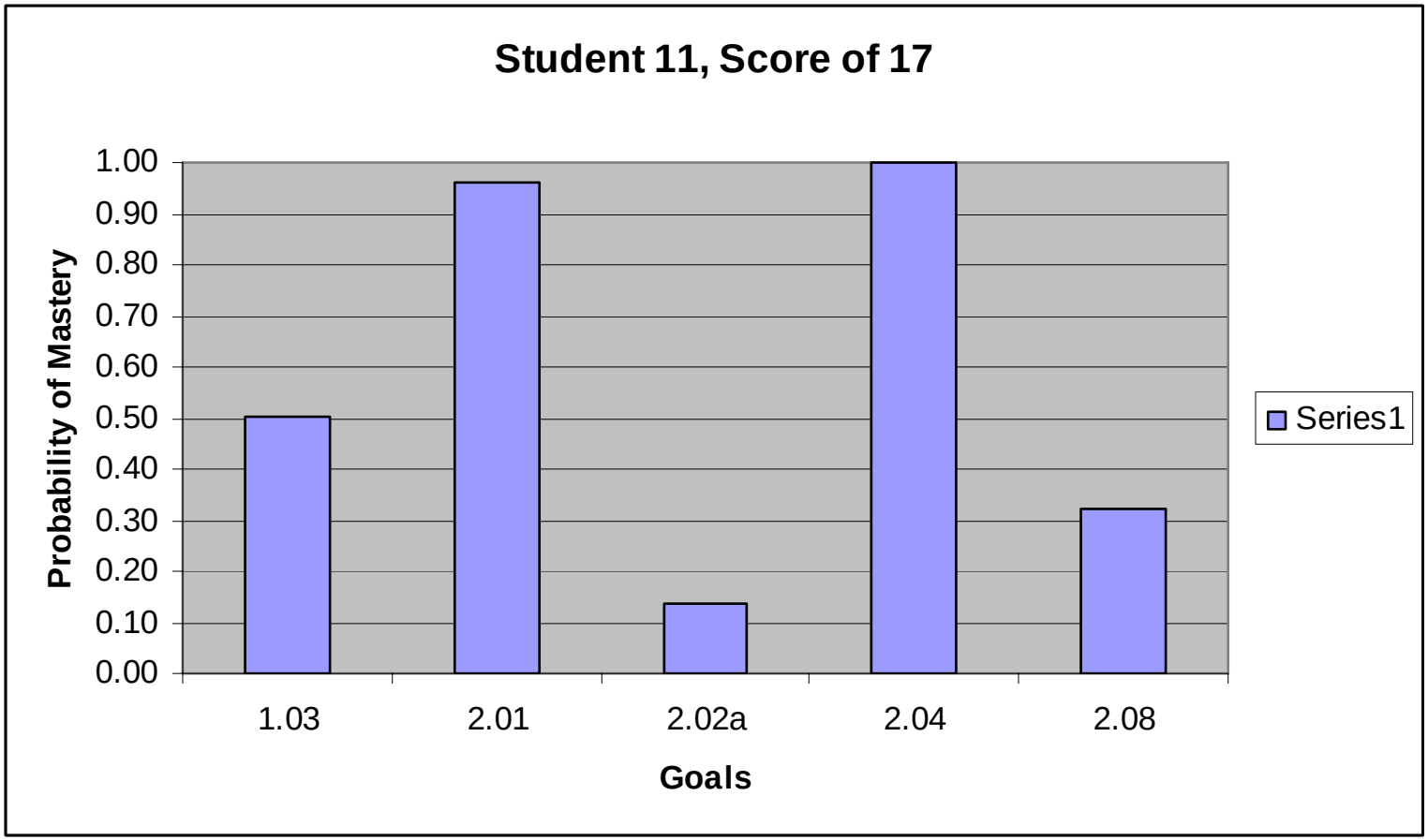
Analyses

- A final meeting is held with the teachers for validation.
  - Of particular interest is a teacher with 31 students who completed the benchmark exam.
  - She was able to directly comment on whether the classifications were reasonable.

Example Feedback



Example Feedback



Additional Feedback

- In addition to talking specifically about each student more general statements related to the class were discussed.
  - For example, those skills that most student seemed to know and those skills that needed to be improved.
- As a preliminary validation, those skills that we identified as needing additional work were confirmed by the teacher.
- For example, of 31 students, we estimated only 2 students had mastered goal 2.08 (Use equations and inequalities with absolute value to model and solve problems: justify results.).
  - This skills had not been covered by the time the benchmark was administered.

Additional Feedback

- Finally, expected benchmark performance for specific students was discussed.

- For example:

  - A student who had not mastered any skills was expected to get between 8 and 9 of these items right.

  - A student who had mastered all five attributes is expected to get between 22 and 23 item correct.

Conclusions

- In general, the procedure used here seems promising.
  - Teachers comments of the format of the questions were encouraging.
  - Responses where generally consistent across teachers.
  - Further refinement is planned.
- However, in our follow up interviews a few issues specific to this benchmark were mentioned.
  - Use of calculators versus actual mastery.
  - Expectation of teachers based on the level of the students.

Acknowledgements

- Jon Templin
- Terry Ackerman
- Deb Bartz
- John Willse