**Silvia E. Moore**

**EDRS 828**

**November 3, 2014**

**Q1 (2 pts.)** If we want to create a new temperature scale by converting Fahrenheit degrees (**Fo**) into new measures (say, denoted **No**), would it be appropriate to use the following transformation: **No =** 2**Fo** – 50? (e.g., if **Fo** = 70o, then **No** = 90o).

**Yes, F0 is reported using an interval scale that has an arbitrary zero that provides an actual distance between points (does not provide ratio information).**

**Transformation is possible with Interval scales if it maintains the interval scale (using operation to average the temperature is permissible).**

**Y = bX + a**

**70=2(70)-50**

**Q2 (1 pt)** Do *test-retest reliability* and *internal consistency reliability* represent the same

concept, so they are interchangeable?

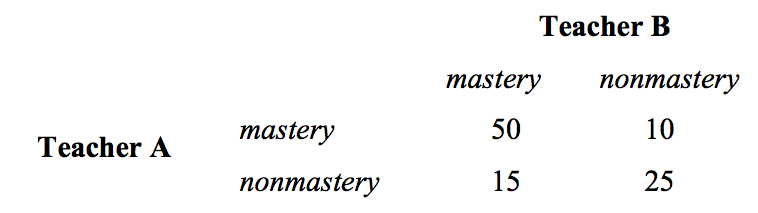
**No, test-retest and internal consistency reliability do not represent the same concept.**

**Test-retest reliability measures if people answer consistently over time on a questionnaire or survey instrument and is affected by errors of ‘carry-over effects’ and therefore is more appropriate for measuring traits that are stable over time.**

**Internal consistency reliability measures the variance correlation among items in a single administration of the instrument and is affected by correlation errors.**

**Q3 (3 pts)** Each of two teachers has classified each of 100 students into two categories (***mastery*** vs. ***nonmastery***) based on their portfolio evaluation of whether the student has demonstrated achievement of designated performance standards. Compute **Cohen’s kappa** to assess the agreement between the two teachers in this evaluation given the classification distribution of the students in the following table (e.g., both teachers classified 50 students into “mastery”).

**Consistent Classification is 38% after controlling for chance…Classification consistency is between .47 and .75.**

****

**. .60 70**

**.40 .65 .35**

**Criterion-referenced reliability for classification:**

**Pe = (60)(65) + (40)(35) = .53**

**P0 = .50 + .25 = .75**

**K=P0 – Pe / 1-Pe**

**K = .75-.53 / 1-.53 = .468**

**Q4 (3 pts.)** Two traits of middle school students, *motivation* and *persistence*, have been assessed by two scales – *Student Questionnaire Scale* (SQS) and *Teacher Opinion Scale* (TOS). If you participate in the validation of these scales, how would you collect convergent evidence and discriminant evidence about the external aspect of validity?

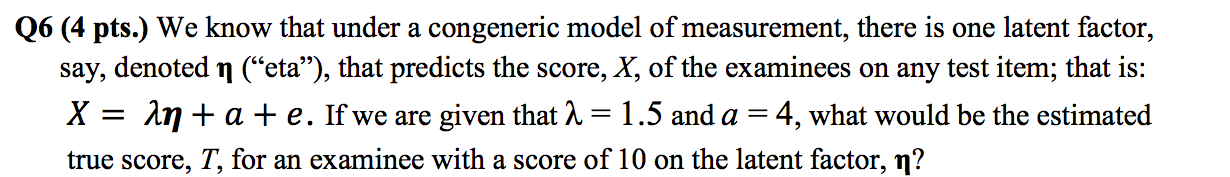
**First, I would have to determine if the instruments measure what they are suppose to measure by examining the scores of the SQS and TOS for high correlation between the SQS and TOS measures of the same construct. ( I assume both measures have been constructed according to a theoretical framework that provides convergent evidence.)**

**Discriminant evidence is examined by the low correlation between the constructs of motivation and persistence measured with the same instrument of different instruments, SQS and/or TOS.**

**Q5 (2 pt.)** If the correlation between true scores and observed scores is .80, what is the reliability of the observed scores?

**The reliability of observed scores is 65%. This is the degree to which the reliability indicates the scores are free of random error:**

**Pxx = (PTX)2 P.80 = (.80)2  = .64**



**True Score is 19**

**T = λη + a**

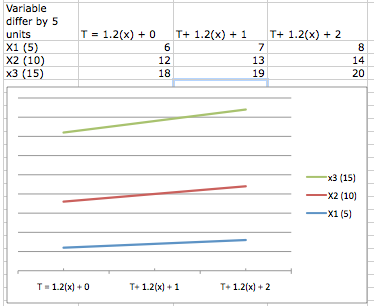
**T = 1.5(10) + 4 = 19**

**Q7 (3 pts)** Assume that a test consists of three components (items or test parts) that represent *essentially tau-equivalent* measures, *X*1, *X*1, and *X*3 (of course, under the congeneric model, say, **X = λη + a + e**). If it is given that λ = 1.2, any two examinees who differ by 5 units on the scale of the latent factor, **η**, will differ by 6 points in their true scores on any of the test measures *X*1, *X*1, and *X*3; (of course, we may have more than three essentially tau-equivalent measures).

**Using the rules of transformation for linear regression: Y= aX + b we prove that scores can be transformed but we must meet assumptions:**

* **The regression coefficient must be equal**
* **Must have tao equivalent of equal factor loading: cronbach’s alpha**

**The chart below shows that if The true scores for items on the latent factor differ by 5 units (X1  = 5 and X2  = 10 and X3  = 15), then their corresponding True scores will differ by 6 points( X1 Differs by 6 units from X2 when graphed using regression; likewise X2 differs from X3 by 6 points).**

****

Point

Diff.

X3  18

X3  12

X3  6

**Q8 (2 pts.)** Under the common factor model (i.e., congeneric model), say,

**X = λη + a + e**, the scale of the latent factor is established by which one(s) of the following conditions imposed to the model: **choices are highlighted**

A. The variance of the observed scores, *X*, is fixed equal to one;

B. Any of the factor loadings, λs, is fixed equal to one;

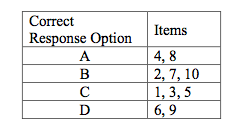
C. Thevarianceofanyerrorterm,*e*,isfixedequaltozero;

D. The variance of the latent factor, η, is fixed equal to one;

E. The variance of any error term, *e*, is fixed equal to one;

F. None of the above

**Q9 (5 pts.)** Conduct CTT item analysis using the computer program jMetrik with the data provided in the tab delimited text file **TESTDATA.txt** (attached) which contains the ID of 190 examinees and 10 multiple-choice items with four response options: A, B, C, D. For each item, one response option is correct and the remaining three options are distractors. The key for the correct response options is as follows:



**A JMerick analysis was conducted to examine the distribution of distracters of an instrument with 10 items. The instrument was administered to 190 students.**

**Results are as follows:**

* **Minimum = 0; this means that there was at least one examinee who could not answer any questions.**
* **Maximum = 10, which means there was at least one examinee who got all items correct.**
* **Mean = 4.9, while the median = 5. The difficulty of the test was at the average difficulty.**
* **Standard deviation = 2.173 and the quartile range = 3.25; this means that the middle 50% scores on the test fall between the 1st and 3rd quartile with a unit of 3.25.**
* **Skewness & kurtosis is not different from zero so we are not concerned about a deviation from normality.**
* **KR21 is reliability of .52**
* **Cronbach Alpha for internal consistency is 55%.. it will vary from .45 to .64.**
* **An examination of item deletion cronbach’s alpha indicates no reason to delete any items based on the argument that alpha will increase substantially if any of the items were deleted.**
* ***Analyzing items for deletion by examining distracters distribution:***

**Only item 1** seems to have distracters whose attractiveness to the responder is evenly distributed, and whose percentage seem to be acceptable 11% to 16%. This distracter has a difficulty leave of 57%, meaning that it is expected that 57% of responders will get the correct answer.

**Item 9** with a difficulty percentage of 47, is a problem as there is a distracter that has an attractiveness level of 41% (the correct answer is 42%). This distractor is chosen at a higher percentage than the other two distracters, A = 5.3%; B = 41% and C =12%.

**Items 2-7** all have distracters with uneven distributions.

* *Item difficulty (ease of item response)*

**Less Difficulty** items were 1, 3, 4, 5, & 7 and could be answered by 54%-63% of examinee

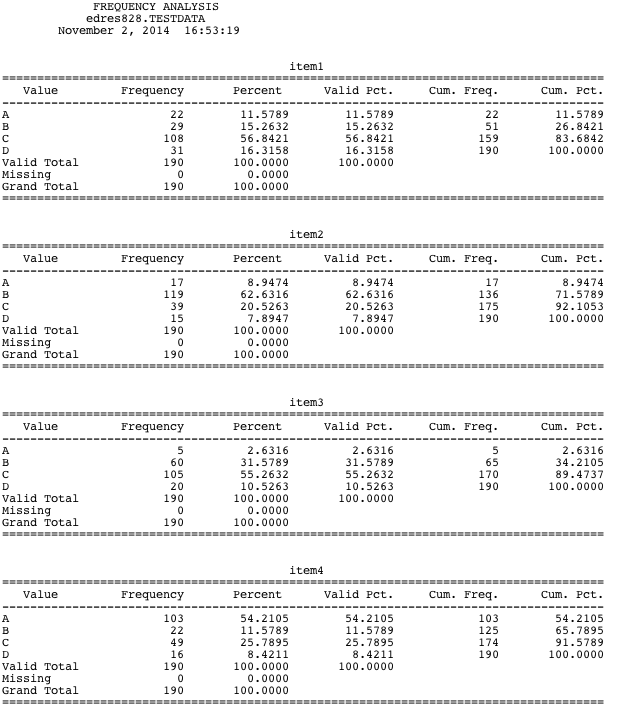
**Medium difficulty**  Items were 6&9: 42% of responders could answer

**Most difficult** Items were 8 & 10, These items could only be answered by 30-33% of responders.

* **Item discrimination** shows us how robust the question is at distinguishing between examinees from lower and upper performing groups. Group 2 items distinguish better then group 1 items:

Group #1 with items 1, 2, 5, 7 & 8 have a discriminant from .1271-.2059;

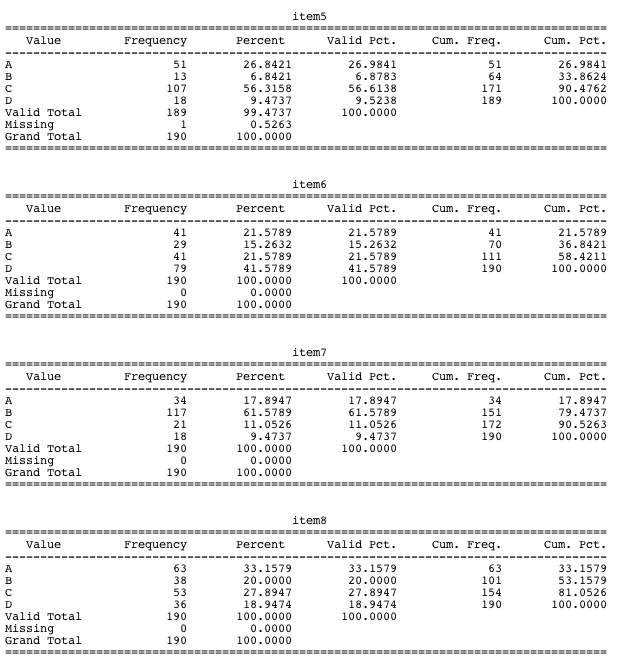
Group #2 with items 4, 6, 9 & 10 have a discrimant from .2858 to .3601.



**Correct response is C distracters are evenly distributed.**

**Correct response is B distracter C most attractive to examinees**

**Correct response is C distracters not evenly distributed (B)**

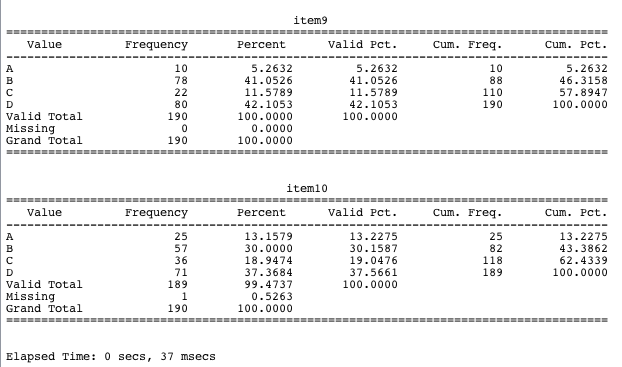
**Correct response is A, distracter C most attractive**

**Correct C, distracter A most attractive**

**Correct D, distracter even but may be too close to correct response**

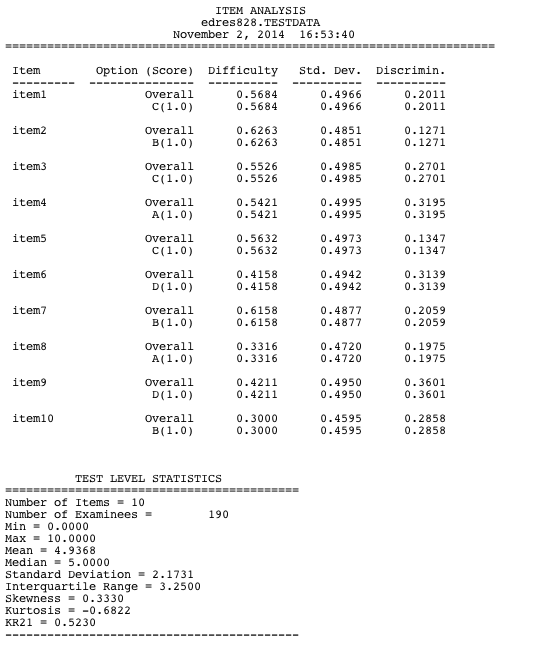
**Correct B, distracter A most attractive**

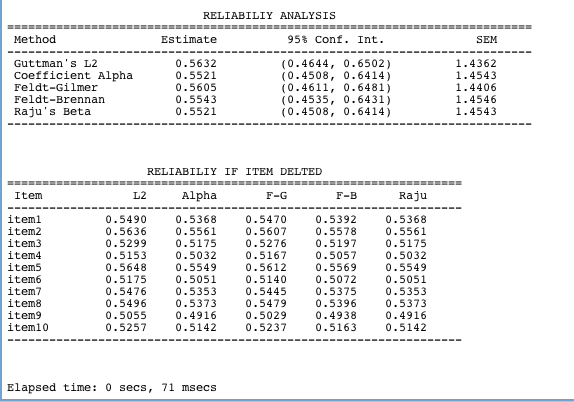
**Correct A, distracters too close to the response**

****

**Correct D, distracter B is *definitely too close* to correct response**

**Correct B, YIKES! Distracter D is the more attractive to responders than the correct answer**





POINT DIFFERENCE

X3 18

X2 12

X1 6