RASCH MODEL ANALYSIS OF PLACEMENT EXAMS ON A SINGLE DIFFICULTY SCALE

W. BRYC

Contents

1. Rasch Model of measurement

1.1. **Introduction.** Rasch model is a statistical model of a test that attempts to describe the probability that a student answers a question correctly. It assigns to every student a real number, a, called the "ability", and to every questions a real number, d, called the "difficulty". Both quantities are measured in the same units, called "logits", and the origin is set arbitrarily. Charts below are rescaled so that one logit corresponds to 100 and the origin was set to 500, i.e. to 5 logits.

Date: 1/26/2001 8:40:34 PM
1/15/2001 12:48:11 AM
Minor revisions of draft of 12/14/2000.

The model assumes that a single student responds correctly to a single question with the probability given by

$$p = \frac{e^a}{e^a + e^d} \tag{1}$$

Thus a student has a 50% chance of correct response to a question of difficulty that matches her ability. Her chances of responding to harder questions decrease rapidly towards 0. Her chances of responding correctly to easier questions increase rapidly to 1. Formula 1 expresses the probability of correct answer through the difference between students ability and question difficulty. Figure **??** on page **??** shows the average score of students of varying abilities on an exam that consists of 20 questions of fixed difficulty 500.

The model assumes that scores of different students on different questions are (stochastically) independent. Thus each test consisting of m questions and taken by n students is described by $n+m$ parameters: abilities $a_1, a_2, \ldots, a_n$, and difficulties $d_1, d_2, \ldots, d_m$. Student abilities and question difficulties are then determined by requiring that the probabilities of the observed scores are maximized[1]

1.2. **Properties of the model.** One can verify the following properties of Rasch model:

(i) Student abilities and difficulties of questions are determined up to an additive constant only. This follows from the form of the probability of correct answer (1), which is a function of the difference $a-d$ of ability and difficulty levels only.

This property is useful for standardizing different tests to construct a larger scale that covers a wider range of abilities.

(ii) For students who answers all questions (no missing data), the ability is a monotone functions of the total score on the test[2].

As a consequence, the cutoff to place a student into a course needs to specify the total score on a test only. (Of course, different versions of the exam may require different cutoffs.)

(iii) The difficulties of the questions do not depend on the abilities of student testers used to determine these difficulties[3] So the difficulties of the questions can be determined without knowing the ability a of a "tester", and testers of unknown ability can be substituted for the target population.

---

[1]In fact, it is known [3] that the maximum likelihood estimators are inconsistent. Instead, a conditional maximal likelihood is used, see [1] (in the examples I used a DOS program BIGSTEPS available at `http://www.winsteps.com`)

[2]This isn't intuitive and requires proof. One would expect that it should matter whether the student answered harder, or easier questions, but it turns out that its the total score only that matters.

[3]This can be seen from computing probabilities of correct answers of a single student of unknown ability a on two questions of difficulties d_1, d_2. If she got exactly one of the two questions right, then the (conditional) probability that that she got the first question right comes out as

$$\frac{\frac{e^a}{e^a+e^{d_1}}\frac{e^{d_2}}{e^a+e^{d_2}}}{\frac{e^a}{e^a+e^{d_1}}\frac{e^{d_2}}{e^a+e^{d_2}} + \frac{e^a}{e^a+e^{d_1}}\frac{e^{d_2}}{e^a+e^{d_2}}} = \frac{e^{d_2}}{e^{d_1}+e^{d_2}}$$

This property is crucial for estimating the relative difficulties of questions using students of unknown ability (like High Schoolers, or random people from the Web).

(iv) The relative abilities of the group of students do not depend on the difficulties of the questions used to test them.

This property makes it possible to asses the abilities of students from the tests which where designed for different levels of ability.

This property helps to determine the cutoffs for placement into Finite-Math/Trig classes versus into Calculus 0.

Properties (iii) and (iv) say that "abilities" of students and "difficulties of questions" can be measured objectively, and independently of each other.

1.3. **Is this a good model?** The answer, according to Rasch is "yes and no". He says that it depends on the test. This mathematical model fits well some tests, and fits poorly some other tests. I suspect that the answer also depends on the population of students; the model should fit better when testing more homogeneous populations.

When applied to multiple choice questions, the Rasch model does not take into account "guessing". It also assumes that students had enough time to work out the questions so that time is not a factor. (Rasch gives formulas with corrections for time. But then one looses all the properties of the model that make it useful for us.)

The model poorly fits tests that have "multi-dimensional" character; however some of these tests (like IQ test) can be split into more homogeneous components and each may fit Rasch model.

The model may fail when applied over wide range of population. For example, questions $1, 2$ may have different difficulties for some students than in general population, sometimes to the point of reversing the order of difficulty.

We will use this model because of its mathematical properties with a clear understanding that it is just an approximation.

- We are interested in analyzing two-three placement tests that are limited in scope and level, thus are perhaps approximately one-dimensional. In future refinements we might want to group questions by a few "topics" if the need multi-dimensional analysis shows up.
- We hope to tailor the test to students of similar math background. But we probably are faced with the fact that difficulties of some questions might get reversed depending on student's math background.
- We will be analyzing students in relatively narrow range of abilities. Thus we will fit one Rasch model to a group of students geared for pre-calculus sequence, and another Rasch model to students geared for calculus. The upper end of the former group may perhaps approximate the lower end of the other group, and we will standardize both models to agree in this range.

2. Summary of Analyses

The charts in the following pages summarize six exams (2 related to College Algebra placement, and 4 related to calculus placement) taken by students enrolled at UC in 1999 and in 2000. We have data on **1988** test results this year, and close to a 1000 responses from previous years.

Rasch analysis allows us also to compare difficulties of questions across the exams. With Rasch analysis we can

- combine items from different exams into a single test
- match questions to student abilities within the range we need to test
- use known cutoffs from one exam to create cutoffs on another exam based on different questions
- space cutoffs to fill in the range of abilities observed, and to match the observed levels of quarterly change in student abilities.

In addition to Rasch analysis we can analyze items on the exams for

- correlation with exam total
- discrimination index
- correlation with course performance
- student average

Eventually, we would like to end up with a large and diverse bank of test items that are well correlated with class grades,and that cover a wide range of difficulty levels.

2.1. **Description of the exams.**

(1) Our Algebra Test 2000 consisted of 20 questions. It was given to a 912 students enrolled in
   - College Algebra I [in 3-rd week?]
   - College Algebra II [in 3-rd week?]
   - Finite Math and Calculus
   - Honors Finite Math
   - Calculus 0 [in 3-rd week?]

   The average difficulty of questions on this exam was (arbitrarily) set to 500, (i.e., five logits) and all other tests were standardized to match this scale.

   The overall average ability of students was found to be about 560. In part of our target population - College Algebra I,II the average was 489.39, with the standard deviation 96.62

(2) In fall of 2000, University College gave a 50-question placement test to 428 students. The difficulties of questions on this test were standardized using several common questions from Algebra 2000.

   The students who took this test are the closest approximation to the population we want to be bale to place into Algebra/College Algebra sequence. Their average was 412.64 and standard deviation 77.31.

(3) Calculus Placement 2000 contained 25 questions. It was given to 648 students enrolled in Calculus 1. About a 100 of these students (typically, at the lower end of scores) later moved to Calculus 0 and took the Algebra 2000 test mentioned above. The latter group was used to standardize the difficulties of questions between the two exams, by assuming that student abilities did not change.[4]

[4]This assumption is somewhat in doubt. But we can estimate the error to be about 30; we arrive at this number by linearly interpolating gains from Calc 0 to be about 17 points per week, and gains in College Alg to be perhaps 5-10 points per week (the upper number is consistent with linear interpolation from Calculus 0 gains; the lower number comes from cutoffs inferred from U College placement, which are about 50 points in 10 weeks).

The scale can be refined by swapping a or two question between Calculus and Algebra Placement exams. This would determine whether indeed all Calculus Exam difficulties are should be shifter up by 30. The precise location of the scales is not of major importance, but it might be helpful to place more precisely cutoff to Business Calc and Trig classes.

(4) Calculus Placement 1999 had two versions with 20 questions each (some overlapped). It was given in 1999 to about 700 Students enrolled in Calculus. This test had several questions in common with Calculus Placement 2000. The common questions were used to standardize other questions on this exam.

In 1999, the exam was given in third week of classes after intensive review. However, under Rasch model the fact that student abilities might have been higher should not affect comparative ranking of difficulties of questions.

(5) Final Exam was given to 84 students in 1999 at the end of Calculus 0. It had 30 questions. The level of difficulty of this exam was standardized on repeated questions from the 1999 Placement exam. This allowed us to measure student abilities at the beginning and at the end of Calculus 0.

Over the period of 8 weeks in a five-credit hour class, the average gain in the ability was 138. This implies a gain of about 17 per week in an intense five-credit hour class, and corresponds to about 10 per week in a three-credit class.

Based on this estimate, we can expect our difficulties of questions in Calculus to be shifted by about 30 points towards the low end

(6) A version of 25-question Michigan Placement test was available online to a number of students. I analyzed 163 tests that were collected before changes in software made it difficult to match student responses to questions. This test was standardized using common questions with other placement exams.

While this is a small data set, it provides alternative pool of questions. Furthermore, since we know the precise cutoff used at Michigan (set at 80%), we can estimate the corresponding level of ability to be about 800.

2.2. **Putting it all on a single scale.** Next few pages give distribution of student abilities and question difficulties in several of the exams we gave. The difficulty scale is standardized so that scores should be comparable, except perhaps for a shift of calculus exams up by about 30 units.

The charts show the distribution of abilities of students on the left hand side of the scale, and the distribution of the difficulties of questions on the right hand side of the scale. Data can be compared across the exams since the vertical axes are set using single scale.

How to read these charts. The ability/difficulty scale is marked at the left side of each chart. Symbols (`X`) show locations of students and questions on the ability scale. More marks horizontally corresponds to more students of a given ability level.

2.2.1. *Abilities of students aiming for College Algebra sequence.* To avoid small numbers, in examples below, I scaled the ability and difficulty measures A, D by a factor of a 100; for mathematical formulas use $a = A/100, d = D/100$. Thus 1 logit corresponds to 100 on the charts.

2.2.2. *U College Placement Exam.* This is probably the best approximation to the population we will need to handle.

```
TABLE 1.1 UColl Placement Exam -standardized on Al ucol-a.txt Nov 22 15:20 2000
INPUT: 428 PERSONS, 50 QUESTIS  ANALYZED: 424 PERSONS, 50 QUESTIS, 2 CATS  v2.82
--------------------------------------------------------------------------------

                        MAP OF PERSONS AND QUESTIS
MEASURE                              |                              MEASURE
  <more> ----------------------  PERSONS-+-  QUESTIS---------------------- <rare>
  800.0                            . +                                800.0
                                     |
                                     |
                                     |
                                     |
                                     |
                                   . |
                                     |
  700.0                              +                                700.0
                                     |
                                     |
                                     |
                                   # | X
                                   # |
                                   # | X
                                   # | X
  600.0                            . + XX                             600.0
                                   # |
                                   # | XXX
                                   # | XXXXX
                                     | XXX
                                 ### | XXXX
                                .### | XX
                              .###### | XX
  500.0                        ###### + X                              500.0
                             ######## | XXXXXX
                              ####### |
                             ######### | XX
                          ############ |
                            ########## | XXX
                         ############# |
                          ############ | XX
  400.0                    ########### +                              400.0
                          .########### | XXX
                            .########## | X
                                ######### |
                              .########## | XXX
                                 .###### |
                                ####### |
                                   ##### | X
  300.0                            ##### + X                          300.0
                                    .## |
                                   .### | X
                                      # | X
                                    .# |
                                     # |
                                     . |
                                    .# | X
  200.0                              . +                              200.0
                                     . |
                                       |
                                       |
                                       |
                                       |
                                       |
                                       |
  100.0                              # +                              100.0
  <less> ----------------------  PERSONS-+-  QUESTIS------------------<frequent>
 EACH '#' IN THE PERSON COLUMN IS   2 PERSONS; EACH '.' IS 1 PERSON
```

2.2.3. *Algebra Placement Exam 2000.* These are combined Students from College Algebra I through College Algebra II through Finite Math through Calculus 0. We should expect that most of these students make it above the College Algebra I cutoff.

```
TABLE 1.1 Algebra 2000 Placement Exam - all together  alg.txt Nov 14 21:48 2000
INPUT: 912 PERSONS, 20 QUESTIS  ANALYZED: 909 PERSONS, 20 QUESTIS, 2 CATS  v2.82
--------------------------------------------------------------------------------

                         MAP OF PERSONS AND QUESTIS
MEASURE                              |                                MEASURE
  <more> ----------------------  PERSONS-+-  QUESTIS---------------------- <rare>
  900.0                           .### +                                  900.0
                                       |
                                       |
                                       |
                                       |
                            .########## |
                                     . |
  800.0                              + |                                  800.0
                                       |
                                     . |
                                     . |
                           ########## |
                                       |
                                     . |
  700.0                 .########### + |                                  700.0
                                     . |
                                     . |
                    .################## |
                                       | XX
                  ################### | X
                                     . |
  600.0           .################### +                                  600.0
                                     # |
           .########################## |
                                    ## | XX
                ####################### | XX
         .############################## | X
                                   .## | X
  500.0      .######################### + X                                500.0
                                  ### | XXXX
                 ################### | X
                     ################ |
                                  .## | XX
                        .############ |
                                   # |
  400.0                    .########## +                                  400.0
                                     . |
                                ###### |
                                     . | XX
                               .##### |
                                       |
                                       |
  300.0                          .## + X                                  300.0
                                       |
                                       |
                                     . |
                                     . |
                                       |
                                       |
  200.0                              . +                                  200.0
                                       |
                                     . |
                                       |
                                       |
                                       |
                                       |
  100.0                                +                                  100.0
  <less> ----------------------  PERSONS-+-  QUESTIS------------------<frequent>
 EACH '#' IN THE PERSON COLUMN IS   3 PERSONS; EACH '.' IS 1 TO   2 PERSONS
```

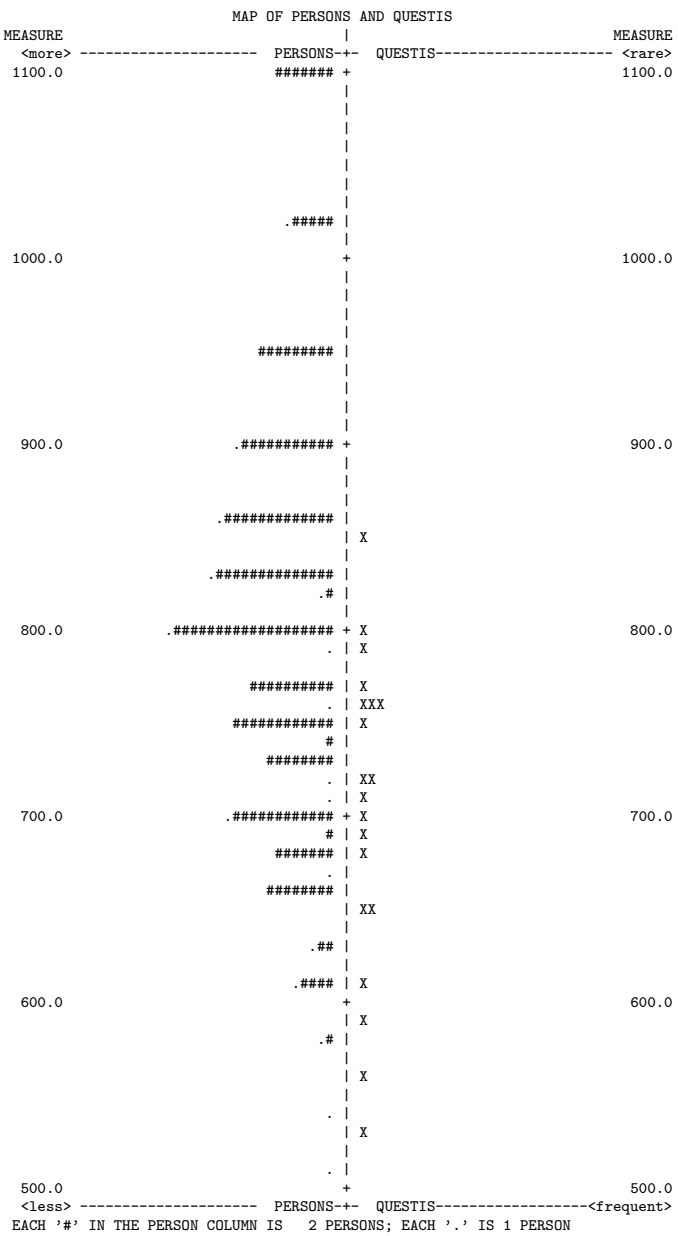
2.2.4. *Abilities of Calculus Students.* Students at the lowest end should perhaps be placed in College Algebra I. But we would expect most of the students to be above the Trig cutoff.

2.2.5. *Calculus Placement 2000.*

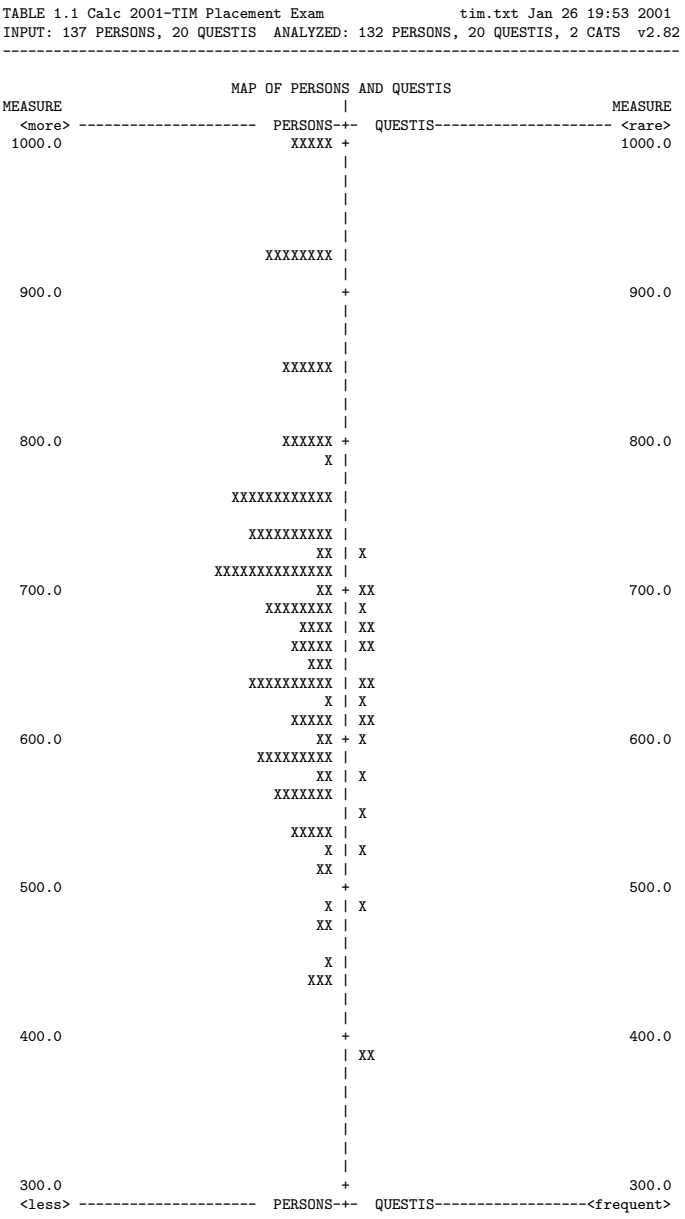
```
TABLE 1.1 Calculus 2000 Placement Exam            calc00.txt Nov 20 19:18 2000
INPUT: 648 PERSONS, 25 QUESTIS  ANALYZED: 646 PERSONS, 25 QUESTIS, 2 CATS  v2.82
--------------------------------------------------------------------------------

                          MAP OF PERSONS AND QUESTIS
MEASURE                                |                                MEASURE
  <more> ----------------------  PERSONS-+-  QUESTIS---------------------- <rare>
 1100.0                               # +                                 1100.0
                                      . |
                                        |
                                      . |
                                        |
                                        |
                                        |
                                  .##### |
 1000.0                                 +                                 1000.0
                                        |
                                        |
                                  .##### |
                                        | X
                                     .  |
                                   #### |
                                        |
  900.0                               . +                                  900.0
                                   #### |
                                      # | X
                             .########### |
                                     ##  |
                              ########## | X
                                     ##  | XX
                               .######## | XXX
  800.0                ################## + X                               800.0
                                    .##  |
                             .########### |
                                    ####  | XX
                          ################ |
                    ###################### | X
                                  ######  | X
                       .################## |
  700.0                           #####  + XX                               700.0
               .####################### | X
                  .###################### | X
                                  .#####  | X
                      ################## |
                                    .##  | X
                  .###################### | X
                           .############## | X
  600.0                               . + X                                600.0
                          ############## | X
                                        |
                                        |
                              ########## | X
                                       # | X
                                .######## |
                                       .  |
  500.0                               .# +                                 500.0
                                  .###  |
                                     .  |
                                     .  |
                                        |
                                     ##  |
                                        |
                                        |
  400.0                                 +                                 400.0
  <less> ----------------------  PERSONS-+-  QUESTIS----------------<frequent>
```

2.2.6. *Calculus Placement 1999.* This is one of the two versions of the exam in 1999. The exam was given after two weeks of intensive review. We might expect these students to have abilities about 35 points higher than students tested in 2000.



2.2.7. *Calculus 1 Trailer.* This is the distribution of abilities of students in trailer Calculus 1.



2.2.8. *Calculus 0 Final in 1999.* This is a group of students that weren't ready for calculus at the beginning of the quarter.

```
TABLE 1.1 Calc 0 Final Exam in 1999 standardized a 260-99.txt Nov 18  0:35 2000
INPUT: 84 PERSONS, 30 QUESTIS  ANALYZED: 83 PERSONS, 30 QUESTIS, 2 CATS    v2.82
-------------------------------------------------------------------------------

                        MAP OF PERSONS AND QUESTIS
MEASURE                                 |                                MEASURE
  <more> ----------------------  PERSONS-+-  QUESTIS---------------------- <rare>
 1100.0                                 +                                 1100.0
                                      X |
                                        |
                                        |
                                        |
                                        |
                                        |
                                      X | X
                                        |
                                        |
                                        |
 1000.0                               X +                                 1000.0
                                        |
                                        |
                                      X |
                                        |
                                     XX | X
                                      X |
                                        |
                                    XXX | XX
                                        |
  900.0                              XX + X                                900.0
                                        |
                                   XXXX | X
                                        |
                                XXXXXXX |
                                        |
                           XXXXXXXXXXX | XX
                                 XXXXXX | XX
                                        | X
                                    XX | XX
  800.0                                 +                                 800.0
                                   XXXX | X
                                 XXXXXX | X
                                        | XXX
                                 XXXXXX |
                                        | XX
                                    XXX |
                                      X | XX
                                XXXXXXXXX |
                                    XXX | XX
  700.0                                 +                                 700.0
                                        | X
                                        |
                                     XX |
                                      X |
                                    XXX |
                                        | X
                                        |
                                        |
                                        |
  600.0                               X +                                 600.0
                                        | X
                                        | X
                                      X | X
                                        |
                                        |
                                        | X
                                        |
                                        |
                                        |
  500.0                                 +                                 500.0
  <less> ----------------------  PERSONS-+-  QUESTIS----------------------<frequent>
```

3. Cutoffs for classes

Factors to be considered:

- What cutoffs were used in past exams?

- Since many of these courses feed into each other, how much can we expect students ability to increase in 10 weeks?
- What are the initial abilities of students that finish a course successfully?
- What are population statistics: how many students do we have in each group, and how many sections of a course at each level are usually offered?
- How many additional preparatory courses can we expect a student to take without disrupting their track towards their major?

One additional piece of information is that a multiple choice test consisting of 20 questions estimates student abilities with errors of about ±50.

3.1. **Calculus cutoffs. Calculus 1 680?** An upper limit for cutoff into Calculus 1 is about 800, based on Michigan placement Test. In last two placements our cutoff scores correspond to student abilities of 680 (in 2000) and 710 (in 1999). Our analysis of data in 1997, 1998, and 1999 suggests that our past cutoffs were at the right level.

Calculus 0 600? We have less data at this end, since we didn't place students out of Calculus 0 in the past. Inspection of abilities of students taking Calculus 0 Final in 1999 indicates that students with initial abilities below 600 tend to fail the class.

Furthermore, the observed average increase of ability of about 140 puts the cutoff into between 800-140=640 and 700-140 = 540 for the placement. On the other end of the spectrum, U College Exam places students into Trig at 580. Assuming 50 points gains in a quarter, this puts 260 at the level of 630. A potential 30 point mismatch in scales between Algebra and Calculus puts the cutoff at 600 in Calculus Placement Scale and at 630 in Algebra Placement Scale.

3.2. **Algebra cutoffs.** Our sole source of data to determine Algebra cutoffs is U College exam.

Using Difficulties of questions and Rasch formula U College real cutoffs correspond to the following abilities

Ability	210	280	350	410	430	470	540	580
Course	Intro I Ext	Intro I	Intro II	Intermediate	Topics	Col Alg I	Col Alg II	Fin Math or Trig
U College (50)	4.71	7.98	12.65	17.74	19.63	23.60	30.74	34.60
Guesser prob	98%	81%	19%	1%	0%			

Assuming 50 points gain per quarter, and taking into account the fact that we cannot estimate student abilities on a more refined scale, we may want to space the courses equal distance apart. Such cutoffs correspond to the following cutoffs on UC College Exam and Algebra 2000 exam.

Ability		250	**300**	350	**400**	420	**450**	500	550	600	680
Course		Intro I Ext	Intro I	Intro II	Intermediate	Topics	Col Alg I	Col Alg II	Fin Math or Trig	Calc 0	Calc 1
U College	out of 50	6.42	9.17	12.65	16.83	18.68	21.59	26.67	31.73	36.39	42.29
Alg 2000	out of 20	2.01	2.98	4.28	5.91	6.64	7.82	9.93	12.05	14.03	16.56

3.2.1. *Predicted placement of UCollege students.*

```
UColl Placement Exam - cutoffs
------------------------------------------------------------------------------

                         MAP OF PERSONS AND PLACEMENT
MEASURE                                 |                                COURSE
  <more> ----------------------  PERSONS-+-  COURSE---------------------- <higher>
  800.0                                . +
                                       |
                                       |
                                       |
                                       |
                                       |
                                     . |
                                       |
  700.0                                +
                                       | CALCULUS I (680)
                                       |
                                       |
                                     # |
                                     # |
                                     # |
                                     # |
  600.0                              . +  CALCULUS 0 (600)
                                     # |
                                     # |
                                     # |
                                       |  TRIG  or FINITE MATH (550)
                                   ### |
                                  .### |
                                .###### |
  500.0                          ###### +  COLLEGE ALGEBRA II (500)
                                ######## |
                                ####### |
                               ######### |
                        ################ |  COLLEGE ALGEBRA I (450)
                              ########## |
                         ############### |  TOPICS IN MATH (420)
                          ############## |
  400.0                    .############## +  INTERMEDIATE ALG (400)
                          .############## |
                            .############ |
                               ######### |
                              .########## |  INTRO ALG II (350)
                                  .###### |
                                 ####### |
                                   ##### |
  300.0                            ##### +  INTRO ALG I  (300)
                                    .## |
                                   .### |
                                      # |
                                     .# |  INTRO I EXT (250)
                                      # |
                                      . |
                                     .# |
  200.0                               . +
                                      . |
                                        |
                                        |
                                        |
                                        |
                                        |
                                        |
  100.0                               # +   PREPARATORY MATH    (no cutoff)
  <less> ----------------------  PERSONS-+-  COURSE -----------------------------<lower>
 EACH '#' IN THE PERSON COLUMN IS   2 PERSONS; EACH '.' IS 1 PERSON
```

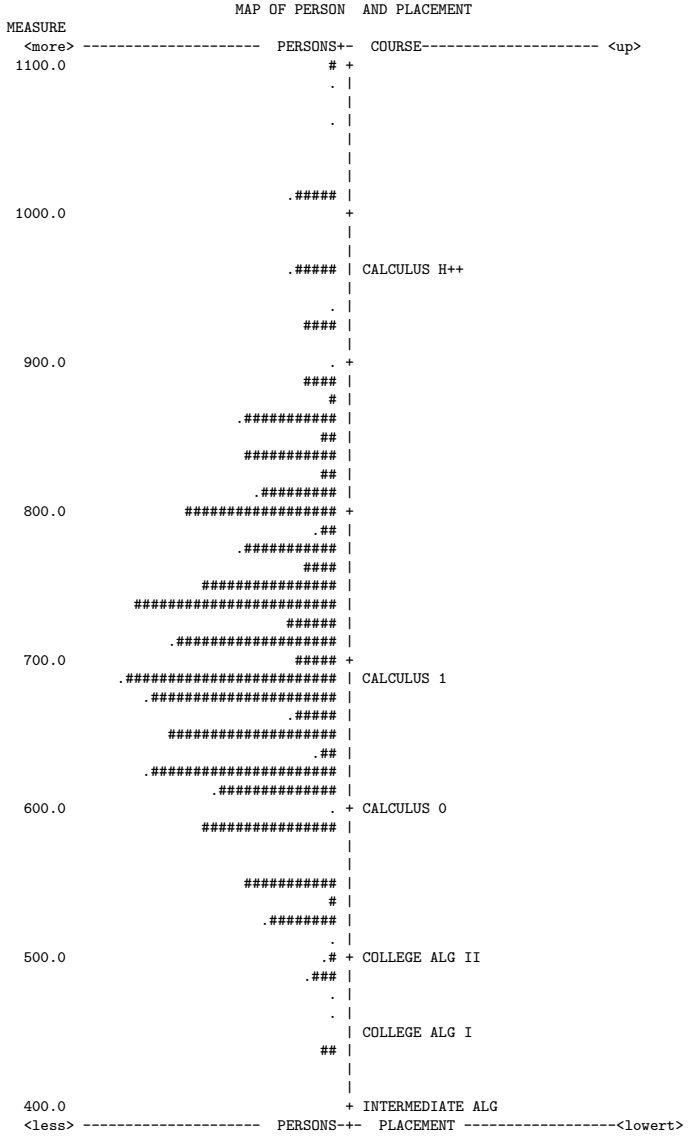
3.2.2. *Predicted placement of Students who took Algebra 2000 exam.*

TABLE 1.1 Algebra 2000 Placement Exam - how would students be placed

```
                          MAP OF PERSON  AND PLACEMENT
MEASURE
  <more> ---------------------  PERSONS+-  COURSE--------------------- <up>
  900.0                              .###|
                                        |
                                        |
                                        |
                                        |
                            .##########
                                      .+
  800.0                               |
                                      |
                                     .|
                                     .|
                           ###############|
                                      |
                                     .|
  700.0                  .################+
                                     .| CALCULUS I (680)
                                     .|
                     .###################|
                                      |
                   ####################|
                                     .|
  600.0             .####################|  CALCULUS 0 (600)
                                     #+
           .############################|
                                    ##|  TRIG  or FINITE MATH (550)
                ########################|
         .##############################|
                                   .##|
  500.0       .#########################|  COLLEGE ALGEBRA II (500)
                                   ###|
                  ####################+
                      ################|  COLLEGE ALGEBRA I (450)
                                   .##|
                         .#############|  TOPICS IN MATH (420)
                                    #|
  400.0                     .##########|  INTERMEDIATE ALG (400)
                                     .|
                                ######|  INTRO ALG II (350)
                                     .+
                                .#####|
                                      |
                                      |
  300.0                            .##|  INTRO ALG I  (300)
                                      |
                                      |
                                     .|
                                     .+
                                      |
                                      |
  200.0                               .|  INTRO I EXT (250)
                                      |
                                     .|
                                      |
                                      |
                                     +
                                      |
  100.0                               |
  <less> ---------------------  PERSONS| PLACEMENT ----------------------------<lower>
```

3.2.3. *Predicted placement of students entering calculus sequence.*

TABLE 1.1 Calculus 2000 Placement Exam



3.3. **Theoretical Curve for range of questions.** Graphs that I made indicate that the resolution can we expect from various arrangement for an exam does not depend to a significant degree on the distribution of difficulties if they are concentrated within an interval of about a 100 units. Theoretical computations are possible with several concentrated settings, or continuous uniform distribution in an interval. Of course, the simplest graph uses all questions of equal difficulty and just graphs expression (1).

By asking questions of constant difficulty d we may expect to cover the range of abilities of about $d-100$ to $d+200$. This is the width of 300, but notice the asymmetry and a bit of optimism, compare Figure **??**, where a bit narrower range of $d-100, d+150$ is marked. A pure-guess strategy gives on average a score of 4, so the "protection" from guessers is low at the low end.

Distributing questions uniformly over the range $d-100, d+100$ does not change the graph significantly: from the graphs I made with Maple I'd estimate the usable interval as $d-120, d+240$. Distributing the questions uniformly over a wider range $d-300, d+300$ we hit a law of diminishing returns, covering the range of $d-180, d+320$.

Mini-recommendations. Ideally, Algebra Placement questions should be restricted to the difficulty range 300 – 400, covering the range of abilities 230 – 590. This is enough to cover the entire range of courses from Intro-Ext I to Trig/Finite Math.

Ideally, all Calculus Placement questions should restricted to the difficulty range of 550–650, covering the abilities in the range 430 – 840. This would cover not only the calculus range (700+), but also quite a bit of lower courses, down to College Algebra II/Trig placement.

Together, the two tests would cover the range 230–840 with an overlap that is probably appropriate: both tests may place students into College Algebra II, Finite Math, or Trig. Of course, this is all nice theory. In practice, we have to use the questions we have.

4. Test construction using Rasch Model

The following is reproduced from the afterward to [6], written by D. Wright.

4.1. **Item Banking.** When a family of test items are constructed so that they can be calibrated along a single common dimension, and when they are employed so that they retain these calibrations over a useful ream of application, then a scientific tool of great simplicity and far-reaching potential becomes available. The "bank" of calibrated items can serve the composition of a wide variety of measuring tests. The tests can be short or long, easy or hard, wide in scope or sharp in focus. As far as "equating" measures is concerned, neither the difficulty nor shape of a test will matter. All possible scores on all possible tests will be completely equated with respect to the measures they imply through the common calibrations of their bank items. Whatever the test, its measures will be expressed on the one common variable defined by the bank.

4.2. **Test Design.** The explicit positioning of test items along the dimension they define makes best test design obvious and easy. If the picture of the variable provided by its items is kept in mind, it is impossible to make serious test design errors. Tests can be targeted on any region along the variable, which is represented by calibrated items. The items chosen to form a particular test can be selected

to spread out over the desired target region in what- ever way is thought to be potentially most informative. Items can be bunched at crucial points along the variable, if that seems important. But the item spacing most natural to common sense, namely, a uniform distribution of item difficulty from one end of the target to the other, approximates the best possible test design in most real situations. If $(M - 2S)$ and $(M + 2S)$ mark the target boundaries (or M and S represent the expected mean and standard deviation of the target), then $(H = M)$ and $(W = 4S)$ specify the difficulty level (height) and scope (width) of the best possible uniform test well enough for all practical purposes.

The number of items chosen for the test depends on the precision of measurement needed. If SE stands for the desired standard error of measurement in logits (or $SER = 1/SE$ for the desired precision in test score), then the number of items L needed for a test to have a standard error of SE logits (or SER scores) over most of its range can be estimated from the formula $L = 6/SE^2$. (For illustration and discussion see chapter 6 of Wright and Stone I979.)

4.3. **Sequential Testing.** The flexibility of test composition makes it possible to put together any kind of test that may be needed, without endangering its measuring connection with the underlying dimension. The basic formula for turning f, the proportion of correct answers on a test of average item difficulty H, into b_f, its corresponding measure in logits, is

$$b_f = H + \ln[f/(1 - f)].$$

(There is also a scaling factor of $[1 + (s/1.7)^2]^{1/2}$ for the second term, which comes into play when the item difficulties are spread out. But when s, the standard deviation of item difficulty spread, is less than 0.5 logits, then this scaling factor is less than 1.05 and can be disregarded.) A simple formula for optimal sequential testing follows directly. If each succeeding item is to be chosen on the basis of prior item performance, then d, the logit difficulty of the best next item to administer, can be estimated well enough for all practical purposes from h, the average logit difficulty of preceding items, and f, the proportion of these items which have been answered correctly. The resulting equation for choosing the best next item is

$$d = h + \ln[f/(1 - f)]$$

and the final measure equals the last difficulty chosen.

4.4. **Self-Tailoring.** Within reasonable limits, the persons being measured can make their own best choice of item difficulty as they go along. For this they are given a booklet of items arranged in order of increasing difficulty and invited to choose their own best starting point. If they feel strong, they may work ahead into harder items until they reach what they find is their best working level for this test. If they feel weak, they may stay with easy items. Capitalization on opportunity is controlled by scoring each person on all of the items contained in the uniform segment they select. But the length of a segment and even the number of items attempted within a segment can be chosen freely by the person taking the test. The segment can be defined by every second, third, or fourth item, as well as by every adjacent one. The person can find his own best level by doing, say, every fourth item from an easy beginning and then complete his testing by working a widening segment of adjacent items.

The most precise measures come from response patterns that end up about half right and half wrong; but all that is necessary for a finite estimate of a person's measure is at least one right and one wrong. Testing may be terminated in the old way by specifying L, the number of items to be taken, or by setting $SE = \frac{2.5}{\sqrt{L}}$, the precision to be reached, or in response to fatigue, or loss of motivation, or end of available time.

4.5. **Response Validation.** Because all bank items are calibrated on the same single dimension, which, in fact, they serve to define, any set of responses to any set of items, however chosen, can be scanned for their pattern validity and, if valid, can be used, along with the standard error of estimation by which every measure must inevitably be qualified, to estimate a measure for the test taker.

The evaluation of response pattern validity is an important, indeed obligatory, step in estimating a measure from a test performance. The items used will vary in their positions along the variable. This will happen when we spread items to cover the expected region of the targeted person. It will also be forced upon us by the inevitable limitation of item resources at any particular position along the variable. More to the point, as we come to appreciate the simplicity and necessity of evaluating response pattern validity, we will make sure that the items we select spread out enough to make an evaluation of response pattern effective.

When test items vary in their difficulty, then we can expect a person who participates in test taking in the way we intend to do better on the easier items than on the harder ones. Because our response model is explicit, this natural expectation can be formulated into a specific statistical test for the fit of any particular response pattern to the model. This allows, indeed obliges, us to evaluate the validity of each and every test performance before reporting any measures estimated from it.

4.6. **Item Bias.** The analysis of response pattern fit allows us to examine each person's item responses in complete detail. If we have any theory about our items that leads us to classify them by response format, page layout, booklet location, or, more importantly, by item text, topic, or approach, then we can analyze for each person the extent to which their re- responses are systematically disturbed by any of our item categories.

Whenever we find a systematic disturbance, then we can estimate the extent to which the unusual category is biased toward or against this particular person. In fact, there is no other statistically sound basis for the objective analysis of item or test bias. Bias estimated from groups of persons can never satisfy the right of each individual to be treated fairly regardless of group membership.

4.7. **Individual Diagnosis.** More important than a search for item bias is the diagnosis of each test taker's strengths and weaknesses and the interpretation of this diagnosis to find what might be the best thing to do next to help this particular individual. Most test takers are associated with educational programs supposedly dedicated to improving their skills. The only justification for testing under these circumstances is the intention to use test results to help test takers. For this, an item content diagnosis of each test taker's response pattern is essential. This analysis is also all that can be done statistically, since the response residuals from the measurement model manifest all of the diagnostic information that the test contains.

5. Historical introduction

Rasch [6] writes

> The work presented in this book has grown out of practical tasks assigned to me as a consultant.

Georg Rasch began his work on psychological measurement in 1945 when he helped Rubin and Rasmussen standardize an intelligence test for the Danish Department of Defence. The goal was to define simultaneously the meaning of the two concepts: *degree of ability* of a students, and *degree of difficulty* of a question. The following quote summarizes the objectives (I adjusted mathematical notation for consistency)

> If the statement that the ability of one person is twice the ability of another person, $a_1 = 2a_2$, say, shall be of any use, it must be valid in connection with more than one problem. It must remain in force *when we present the persons with several problems of the same kind*, e.g. the separate items of BPP-N[5] Thus, to attach a meaning to the statement we must be able to confront the persons with a battery of test problems, preferably of widely varying difficulty, which can act as a measuring instrument.
>
> ...
>
> As the next requirement of the concept we may propose that when one person is twice as able as another and one problem is twice as difficult as another, *the first person shall solve the first problem just as "easily" as the other person solves the second problem...*
>
> ...
>
> *...If we maintain that the abilities of two persons have certain values and the difficulties of some test items certain other values, it must be possible to check whether this is true or not.*

Rasch proceeds with an argument that the correct model must be statistical in nature and rely on assigning the probability of solving a problem. He says

> ...
>
> According to our assumptions this probability must be determined exclusively by the degree of ability of the person and the degree of difficulty of the problem ...

He ends up proposing the simplest he knows formula, equivalent to (1).

References

[1] E. B. Andersen The Numerical Solution of a Set of Conditional Estimation Equations. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 34, No. 1. (1972), pp. 42-54.

[2] Embretson, Susan E Item response theory for psychologists / Susan E. Embretson, Steven P. Reise Mahwah, N.J. : Lawrence Erlbaum Associates, Publishers, 2000

[3] Ghosh, Malay Inconsistent maximum likelihood estimators for the Rasch model (Statistics & Probability Letters, Volume: 23, Issue: 2, May 1, 1995, pp. 165-170)

[4] Haladyna, Thomas M Developing and validating multiple-choice test items / Thomas M. Haladyna Mahwah, N.J. : L. Erlbaum Associates, 1999

[5] Hashway, Robert M Assessment and evaluation of developmental learning : qualitative individual assessment and evaluation models / Robert M. Hashway Westport, Conn. : Praeger, 1998

[5]BPP-N is one of the forms of an IQ test analyzed by Rasch

[6] G. Rasch, Probabilistic Models for Some Intelligence and Attainment Tests, The University of Chicago Press, Chicago 1980.
[7] Wainer, Howard Computerized adaptive testing : a primer Lawrence Erlbaum Associates, 2000
[8] Handbook of modern item response theory / Wim J. van der Linden, Ronald K. Hambleton, editors New York : Springer, 1997.
[9] Wright B.D. & Stone M.H. Best Test Design, Chicago: MESA Press, 1979.
[10] Wright B.D. & Masters G.N. Rating Scale Analysis, Chicago: MESA Press, 1982.