TEST CONSTRUCTION FOR A DESIRED AVERAGE SCORE GOAL

Ann L. Ellis, Moyes College of Education, Weber State University, Ogden, UT, 84408, 801-626-7343, aellis4@weber.edu
Peter M. Ellis, Emeritus, Huntsman College of Business, Utah State University, Logan, UT 84322-3555, peter.ellis@usu.edu

ABSTRACT

This work shows how items can be selected from a test bank so that an examination which seeks attainment of a given outcome goal might be constructed. The model presented here develops a general model which is easily adapted to the needs of a particular user. It permits consideration of separate test sections and the required inclusion of a specified number of items from each section. It shows how selection of competing items might be accomplished.

INTRODUCTION

Many university courses have test banks available for the purpose of providing examinations in a convenient manner. Within the general requirement that the test outcome should provide a measure of student attainment of course content, there may be several specific aims of interest. The objective of interest here is to select test questions with the goal of having the average test result from all participants be as close as possible to a designated target percentage. Such a construction is beneficial in situations where performance consistency is sought. This might be important when testing different groups, such as common examinations for several sections of a university course. It would also be helpful when testing different groups at different points in time, such as annual pre-admission testing at professional schools.

The approach being described here envisions the development of a test bank of a set of items which have been pretested. During test development, item response theory (IRT) is applied in order to predict the probability that a participant with a specific ability level will correctly answer a given item [1]. A large pool of potential test items is field tested and item responses are analyzed for such indices as difficulty, discrimination power, reliability and validity. Items not meeting desired values are eliminated. Tests are then constructed by selecting items from the remaining pool. The following describes the item selection process.

For any item, there might exist differing percentages of correct responses between participants who receive differing overall test scores. There will be cohorts ranging from high achievers to low achievers. For example, the high achieving group might be the top quartile, the middle achievers would come from the second and third quartiles and the low achievers would then come from the lowest quartile. One simple parameter is the index of discrimination **D**. It is determined as $\mathbf{D} = p_u - p_l$, where p_u is the proportion in the upper group who answer the item correctly and p_l is the proportion of the lowest group who answer the item correctly [2].

This work will provide for the existence of C cohorts, and the numerical example will set C = 3. Also, the model includes K test sections. These might be thought of as different chapters or different chapter sections to be included in the test material. In the application shown here, K was set to 3. For any test item I, the average percentage of correct response from cohort c is given as D_{ic}. This value will be known because of the pretesting of the items in the test bank.

There is not a great amount of background research available concerning the approach to test construction where test outcome goals can be attained through test design. Hambleton, Swaminathan & Rogers [3] presented the subject matter of item response theory. Theunissen [4] was the first to apply integer programming to this problem. Van der Linden [5] [6] followed with applications to psychological and educational testing. The work developed here presents the development a performance goal objective in the optimization process. Armstrong, Belov & Weissman [1] also presented a bivalent integer approach.

THE MATHEMATICAL MODEL

Several variables and parameters are needed in the model. They are:
S = the number of sections on the test
s = the section number on the test
C = the number of cohorts taking the test
D_{ics} = average correct response percentage on item i from cohort c in section s of the test
N_s = number of items to include from section s
$\underline{N_s}$ = specified value of N_s
n_s = number of items available in the test bank for section s
X_{is} = 1 if item i in section s is of the test bank is included, and = 0 if not
m = number of items from a restricted set that may be permitted for inclusion in the test
S_{cs} = aggregate score from cohort c in section s

The general objective function of the formulation will be designated as $f(\underline{X})$, where $\underline{X}$ is the vector of the X_{is} values. With that, the general mixed bivalent integer formulation of the item selection problem is:

maximize or minimize $Z = f(\underline{X})$ (Ia)
subject to:

$$\sum_{i=1}^{n_s} X_{is} = N_s \tag{Ib}$$

$$\sum_{i=1}^{n_s} D_{ics} X_{is} - S_{cs} = 0 \tag{Ic}$$

$$N_s = \underline{N_s} \tag{Id}$$

all Ns, $S_{cs} \geq 0$, all $X_{is} = \{0,1\}$

This is a mixed bivalent integer formulation where the Ns and S_{cs} are continuous variables and the X_{is} variables are bivalent. Constraints (Ib) establish the number of items that are to be chosen from each of the S test sections. The constraints of (Ic) calculate the expected total score for each cohort c. The limits provided in (Id) dictate the number of items that are to come from each section s. Note that the total number of items included on the test is the sum of the $\underline{N_s}$ values.

The model presented here seeks to minimize the positive or negative deviation in overall average test score from a desired goal. Often test preparers would like to have outcome consistency over several test experiences. This can come about by selecting a desired average score for all test takers. Denote this desired average as G. For instance, G = 80 would indicate that test questions should be selected so that the overall average for all test takers is as near to 80 as possible. The number of test items will be $N = N_1 + \ldots + N_S$. If the actual average score is above the target value, the excess amount is designated as GP. If the actual average is below the target value the shortage amount is designated as GM.

The aggregate expected score on test section s for cohort c is $A_{cs} = S_{cs}/N_s$. Therefore, the average score for all cohorts in test section s is

$$A_s = \sum_{c=1}^{C} (A_{cs} / C)$$

This yields the average test score for all test participants to be $A = \sum_{s=1}^{S} (A_s / S)$. The amount of deviation from the goal G is then written as

$$A - GP + GM = G \qquad \text{(Ie)}$$

This new equation must be added to the constraint set. Then, the objective function is

$$\text{minimize} \quad z = GP + GM \qquad \text{(Ia5)}$$

AN EXAMPLE OF THE MODEL

Let there exist a test bank with 99 questions. Further, there are S = 3 test sections and C = 3 cohort groups. These cohort groups represent the top, average and low achievers. Each test item has been pretested so that expected percentages of correct scores D_{ics} on each question i in section s by cohort c is known. There are 33 items in each of the three test sections and the test is to use exactly 10 questions from each section (N1 + N2 + N3 = 10 + 10 + 10 = 30). The goal is to produce a test where the overall average score is G = 80%.
A Monte Carlo simulation has been used to generate the D_{ics} values. For any item the expected percentage correct for cohort 3 was generated randomly with the formula D3 = INT(100*(RN +(.6)*(1-RN))), where RN is a uniformly distributed pseudo-random number on (0,1). For cohort 2 the expected correct percentage is D2 = INT(100*(.9*D3 + RN*(D3-.9*D3))), and D1 is generated in the same way. The values obtained from the simulation are the ones found in constraints 5) – 13).

THE TARGETED AVERAGE SCORE MODEL

The model with goal (Ia5) is shown in listing 1 and the computer output is shown in listing 2. The objective function is shown in 1) of listing 1. We seek to develop a test with an overall average score of 80%. The constraints of 2) to 4) set the item numbers to be chosen from each section. The constraints in 5) to 13) specify the expected aggregate scores in each section for each cohort. These values come from the Monte Carlo simulation, as shown in table 1. Next, 14) – 16) bring about the requirement that each section of the test must include exactly 10 items. Constraints 17) to 25) calculate the average score A_{cs} for cohort c in test section s. Constraints 26) to 28) yield the average score for each test section for all cohorts and constraint 29) obtains the overall test average for all participants. Constraint line 30) shows that the deviation of the actual test average from the goal of 80% to be either GP or GM. Finally, the several X_{is} are also designated as bivalent.

Listing 1
The Mixed Integer Formulation For the Goal of Having the Overall Test Score of 80%

```
      MIN     GP + GM
 SUBJECT TO
       2)   X011 + X021 + X031 + X041 + X051 + X061 + X071 + X081 + X091
     + X101 + X111 + X121 + X131 + X141 + X151 + X161 + X171 + X181 + X191
     + X201 + X211 + X221 + X231 + X241 + X251 + X261 + X271 + X281 + X291
     + X301 + X311 + X321 + X331 - N1 =    0
       3)   X012 + X022 + X032 + X042 + X052 + X062 + X072 + X082 + X092
     + X102 + X112 + X122 + X132 + X142 + X152 + X162 + X172 + X182 + X192
     + X202 + X212 + X222 + X232 + X242 + X252 + X262 + X272 + X282 + X292
     + X302 + X312 + X322 + X332 - N2 =    0
       4)   X013 + X023 + X033 + X043 + X053 + X063 + X073 + X083 + X093
     + X103 + X113 + X123 + X133 + X143 + X153 + X163 + X173 + X183 + X193
     + X203 + X213 + X223 + X233 + X243 + X253 + X263 + X273 + X283 + X293
     + X303 + X313 + X323 + X333 - N3 =    0
       5)   53 X011 + 55 X021 + 65 X031 + 71 X041 + 82 X051 + 56 X061
     + 59 X071 + 88 X081 + 75 X091 + 66 X101 + 56 X111 + 56 X121 + 59 X131
     + 65 X141 + 86 X151 + 84 X161 + 60 X171 + 60 X181 + 82 X191 + 77 X201
     + 77 X211 + 64 X221 + 63 X231 + 74 X241 + 69 X251 + 82 X261 + 74 X271
     + 50 X281 + 91 X291 + 73 X301 + 68 X311 + 61 X321 + 74 X331 - S11
     =    0
       6)   59 X011 + 60 X021 + 67 X031 + 79 X041 + 84 X051 + 62 X061
     + 64 X071 + 90 X081 + 83 X091 + 73 X101 + 63 X111 + 59 X121 + 66 X131
     + 66 X141 + 90 X151 + 88 X161 + 65 X171 + 62 X181 + 89 X191 + 85 X201
     + 86 X211 + 67 X231 + 79 X241 + 76 X251 + 85 X261 + 81 X271 + 55 X281
     + 92 X291 + 74 X301 + 71 X311 + 68 X321 + 81 X331 - S21 =    0
       7)   65 X011 + 65 X021 + 69 X031 + 80 X041 + 85 X051 + 68 X061
     + 65 X071 + 99 X081 + 87 X091 + 81 X101 + 68 X111 + 62 X121 + 68 X131
     + 72 X141 + 95 X151 + 93 X161 + 68 X171 + 68 X181 + 99 X191 + 86 X201
     + 93 X211 + 71 X221 + 76 X231 + 86 X241 + 77 X251 + 91 X261 + 85 X271
     + 62 X281 + 97 X291 + 80 X301 + 75 X311 + 76 X321 + 83 X331 - S31
     =    0
       8)   68 X012 + 61 X022 + 74 X032 + 69 X042 + 59 X052 + 87 X062
     + 65 X072 + 72 X082 + 68 X092 + 56 X102 + 62 X112 + 56 X122 + 70 X132
     + 67 X142 + 58 X152 + 69 X162 + 95 X172 + 78 X182 + 91 X192 + 66 X202
     + 73 X212 + 62 X222 + 61 X232 + 68 X242 + 67 X252 + 78 X262 + 65 X272
     + 66 X282 + 74 X292 + 80 X302 + 63 X312 + 56 X322 + 63 X332 - S12
     =    0
       9)   71 X012 + 68 X022 + 81 X032 + 77 X042 + 60 X052 + 92 X062
     + 69 X072 + 74 X082 + 74 X092 + 57 X102 + 66 X112 + 59 X122 + 73 X132
     + 70 X142 + 64 X152 + 71 X162 + 96 X172 + 80 X182 + 97 X192 + 67 X202
     + 79 X212 + 64 X222 + 63 X232 + 70 X242 + 74 X252 + 81 X262 + 66 X272
     + 68 X282 + 78 X292 + 88 X302 + 65 X312 + 59 X322 + 68 X332 - S22
     =    0
      10)   75 X012 + 76 X022 + 83 X032 + 84 X042 + 62 X052 + 95 X062
     + 74 X072 + 80 X082 + 82 X092 + 60 X102 + 72 X112 + 63 X122 + 74 X132
     + 71 X142 + 67 X152 + 78 X162 + 99 X172 + 89 X182 + 98 X192 + 72 X202
     + 88 X212 + 71 X222 + 70 X232 + 76 X242 + 77 X252 + 83 X262 + 70 X272
     + 69 X282 + 80 X292 + 93 X302 + 66 X312 + 61 X322 + 72 X332 - S32
     =    0
```

```
 11)   77 X013 + 56 X023 + 73 X033 + 63 X043 + 85 X053 + 58 X063
+ 86 X073 + 82 X083 + 78 X093 + 84 X103 + 81 X113 + 74 X123 + 71 X133
+ 80 X143 + 61 X153 + 66 X163 + 84 X173 + 50 X183 + 81 X193 + 63 X203
+ 79 X213 + 61 X223 + 51 X233 + 59 X243 + 74 X253 + 53 X263 + 80 X273
+ 77 X283 + 74 X293 + 55 X303 + 74 X313 + 75 X323 + 71 X333 - S13
=    0
 12)   86 X013 + 63 X023 + 81 X033 + 69 X043 + 87 X053 + 62 X063
+ 91 X073 + 89 X083 + 80 X093 + 86 X103 + 85 X113 + 81 X123 + 79 X133
+ 89 X143 + 68 X153 + 74 X163 + 86 X173 + 55 X183 + 85 X193 + 67 X203
+ 88 X213 + 64 X223 + 56 X233 + 62 X243 + 77 X253 + 56 X263 + 83 X273
+ 84 X283 + 82 X293 + 57 X303 + 76 X313 + 78 X323 + 75 X333 - S23
=    0
 13)   94 X013 + 68 X023 + 84 X033 + 76 X043 + 92 X053 + 65 X063
+ 96 X073 + 92 X083 + 89 X093 + 92 X103 + 90 X113 + 86 X123 + 81 X133
+ 99 X143 + 72 X153 + 75 X163 + 96 X173 + 61 X183 + 93 X193 + 68 X203
+ 97 X213 + 70 X223 + 61 X233 + 68 X243 + 78 X253 + 62 X263 + 87 X273
+ 87 X283 + 88 X293 + 61 X303 + 84 X313 + 83 X323 + 78 X333 - S33
=    0
 14)   N1 =    10
 15)   N2 =    10
 16)   N3 =    10
 17) - S11 + 10 A11 =    0
 18) - S21 + 10 A21 =    0
 19) - S31 + 10 A31 =    0
 20) - S12 + 10 A12 =    0
 21) - S22 + 10 A22 =    0
 22) - S32 + 10 A32 =    0
 23) - S13 + 10 A13 =    0
 24) - S23 + 10 A23 =    0
 25) - S33 + 10 A33 =    0
 26) - A11 - A21 - A31 + 3 A1 =    0
 27) - A12 - A22 - A32 + 3 A2 =    0
 28) - A13 - A23 - A33 + 3 A3 =    0
 29) - A1 - A2 - A3 + 3 A =    0
 30) - GP + GM + A =    80
END
```

The optimal solution portrayed in listing 2 includes test items 3, 6, 8, 15, 16, 17, 19, 26 , 29 and 32 from section 1. the items from section 2 are numbers 4, 6, 15, 17, 19, 23, 24, 26, 28 and 31. For section 3 the selected items are 1, 3, 5, 7, 9, 11, 13, 17, 20 and 28. These all have the associated X_{is} variable equal to 1. The objective function has the optimal value 0. Therefore, the actual average test score actually equaled the desired value of 80%. This is also seen in listing 2, where A = 80. For further exploration, the model will be used again under the requirement that the desired overall test average be as close as possible to 90%. The only change is in line 30) of listing 2), which will now be - GP + GM + A = 90. The computer solution is given in listing 3. The objective function value in listing 3 is 3.411. The listing also gives GM = 3.41 and A = 86.58. Therefore, the selected test items yield an overall test average of 86.58%, which is 3.411% short of the 90% goal. No other collection of test questions will yield an overall average closer to 90%. Listing 3 shows that the selected items from section one are numbers 5, 8, 9, 15, 16, 19, 20, 21, 26 and 29. From section 2 the included items are 3, 4, 6, 17, 18, 19, 21, 26, 29 and 30. The items drawn from section 3 are 1, 5, 7, 8, 10, 11, 14, 17, 19 and 21.

CONCLUSION

Test item selection for the purpose of pursuing the goal of having an overall test average come as close as possible to a stated percentage has been developed here. It is carried out with a mixed integer programming model. Two example problems were solved. The first yielded a test that fully accomplished the goal of having an overall actual average be 80%. The second had the goal of having the overall test average be 90%. The selected test items yielded an overall performance average of 86.58%, just 3.41 points short of the 90% goal. The general model allows establishing several test sections, such as different chapters of the textbook. It also admits treatment of a specified number of performance cohorts. These might be considered to be groupings ranging from high performers to low performers. It is easy to contemplate the expansion of this model to other testing goals. Listings 2 and 3 are available from the authors.

REFERENCES

[1] Armstrong, R., Belov, D. & Weissman, A. Developing and Assembling the Law School Admission Test. *Interfaces*, March-April 2005, *35*, 140-151.

[2] Crocker, L. & Algina, J. *Introduction to Classical and Modern Test Theory*. Harcourt Brace Jovanovich College Publishers, Fort Worth, TX, 1986.

[3] Hambleton, R. K., Swaminathan, H. & Rogers, H. J., *Fundamentals of Item Response Theory*. Sage Publications, Newbury Park, CA, 1991.

[4] Theunissen, T. Binary Programming and Test Design, *Psychometrica*, 1985, *50* (4), 411-420.

[5] Van der Linden, W. J. Optimal Assembly of Psychological and Educational Tests. *Appl. Psych. Measurement*, 1998, 22 (3), 195-211.

[6] Van Der Linden, W. J., Optimal Assembly of Tests With Item Sets. *Appl. Psych. Measurement*, 2000, *24* (3), 225-240.