Instructor's Guide for TAP

George Johanson and Gordon Brooks

You should experiment/play-with the TAP software for awhile. There are a number of features to try and it is quite easy to do this if you use that portion of the software that simulates or generates data. For the classroom, you might begin by using the defaults to generate the following sort of data and data analysis.

The initial portion of the complete results should resemble the following:

```
TITLE:   Generated Data:
COMMENT: Cases=20, Items=30, Seed=350712,
*******************************************************************************
Examinee Analysis
*******************************************************************************

                                          ~68% C.I.     ~95% C.I.
Examinee             Score Percent        (Raw Score)   (Raw Score)
-------------------- ----- ------- ----- ------------- -------------
Person_001             17   56.67%        (14.9- 19.1)  (12.8- 21.2)
...
Person_020             26   86.67%        (23.9- 28.1)  (21.8- 30.2)
===============================================================================

Number of Examinees = 20
Minimum Score       = 12.000 = 40.0%
Maximum Score       = 29.000 = 96.7%
Median Score        = 23.500 = 78.3%
Mean Score          = 22.850 = 76.2%
Standard Deviation  = 3.940
Variance            = 15.527
Skewness            = -0.874
Kurtosis            = 0.733
```

As a beginning, you can have your students briefly describe these examinee scores using the statistics above. The negative skewness might be explained or described as being due to a slight 'ceiling effect' in that the mean score for this examination is well above 50%. This could be the result of a relatively easy test and/or a relatively skilled group of examinees.

Confidence Intervals

To make sense of the confidence intervals, you might conduct a 'thought experiment' where you hypothesize about the score distribution of, say, Person_001 where this individual is retested many times without recall of prior testing or fatigue. The mean of this hypothetical (and normally distributed) distribution of scores is called the individual's 'true score'. The probability that Person_001's true score is captured by the ~68% confidence interval (about the raw score of 17) for this individual is approximately 0.68. A similar interpretation is possible for the ~95% confidence interval. A brief discussion of the desired qualities (i.e., being small) of such

confidence intervals is useful. If you want, you can compute such intervals using the standard error of measurement, SEM, where:

$$SEM = s_X \sqrt{1 - r_{XX'}}$$

and s_X is the standard deviation of raw scores and $r_{XX'}$ is the reliability. It is informative to note that this can be thought of as the standard deviation of the (hypothetically) many scores of individuals such as Person_001. An ~68% confidence interval can be constructed using the following:

$$C.I. = raw\ score \pm P_z\%SEM$$

where $P_z\%$ is the z-score that, two-sided, encloses and area of P percent in a standard-normal distribution. A later portion of the output indicates that:

```
KR20 (Alpha)             = 0.719
KR21                     = 0.672
SEM (from KR20)          = 2.089
```

From our formula for SEM, we can confirm the SEM value by computing:

$$SEM = s_X \sqrt{1 - r_{XX'}} = 3.940\sqrt{1 - 0.719} = 2.089$$

and the C.I.,

$$C.I. = raw\ score \pm P_z\%SEM = 17 \pm 1SEM = 17 \pm 2.1 = (14.9, 19.1)$$

Using a $P_z\%$ value of 1.96, the ~95% C.I. from the output can also be confirmed. As with all confidence intervals, a key to an accurate interpretation is that we not encourage/allow the conception of a 'fixed' interval that catches a 'varying' true score a certain percentage of the time, but rather a fixed true score for each individual that is caught by a multitude of such intervals a certain percentage of the time. Note that we are assuming that the magnitude of such intervals (or the variation of repeated scores) is uniform for all persons and all raw scores…a pretty strong assumption.

Graphs

The next portion of the output is a bar graph of raw scores:

```
Bar Graph
*********

Score    Count  Graph (each @ represents 1 case)
-------  -----  --------------------------------

 12.00       1    @
 13.00       0
 14.00       0
 15.00       0
 16.00       0
 17.00       1    @
 18.00       0
 19.00       1    @
 20.00       2    @@
 21.00       2    @@
 22.00       2    @@
 23.00       1    @
 24.00       1    @
 25.00       3    @@@
 26.00       4    @@@@
 27.00       0
 28.00       1    @
 29.00       1    @
```

The negative skewness is rather apparent in this graph. The Stem-and-Leaf display also shows this:

```
Stem-and-Leaf Display
*********************

Stem Leaves (width=10)
---- -----------------

  1 . 2
  1 . 79
  2 . 00112234
  2 . 555666689
```

Item Analysis: Part I

The next portions of the output are the *Item and Test Analysis* and the *Additional Item Analysis*.

```
*****************************************************************************
Item and Test Analysis
*****************************************************************************

        Number  Item    Point   Disc.  # Correct  # Correct   KR20 if Item
Item    Correct Diff.  Biserial Index in High Grp in Low Grp    Deleted
------- ------- ------ -------- ----- ---------- ---------- ------------
Item_01    17    0.85   -0.336 -0.167   5 (0.83)    5 (1.00)     0.748
Item_02    18    0.90    0.241  0.200   6 (1.00)    4 (0.80)     0.716
…
Item_30    14    0.70    0.197  0.233   5 (0.83)    3 (0.60)     0.724
=============================================================================

Number of Items           = 30
Mean Item Difficulty      = 0.762
Mean Item Discrimination  = 0.308
Mean Point Biserial       = 0.329
KR20 (Alpha)              = 0.719
KR21                      = 0.672
SEM (from KR20)           = 2.089
High Grp Min Score (n=6)  = 26.000
Low Grp Max Score (n=5)   = 20.000

Minimum Item Diff.   =  0.400, Maximum Item Diff.   = 1.000
Minimum Disc. Index  = -0.167, Maximum Disc. Index  = 0.833
Minimum Pt. Biserial = -0.336, Maximum Pt. Biserial = 0.632

To obtain a Reliability of .80, the test must be 1.56 times longer,
  for a total of 47 items of similar quality to those in the test now.

To obtain a Reliability of .90, the test must be 3.52 times longer,
  for a total of 106 items of similar quality to those in the test now.


*****************************************************************************
Additional Item Analysis
*****************************************************************************

        Scale Mean if  Scale SD if    Corrected    SEM if Item   Biserial
Item    Item Deleted   Item Deleted  Pt. Biserial    Deleted    Correlation
------- -------------  ------------  ------------  -----------  -----------
Item_01    22.000          4.074        -0.412        2.045       -0.516
Item_02    21.950          3.879         0.168        2.068        0.417
…
Item_30    22.150          3.877         0.082        2.038        0.260
=============================================================================

Mean Biserial Correlation = 0.481
Minimum Biserial Corr.    = -0.516
Maximum Biserial Corr.    = 1.000
```

Difficulty and Discrimination Indices

There is a large amount of information in this section of the output. We can use these TAP results with students to explain item difficulty and discrimination. In particular, with the first portion of the complete results, you will likely want to:

- Define item difficulty, p_i, as the proportion of person having item 'i' correct. While not shown on this edited section of output, note that everyone had item #13 correct and therefore the item difficulty is reported as 1.0. You might also note that item 'difficulty' is a bit of a misnomer since a larger value for this index indicates an easier item.
- Define item discrimination index, Di, as the difference, $p_H - p_L$, where p_H is the item difficulty in the highest scoring (top 27%) group and p_L is the item difficulty in the lowest scoring (bottom 27%) group. The item difficulties in the high and low groups are labeled '# Correct in High Group' and '# Correct in Low Group', respectively, in the output. Note that item #1 has a negative value for this index indicating that a higher proportion of persons in the presumed the less knowledgeable group had success than those in the presumed more knowledgeable group. This is a problem. Items such as this are undesirable in any context since they function in precisely backwards direction from that which is desired. We will revisit this later.
- The point-biserial correlation, pt-bis$_i$, is computed as a Pearson correlation between item 'i' (scored 0/1) and the total score. The corrected point-biserial correlation, c-pt-bis$_i$, is computed as a Pearson correlation between item 'i' (scored 0/1) and the total score of all items except item 'i', itself. Note the value of this index for item #1.
- Define the biserial correlation, bis$_i$, as an 'adjusted' point-biserial correlation between item and total score where the purpose of the adjustment is to estimate the value of a Pearson correlation between the item and total as if the item scores were normally distributed and not binary.
- Note that the item discrimination index, point-biserial correlation, corrected point-biserial correlation, and biserial correlation are all used similarly as measures of an item's ability to discriminate among the more and less knowledgeable examinees.
- Also note that these indices, in contrast to the item difficulty, are, conceptually, norm-referenced by their very nature. That is, we assume that one purpose of the assessment under consideration is to identify individual differences among examinees.

Some Small Sample Cautions

Generally speaking, we must always be very cautious when interpreting parameter estimates from small samples. Standard errors of these estimates will often be so large that conclusions about the value (or, maybe even the direction) of the estimate will be in question. For example, an estimate of the standard error of a proportion, *p*, is given by:

$$s_p = \sqrt{\frac{p(1-p)}{n}}$$

So, if we have, say, an item difficulty of 0.90 (as we do with Item_02 for these data), then an ~95% C.I. would be: $p_i \pm 1.96 s_p$. Or, for our example the standard error is:

$$s_p = \sqrt{\frac{0.90(1-0.90)}{20}} = 0.067$$

Therefore, the ~95% C.I. is $0.90 \pm 1.96(0.067)$, or 0.90 ± 0.13, or (0.77, 1). We would too often be incorrect to assume that item difficulty estimates within this interval derive from different population values. Certainly, it would be risky to suppose that Item_01 (with difficulty 0.85) is really easier than Item_02 with a sample of only 20 examinees.

An even more dramatic example would be Item_15 with difficulty 0.50. The calculation in this case results in an ~95% C.I. that is even larger: (0.28, 0.72). Given all of this, we can still find useful implications within small sample sets of data, but we need to be quite a bit more conservative (or tentative) than we might be with a larger N.

Reliability

A key observation at this point is that the preferred measure of (internal consistency) reliability, Kuder-Richardson's formula #20 or, alternatively, Cronbach's Coefficient Alpha is computed to be 0.719 for these data. KR-20 and Alpha are identical for such binary data as these. KR-20 can be computed only for binary data while Alpha can be computed for items that are scored in virtually any fashion. KR-21 is easier to compute by hand (it doesn't require item-level data as is the case with KR-20 and/or Alpha). KR-21 will always be less than or equal KR-20 (equality prevails when all items have the same difficulty…differences in the indices are greater as item difficulty becomes more diverse). For almost all applications, KR-20 is the preferred index.

Explaining the conceptual meaning of indices of internal consistency can be difficult, but if you have already discussed internal consistency using the concept of a split-half correlation or reliability index, you can indicate that, if you use one particular way of computing a split-half correlation, then Alpha and KR-20 can be thought of as the mean of all possible split-halves. That is, Alpha is conceptually similar to a split-half, but has the advantage of not being dependent upon any particular splitting scheme. Alpha can also, approximately, be thought of as the value of the mean correlation among items increased using the Spearman-Brown prophesy

formula with 'L' equal to k, the number of items on the test. That is, Alpha/KR-20 is a rather direct reflection of item homogeneity where 'homogeneity' is interpreted as a correlation.

Spearman-Brown Prophesy Formula

To illustrate the prophesy computation for longer/shorter tests, you can change the defaults of TAP to N=200, k=30, seed=4487, #options=4, moderate difficulty and note that KR-20 is now 0.728. We will use the Spearman-Brown formula to estimate KR-20 for a test with 60 items.

Suppose test A has reliability $r_{AA'}$ and k_A items. If we increase or decrease the length of test A by adding or deleting items that are similar in quality to the existing items, then an estimate of the reliability of the new test, B, with k_B items is:

$$r_{BB'} = \frac{Lr_{AA'}}{1+(L-1)r_{AA'}}$$

where L is the ratio of the number of items on test A to test B or $L = \frac{k_B}{k_A}$. In this case, L=2.

For our new data, we have: $r_{BB'} = \frac{2(0.728)}{1+(2-1)(0.728)} = \frac{1.456}{1.728} = 0.843$

Now, choose *Generate Sample Data* in TAP and then *Go to Data Editor*. Change the number of items, k, from 30 to 60 and then select *Generate*, followed by *OK*, and finally *Analyze*. The resulting KR-20 is 0.839…very close to the prophesized value of 0.843.

You can also confirm, say, the prophesized L of 3.52 for a reliability of 0.90 for the original data by computing:

$$r_{BB'} = \frac{3.52(0.719)}{1+(3.52-1)(0.719)} = \frac{2.531}{2.812} = 0.900$$

Alternately, you can solve for L in the above formula to get:

$$L = \frac{r_{BB'}(1-r_{AA'})}{r_{AA'}(1-r_{BB'})} = \frac{0.90(1-0.719)}{0.719(1-0.90)} = \frac{0.2529}{0.0719} = 3.517$$

Miscellaneous Observations

It is interesting to point out to students that the mean item difficulty (over persons), 0.762 is precisely the same as the mean person percentage correct (over items), 76.2%. That is, there is a pleasant duality…examinees take items and items 'take' examinees…and the performance of either can be expressed as a proportion or percentage of successes.

Point-biserial correlations are simply computed as Pearson product-moment correlations between an item response (with binary, say, 0/1 scoring) and total or raw scores over items. The corrected version is the correlation between the item and the total or raw score of all of the <u>other</u> items. Biserial correlations can be conceptualized as attempts to estimate or recapture value of correlations between continuously scored and normally distributed item constructs and total scores when the only data available are the Pearson or point-biserial correlations between the (artificially) dichotomized item scores and the total scores. The following SPSS commands (syntax) will compute biserial correlations from point-biserial correlations where 'diff' is the item difficulty:

```
COMPUTE p = diff .
COMPUTE q = 1-diff .
COMPUTE z = IDF.NORMAL(p,0,1) .
COMPUTE y = exp(-z*z/2)/sqrt(2*3.14559) .
COMPUTE spss_bis = ptbis*sqrt(p*q)/y .
EXECUTE .
```

The 'y' in the above is the ordinate or height of the standard normal curve at p, the item difficulty.

TAP works very well for constructed response item formats such as short-answer or completion as long as you score items right-wrong (no partial credit). Enter '1' for a correct item for each person and '2' for incorrect. The answer key is then simply a series of 1's.

Item Analysis: Part II

The following section of the output indicates precisely which options examinees selected for each item. We show a selection of items here:

```
Options Analysis
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Item Frequencies and Percentages (* indicates correct answer) -- page1
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Item Group       Option 1    Option 2    Option 3    Option 4    Option 5
---- ---------- ----------- ----------- ----------- ----------- -----------
  1  TOTAL        0 (0.000)   1 (0.050)   0 (0.000)   2 (0.100)  17*(0.850)
     High Group   0 (0.000)   1 (0.167)   0 (0.000)   0 (0.000)   5 (0.833)
      Low Group   0 (0.000)   0 (0.000)   0 (0.000)   0 (0.000)   5 (1.000)
     Difference   0 (0.000)   1 (0.167)   0 (0.000)   0 (0.000)   0(-0.167)

  2  TOTAL       18*(0.900)   1 (0.050)   1 (0.050)
     High Group   6 (1.000)   0 (0.000)   0 (0.000)
      Low Group   4 (0.800)   1 (0.200)   0 (0.000)
     Difference   2 (0.200)  -1(-0.200)   0 (0.000)

…

  4  TOTAL       15*(0.750)   5 (0.250)
     High Group   5 (0.833)   1 (0.167)
      Low Group   3 (0.600)   2 (0.400)
     Difference   2 (0.233)  -1(-0.233)

…

 13  TOTAL        0 (0.000)   0 (0.000)  20*(1.000)   0 (0.000)
     High Group   0 (0.000)   0 (0.000)   6 (1.000)   0 (0.000)
      Low Group   0 (0.000)   0 (0.000)   5 (1.000)   0 (0.000)
     Difference   0 (0.000)   0 (0.000)   1 (0.000)   0 (0.000)

 …

 30  TOTAL        3 (0.150)   2 (0.100)  14*(0.700)   1 (0.050)
     High Group   0 (0.000)   1 (0.167)   5 (0.833)   0 (0.000)
      Low Group   1 (0.200)   1 (0.200)   3 (0.600)   0 (0.000)
     Difference  -1(-0.200)   0(-0.033)   2 (0.233)   0 (0.000)
```

An item like #13 above was correctly answered by all examinees. We see that there are 6 persons in the High Group, 5 persons in the Low Group, and, since there are 20 persons total, 9 persons who are in neither the High nor Low Groups. Since the keyed response or correct answer is indicated with an '*', we can see that no one chose any of the three distractors (a, b, d) for this item. The Difference between the High and Low Groups is nil and, in this case, is a wee bit misleading since the value of '1' simply reflects the fact that the groups are of different sizes. We might be well advised to focus instead on the proportion difference in (parentheses) which is zero as we would expect.

Item #4 had only one option other than the correct one. We note that the Difference in High and Low Group proportions correct is positive for the correct response and negative for the distractor. This is a desirable outcome. Persons in the more knowledgeable High Group should get the item correct more often than persons in the less knowledgeable Low Group. Conversely, the incorrect option should be more attractive to the less knowledgeable. The magnitude of the difference is a measure of the item's ability to discriminate or distinguish among these more and less skilled groups of examinees. Larger (absolute) values are preferred. For example, we would conclude that Item_04 may discriminate somewhat better among these examinees than Item_13.

Item #1 may be a flawed item. The discrimination index is negative (-0.167) indicating that the more knowledgeable respondents favored another option to a greater extent than the less knowledgeable for this item. It would be informative at this point to ask for possible reasons for this item's apparent failure. You might hear or suggest that this was:

- a misleading/tricky item or poor item where, say, improper phrasing misled some students
- a teacher error and/or miskeying where the teacher either got the answer wrong or made a transcription error
- an instructional glitch where the teacher actually either incorrectly taught the concept or skill or, perhaps, emphasized some (minor?) aspect of the concept in such a way as to make some examinees feel the keyed response was not the best answer
- an anomaly in which the item is really fine, but where a few of the more knowledgeable students, perhaps by chance, slipped and, simultaneously, some few of the less knowledgeable students just got lucky.

The sort of *post mortem* required to determine the precise cause of failure of particular items can be very informative for both subsequent instruction and assessment activities. Students are often both willing and able to inform you why an item didn't work as intended. There are instructional benefits to such explorations with students as well.

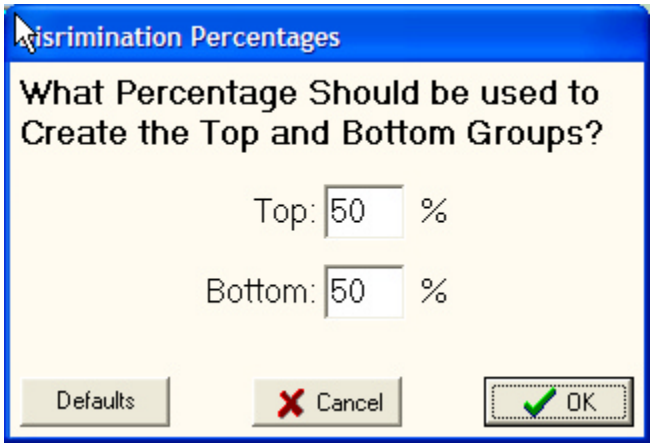
If you want, it is quite easy to use TAP to generate data with 'flawed' items for students to practice their skills at detection. Simply generate data as desired and then change an answer or two on the answer key while in the Data Editor. If you peek at the right section of the output beforehand, you can choose a distractor that worked well (that is, where the Difference proportion is much less than zero) for the new answer to make the item worse.

It is informative for students to see the impact of these poor items on the reliability of the test. For example, on the *Item and Test Analysis* section of the TAP output note that simply eliminating Item_01 would increase Alpha or KR-20 from 0.719 to 0.748 while eliminating items which discriminate positively will decrease reliability. That is, while Spearman-Brown implies that more items give greater reliability, this is only true when the items are of similar or higher quality (discrimination) than those items already composing the test. If you drop or improve poor items, you will improve the reliability of the test. If you add poor items or degrade good items, you will reduce the reliability of the test.

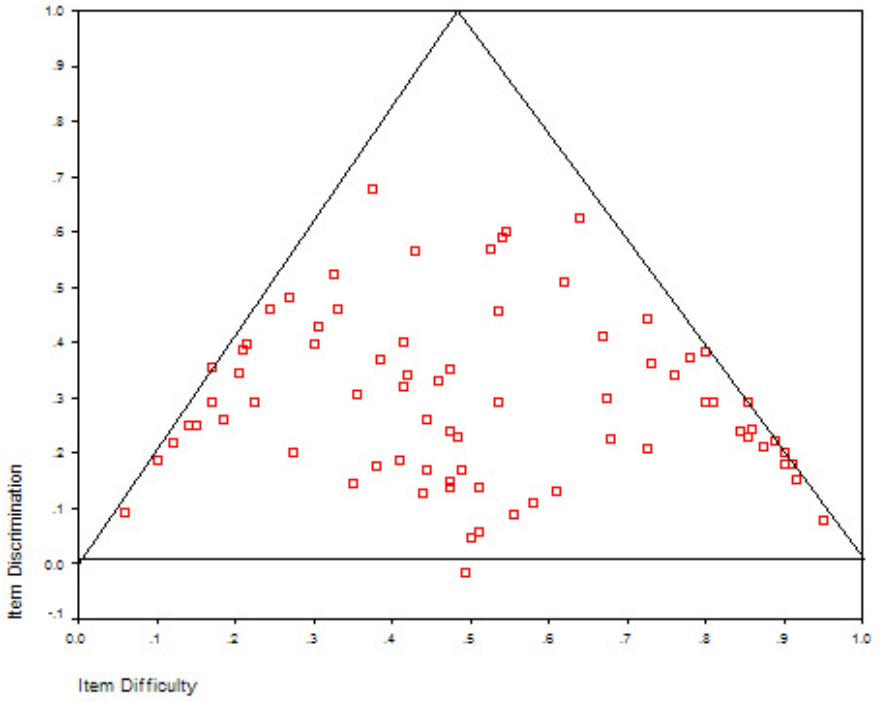
Comparisons of the Item Discrimination Indices

Portions of this section may have more detail than you feel you need or desire. Please feel free to read selectively.

The data used in this section, *InGuide.tap*, are included with the software. To replicate the information below, you will need to go to the *TAP* pull-down *Options* and choose 'Set Percentages for Item Discrimination'. Set these for a 50% high and 50% low group as indicated below:



You will see a warning that this leaves few persons in the middle group. This is OK. You will then need to *Analyze* these data. If you choose the *Save the Item Statistics* option from under the pull-down *File* option, you can easily bring the item statistics into, say, SPSS and create the following scatterplot:

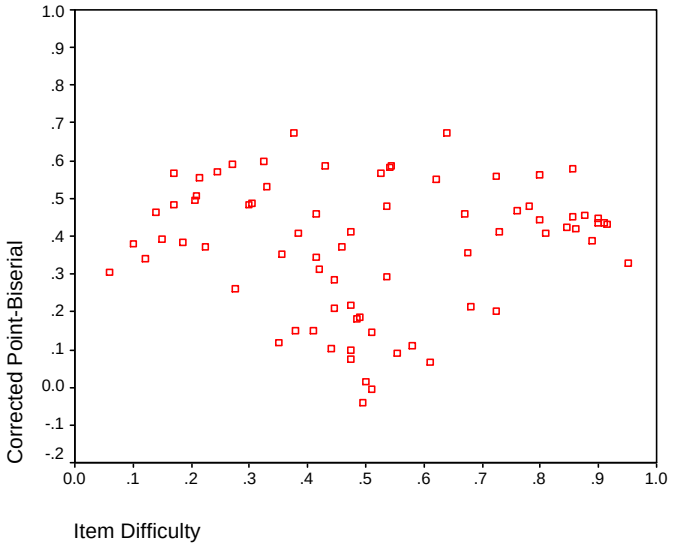


The key observation for students is that all of the points with positive discrimination are within the superimposed isosceles triangle. Similarly, all point with negative discriminations would be within a congruent inverted triangle. The boundaries of the triangle, in fact, represent maximum possible discriminations given the item difficulty. To help see this, have students compute maximum discrimination values when, say, there are 100 students taking an item and the item difficulty is 0.60. Logic would indicate that the most highly discriminating items will have all students in the higher group with the item correct (N=50) and all students in the lower group with the item incorrect (N=50). This is possible only when the item difficulty is 0.50 and not when the difficulty is 0.60. In the case of an item with difficulty 0.60, the maximum value for discrimination will occur when all in the higher have the item correct and only the minimum necessary number of students or examinees, N=10, have the item correct and are in the lower group. The item difficulty in the higher group would be 1.00 and 0.20 in the lower group in this best-of-all-possible cases. The discrimination index would, therefore, be 0.80 (1.00 – 0.20) at best. Note that there is necessarily no middle group for this illustration to work. That is, if you use the default setting with upper and lower 27% groups, then you will very likely see a similar plot, but there may well be points outside of the triangle(s) above.

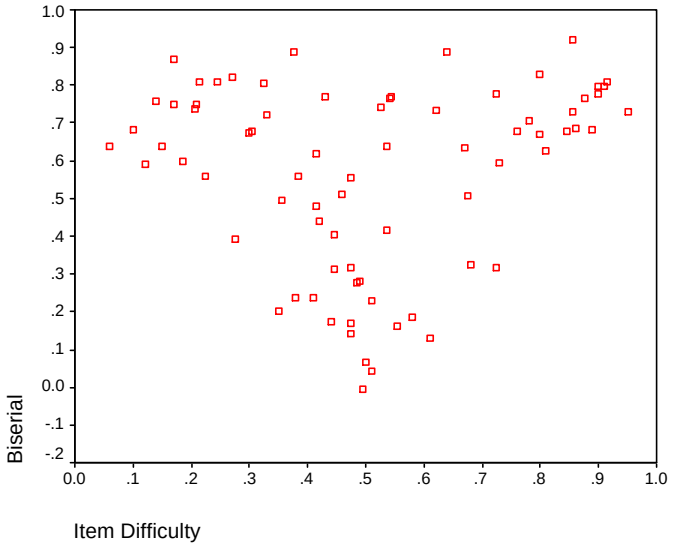
If this exercise is continued for other values of item difficulty, say, 0.20, it is easy to see the triangle 'appear'. The point of all of this is that item discrimination index (the high-minus-low group item difficulty index) is highly dependent upon, or at least constrained by, item difficulty. It is simply impossible for this index to be very large as items become quite easy or quite difficult.

Plot the corrected point-biserial item-total correlations (as below). You will see that these values are seemingly within an ellipse and are also restricted to have maximum values near 0.80. The

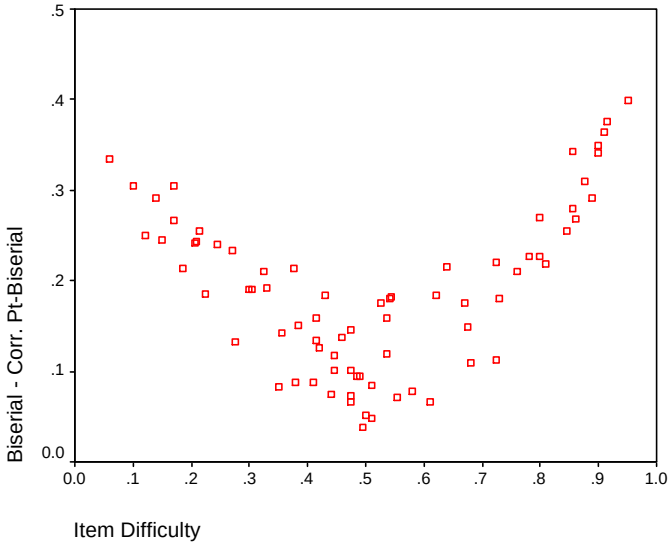
restricted nature at the ends of the item difficulty scale is somewhat less dramatic than with the discrimination indices, but present, nonetheless.



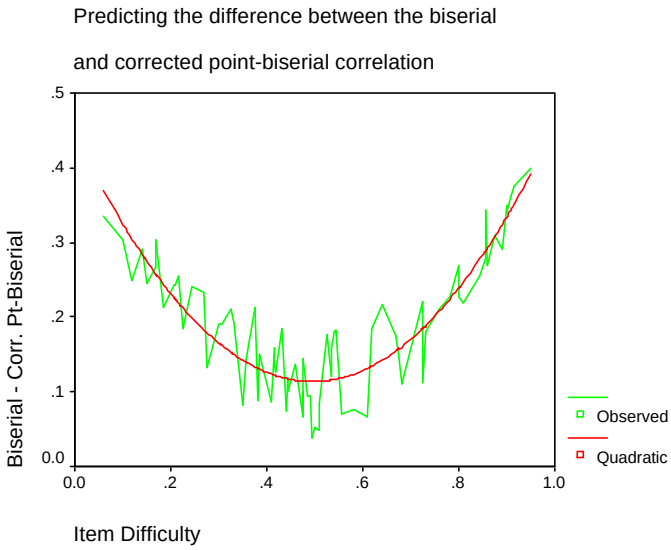
The plot for the biserial correlations seems to show virtually no relationship with item difficulty.



Perhaps the clearest way to see the difference between the biserial and point-biserial is to actually plot this difference verses the item difficulty.



If you fit a quadratic equation to this scatter using least squares, you get a very nice approximation. The multiple correlation is 0.90 ($R^2 = 0.81$) using both item difficulty and the square of item difficulty to predict the differences. The fitted quadratic is shown below:



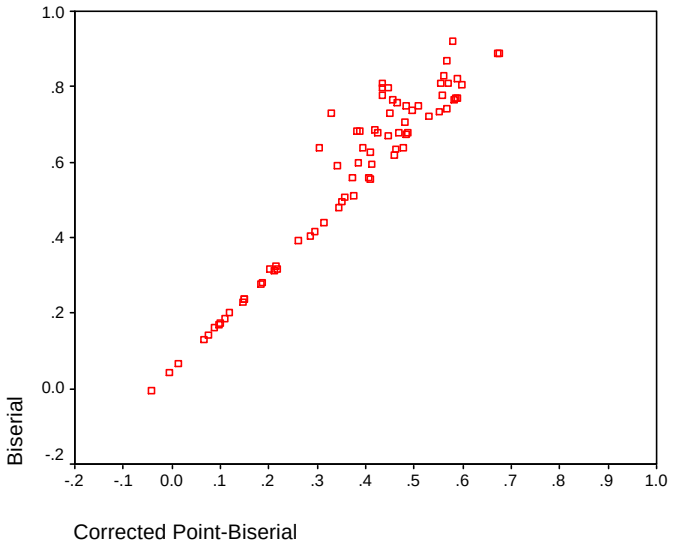
We have seen how these various discrimination indices relate to item difficulty, but an even more interesting question is: How do they relate to one another?

Correlations

		Point-Biserial	Biserial	Corrected Point-Biserial	Discrimination
Point-Biserial	Pearson Correlation	1	.956**	1.000**	.847**
	Sig. (2-tailed)	.	.000	.000	.000
	N	75	75	75	75
Biserial	Pearson Correlation	.956**	1	.963**	.665**
	Sig. (2-tailed)	.000	.	.000	.000
	N	75	75	75	75
Corrected Point-Biserial	Pearson Correlation	1.000**	.963**	1	.836**
	Sig. (2-tailed)	.000	.000	.	.000
	N	75	75	75	75
Discrimination	Pearson Correlation	.847**	.665**	.836**	1
	Sig. (2-tailed)	.000	.000	.000	.
	N	75	75	75	75

**. Correlation is significant at the 0.01 level (2-tailed).

The more accurate value for the correlation between the point-biserial and the corrected point-biserial is 0.9996778906938. Not too shabby! The uncorrected correlations, however, will always be larger than the corrected correlations (differences range in this example from 0.02 to 0.04 for the 75 items), but the linear relationship between them is very strong. The biserial and either the point-biserial or the corrected point-biserial are also strongly related (0.956 and 0.963, respectively). The plot in this case is informative as well.



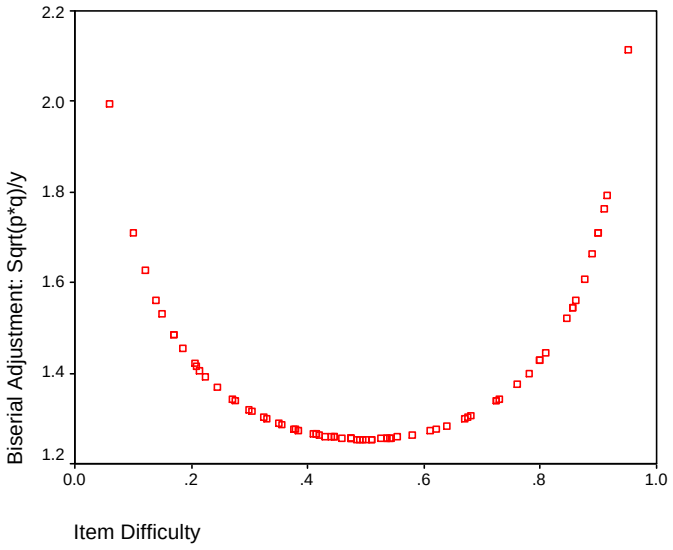
Note that for the smaller pairs of values the relationship is exceptionally linear, but that the biserial correlations can have greater values than might be expected or predicted from a linear relationship determined by the pairs of smaller values. In the table below we see that the biserial

correlations are greater than both the corrected and uncorrected point-biserial correlations. The differences between the corrected point-biserial and biserial correlations range from 0.04 to 0.40 for these items with a mean difference of about 0.19.

Descriptive Statistics

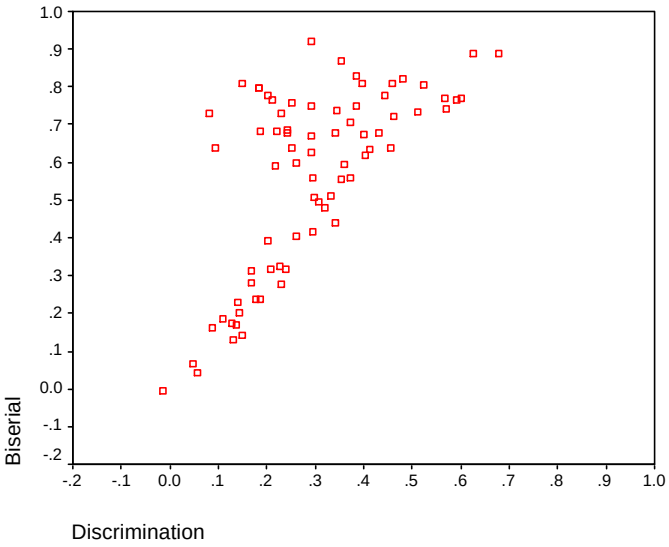
	N	Minimum	Maximum	Mean	Std. Deviation
Point-Biserial	75	.00	.70	.4005	.16914
Biserial	75	.00	.92	.5621	.24395
Corrected Point-Biserial	75	-.04	.68	.3724	.17471
Discrimination	75	-.02	.68	.2936	.14928
Valid N (listwise)	75				

To understand the nature of the relationship between the point-biserial and biserial, we return briefly to the definition of the biserial correlation (given earlier as SPSS syntax). The key is the adjustment to the point-biserial. We plot the values of this adjusting factor, sqrt(p*q)/y, versus item difficulty.



The minimum adjustment occurs for items with difficulty near 0.50 and increases for both more and less difficult items to compensate for the restrictions in the same region that we have seen imposed by the point-biserial correlations. For these items, the minimum adjustment is 1.25 and the maximum is 2.11.

The weakest linear relationship (0.665) is that between the discrimination index from high and low scoring groups and the biserial correlation. The plot is somewhat similar to the prior plot of biserial versus corrected point-biserial.



Three Suggestions for Use of the Various Indices

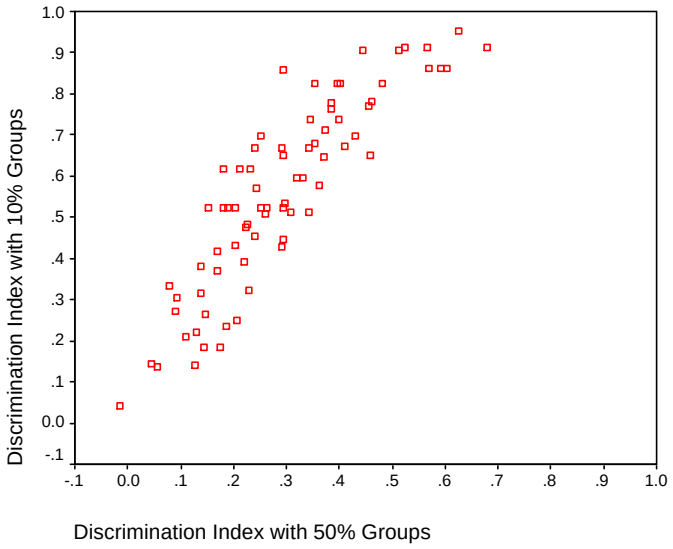
First, it would seem to make little difference whether you use point-biserial correlations or corrected point-biserial correlations for intra-item comparisons. However, simple point-biserial correlations are clearly inflated since you are computing a correlation between an item measure and a total score containing that item. This correlation would therefore be positive even if all of the other items were independent of the item under consideration. Many would seem to prefer the corrected point-biserial correlation for this reason. We concur and also would suggest the use of the corrected index over the uncorrected.

Second, you need to be aware that the discrimination indices will vary depending upon how we define the high and low groups. Generally speaking, this index will be greater when the differences are greater in the item difficulties for the high and low groups and this occurs when the groups are more 'selective' or formed with smaller percentages of the examinee group. For example, with our current data and upper and lower 10% groups we have:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
DISCRIM	75	.04	.95	.5598	.23083
Valid N (listwise)	75				

Recall that from the 50% high and 50% low groups we had a mean item discrimination value of 0.2936. Is only the mean discrimination impacted? That is, perhaps there a very strong linear association between these indices…so strong in fact that one index is essentially a linear transformation of the other? Nope. The correlation between the 50% discrimination indices and the 10% discrimination indices for these data is 0.881 and the following plot shows the scatter.



It is difficult to set guidelines for acceptable magnitudes for discrimination indices unless you know the definition of high and low groups. In any event, constructing an index where 80% of the examinees are excluded may create sample size (i.e., unstable estimate) problems for many data sets.

A further complication is the relationship of the discrimination index to the other indices when using more selective (say, 10%) groups. Compare this table to the previous table of associations among the discrimination indices.

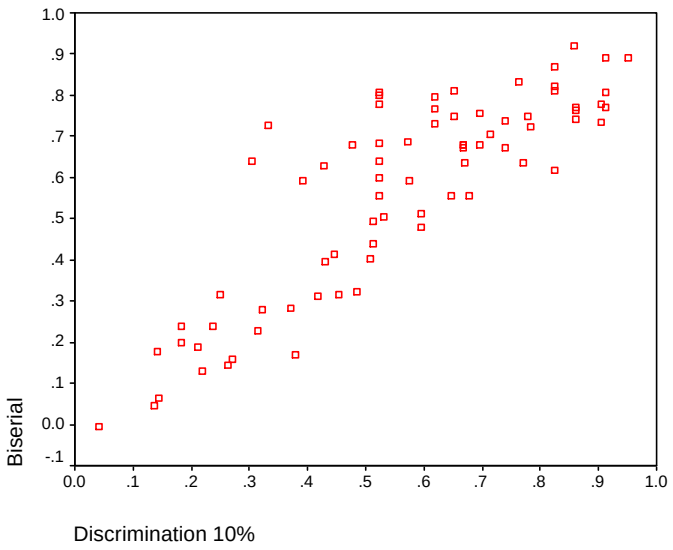
Correlations

		Point-Biserial	Biserial	Corrected Point-Biserial	Discrimination 10%
Point-Biserial	Pearson Correlation	1	.956**	1.000**	.951**
	Sig. (2-tailed)	.	.000	.000	.000
	N	75	75	75	75
Biserial	Pearson Correlation	.956**	1	.963**	.856**
	Sig. (2-tailed)	.000	.	.000	.000
	N	75	75	75	75
Corrected Point-Biserial	Pearson Correlation	1.000**	.963**	1	.946**
	Sig. (2-tailed)	.000	.000	.	.000
	N	75	75	75	75
Discrimination 10%	Pearson Correlation	.951**	.856**	.946**	1
	Sig. (2-tailed)	.000	.000	.000	.
	N	75	75	75	75

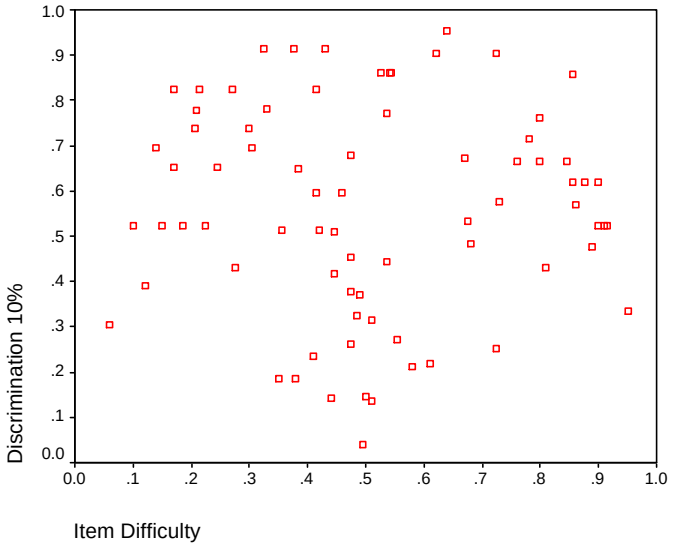
**. Correlation is significant at the 0.01 level (2-tailed).

The relationship of the discrimination index to the other indices is generally enhanced using more selectively formed high and low groups. The most modest relationship is still that between

the discrimination index and the biserial correlation (0.856), but it is greater than it was with the 50% groups (0.665) and less 'patterened'.



This is because the 'triangle' restrictions no longer are applicable. See the following scatterplot.



The discrimination index would seem to have some more desirable properties when more selective groups are employed. True, but then again these indices are less accurately estimated when you discard a large proportion of the participants. In typical classroom situations where sample sizes will be small, you may well feel you need to include everyone.

Because of the 'triangle' restrictions imposed on the 50% discrimination index we might be moved to use a more selective, say, 10% discrimination index, however, the small sample limitations of this index are discouraging. We are thus moved to our second suggestion of

recommending the use of corrected point-biserial over the different sorts of high-low group discrimination indices for most work.

We realize that this is also limited or restricted for item difficulties that are more extreme. If it is reasonable to theorize an underlying normal distribution for each item (infinitely parsed partial credit as opposed to right-wrong scoring), then the biserial would seem to be a good choice for use with items of more extreme difficulty. Item #9 gives us some insight into these choices. The difficulty of this item is 0.92. It is quite an easy item in comparison to the others. The corrected point-biserial correlation is 0.43. this item might seem to be comparable in discrimination to, say, item #16 where the difficulty is 0.42 and the corrected point-biserial is 0.46. That is, item #16 is quite a bit more difficult than item #9, but seems to discriminate ever so slightly better. The value of KR-20 would be the same if either item is deleted. If you consult the biserial correlations for these items you discover something a bit different. The biserial for item #9 is 0.81; the corresponding value for item #16 is quite a bit less at 0.62. Item #9 may be the more discriminating item after all if you are willing to accept the assumptions necessary for the use of the biserial correlation. At the very least, it is clear that using the point-biserial to evaluate the discriminating ability of item #9 paints a rather different picture than using the biserial. Using both indices is simply more informative than either index alone.

In this labored way we have arrived at our third suggestion: While it is generally fine (and perhaps even advisable) to use the corrected point-biserial as the measure of item discrimination, when items are either quite difficult or quite easy it is informative to also consult the biserial correlation.

The Grading Options

Under the *Options* pull-down you will find 'Set Percentages for Grades'. If you select this, you will be able to use percentage-correct standards for each of several grading options. First, choose the 5 category grading option (A, B, C, D, F) with the default percentages. Second, also under *Options*, select 'Use Letter Grades in Confidence Bands'. Third, create new data using the TAP defaults with the data generator (for this example, the seed was: 52255). If you view the output, you will see:

```
TITLE:   Generated Data:
COMMENT: Cases=20, Items=30, Seed=52255,
*************************************************************************
Examinee Analysis
*************************************************************************

                              Ltr    ~68% C.I.     ~68% C.I.    ~68% C.I.
Examinee        Score Percent Grade  (Raw Score)    (Percents)   (Letters)
--------------- ----- ------- ----- ------------ ------------ -----------
Person_001        27   90.00%   A    (24.9- 29.1)  (83.1- 96.9)   (B , A )
Person_002        28   93.33%   A    (25.9- 30.1)  (86.4-100.2)   (B , A )
Person_003        15   50.00%   F    (12.9- 17.1)  (43.1- 56.9)   (F , F )
Person_004        16   53.33%   F    (13.9- 18.1)  (46.4- 60.2)   (F , D )
…
Person_020        24   80.00%   B    (21.9- 26.1)  (73.1- 86.9)   (C , B )
=========================================================================
Percentage Grading Scale:  A = 90, B = 80, C = 70, D = 60
```

Notice that the first person would receive a letter grade of 'A' for this test since the standard (90%) for this grade was obtained. However, a ~68% confidence interval would include percentages less than 90%. Therefore, the letter grade confidence interval for this person would be (B, A). As was noted in the prior discussion on confidence intervals, we can interpret this to mean that the probability that the 'true' letter grade for an examinee is in this ~68% interval is 0.68. Or, if the student were retested (without fatigue or memory of the experience) again and again, then we might expect approximately 2/3 of the resulting letter grades to be 'A' or 'B'. The message is that while the person received a letter grade of 'A', our confidence in this grade is less than, say, in the letter grade of the third person where the ~68% confidence interval contains only a single grade (F, F).

For the 20 examinees in this example, only 5 have intervals containing a single letter grade. There are also intervals that contain three letter grades such as (C, A). Of course, ~95% intervals would be substantially larger.

If more grading confidence is desired, the reliability of the test can be increased which will decrease the standard error of measurement (SEM) and the size of the confidence interval. Also, the number of grading categories can be reduced. With just two categories (and using the default 70% passing score), the number of persons with intervals having a single letter grade is 13. Increasing the number of grading categories to 12 will result in only 4 intervals containing a single letter grade (F, in this case).

The following chart is produced:

```
Bar Chart for Letter Grades
***************************

Grade  Count  Graph (each @ represents 1 case)
-----  -----  --------------------------------

  A        6   @@@@@@
  B        4   @@@@
  C        2   @@
  D        3   @@@
  F        5   @@@@@
```

Notice that the distribution of grades is rather 'U' shaped for these data and grading standards. This may be surprising since the raw score distribution appears to be somewhat uniform and the stem-and-leaf plot would suggest a more 'mounded' or normal distribution.

Table of Specifications Option in the Data Editor

If you select the 'Set Table of Specifications' option from within the Data Editor, you will be able to see mean item difficulties for both the Content Areas and for the Cognitive Levels. That is, using the file InGuide2.tap, we see that there were 4 Content Areas (Addition, Subtraction, Multiplication, and Division) with 2 Cognitive Levels. The 40 items of this test are uniformly distributed across both the content and cognitive cells. For example, there are 10 questions on arithmetic, 5 at the knowledge level and 5 applications. Overall, there are 8 cells in the table of specifications with 5 items in each cell.

If you look at the analysis for these data, you see:

```
For best results, set font to COURIER NEW 10 Point
TITLE:   Generated Data:
COMMENT: Cases=100, Items=40, Seed=111282,
*************************************************************************
Examinee Analysis
*************************************************************************

                              Ltr    ~68% C.I.     ~95% C.I.
Examinee        Score Percent Grade  (Raw Score)   (Raw Score)
--------------- ----- ------- ----- ------------- -------------
Person_001        29   72.50%   C    (26.5- 31.5)  (23.9- 34.1)
Person_002        33   82.50%   B    (30.5- 35.5)  (27.9- 38.1)
…
Person_100        33   82.50%   B    (30.5- 35.5)  (27.9- 38.1)
=========================================================================

Mean Item Difficulty      = 0.727
```

```
*************************************************************************
Content Area Analysis
*************************************************************************

                      Average
Content Area          Difficulty   Items in Content Area
--------------------  ----------   -----------------------------
addition              0.701        1,2,9,10,17,18,25,26,33,34
subtraction           0.637        3,4,11,12,19,20,27,28,35,36
multiplication        0.742        5,6,13,14,21,22,29,30,37,38
division              0.828        7,8,15,16,23,24,31,32,39,40


                      Average
Cognitive Level       Difficulty   Items in Cognitive Level
--------------------  ----------   -----------------------------
Knowledge             0.754
1,3,5,7,9,11,13,15,17,19,21,23,25,27,29,31,33,35,37,39
Beyond Knowledge      0.700
2,4,6,8,10,12,14,16,18,20,22,24,26,28,30,32,34,36,38,40
```

The mean item difficulties across content areas inform us that the subtraction items were most difficult for this group of students, followed by addition and then multiplication. Division was the content area where the students were most successful. This may or may not be surprising, but it will likely be useful information for several instructional purposes. Similarly, the application items were just a bit more difficult for these students than were the more factual items. This also might or might not have been expected and/or desired. There may be implications for remediation. Or, there may be reason for celebration.