## Introduction to the online scan of “Quantitative Aspects of Psychological Assessment”

I wrote this book and had it published in 1972 by Duckworth in London, and Barnes and Noble in New York. In its lifetime it has been used as a text in some UK, North American and Australian post-graduate Clinical Psychology courses.

There have, of course, been tremendous changes in psychological statistics since that date, but fewer than one might expect in elementary psychometric theory.

Nevertheless the book is being up-dated and a revised edition should be available shortly.

Much of the new material to be added is available elsewhere on this site, and it will therefore be useful to alert readers to the Modules containing new material relating to the chapters of the book.

**Chapters 5 and 6. Correlation and Regression**:

The two Modules on this site which contain some material not in the book are **Correlation** and **Correlation 2**.

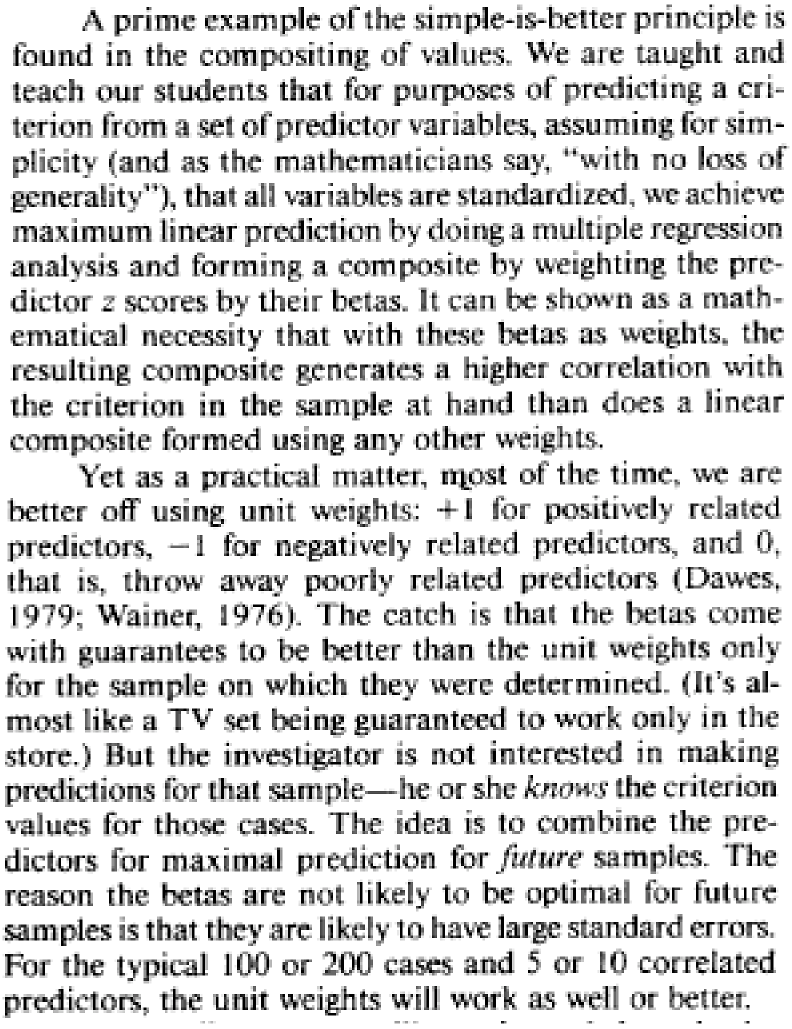
In these modules you should find material on the different sorts of correlation coefficient, e.g. **point-biserial, biserial, Phi** and a little more on **Eta** than is mentioned in the book.

I have also included a discussion of the **Binomial Effect Size Display**.

**Chapter 8 Multiple Regression and Prediction**

The use of computers has revolutionised what can be done in the area of multivariate prediction. Nevertheless it is still possible to argue that in the assessment of individuals most of the multiple regression equations we are likely to use are fairly simple ones.

And, I am mindful of the fact that Jacob Cohen, one of the great contributors to this field, is on record as saying (in a paper well worth reading):



> A prime example of the simple-is-better principle is found in the compositing of values. We are taught and teach our students that for purposes of predicting a criterion from a set of predictor variables, assuming for simplicity (and as the mathematicians say, "with no loss of generality"), that all variables are standardized, we achieve maximum linear prediction by doing a multiple regression analysis and forming a composite by weighting the predictor *z* scores by their betas. It can be shown as a mathematical necessity that with these betas as weights, the resulting composite generates a higher correlation with the criterion in the sample at hand than does a linear composite formed using any other weights.
>
> Yet as a practical matter, most of the time, we are better off using unit weights: +1 for positively related predictors, −1 for negatively related predictors, and 0, that is, throw away poorly related predictors (Dawes, 1979; Wainer, 1976). The catch is that the betas come with guarantees to be better than the unit weights only for the sample on which they were determined. (It's almost like a TV set being guaranteed to work only in the store.) But the investigator is not interested in making predictions for that sample—he or she *knows* the criterion values for those cases. The idea is to combine the predictors for maximal prediction for *future* samples. The reason the betas are not likely to be optimal for future samples is that they are likely to have large standard errors. For the typical 100 or 200 cases and 5 or 10 correlated predictors, the unit weights will work as well or better.

In other words, in many situations, weighting the Z scores on variables we know to be correlated with the criterion by +1 if the correlation is positive; -1 if the correlation is negative; and 0 if the correlation is insignificant (i.e., ignoring); we can very often obtain a simple composite variable which correlates virtually as highly with the criterion as a more complicated weighted-composite variable.

The paper, which you should read if you can get hold of a copy, is:

> Cohen, J (1990) What I have learned so far. ***American Psychologist***, 1990, ***45*** (12), 1304 -1312

The two references cited in the selection above are:

> Dawes, R. M. (1979) The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571 - 582
>
> Wainer, H. (1976) Estimating coefficients in linear models. It don't make no nevermind. *Psychological Bulletin, 83*, 213 – 217

**Chapter 9  Composite Scores**

This is a little changed but in the Module I have added colour to simplified tables in an effort to aid comprehension.

I have also added a new, but perhaps important, section which deals with viewing difference scores as composites, in the context of the problem of the abnormality of a difference between two test scores.

At least it provides a different way of looking at this problem.

**Chapter 11 Reliability**

The main update material here is to be found in the Reliability Module.

I have included a much lengthier discussion of the **standard error measurement** and the possible ways of using it.

The references which will give you an idea of the standard error of measurement debate are:

Lord, F. M. and Novik, M. R. (1968) *Statistical theories of mental test scores*. Menlo Park, California: Addison Wesley

Stanley, J. C. (1971), Reliability. In Thorndike, R. L. (ed) *Educational measurement*. (Second Edition, pp. 356 – 442), Washington, DC, American Council on Education.

Nunally, J. C. (1978) *Psychometric theory*. (Second edition) New York, McGraw-Hill

Dudek, F. (1979) The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin*, **86**, 335 – 337

Glutting, J. L., McDermot, P. A., and Stanley, J. C. (1987) Resolving differences among methods of establishing confidence limits for test scores. *Educational and Psychological Measurement*, **47**, 607 – 614

Charter, R. A. and Feldt, L. S. (2001) Confidence intervals for true scores: Is there a correct approach? *Journal of Psychoeducational Assessment*, **19**, 350 – 364

There is also more, than in the book, on **Coefficient Alpha**.

## Chapter 12 Validity

One of the main changes required in this chapter is some discussion of ecological validity, or rather the ecological validity of psychological tests.

Ecological validity refers to how well the test or tests predict real life behaviour. For example, suppose that you have administered a series of tests of memory and other functions and that all the tests have been established a valid measures of dementia.

You notice that the person you have assessed has not only shown clear evidence of dementia on your test, but has indeed done very badly on them even for somebody with dementia.

It would be tempting to conclude from this that the testy results imply very deteriorated performance in real life tasks. And psychologists have often made this mental jump. (and not only in the case of dementia).

But, as some of you will have realised, we are dealing here with a concurrent validity problem. The tests correlate with dementia, and dementia correlates with deteriorated everyday life behaviours, therefore the worse the test performance, the worse should be the real life behaviours.

Stretching our minds back to the How to … B (or looking it up) and supposing that our test results correlate .70 with a diagnosis of dementia and that a diagnosis of dementia correlates about .70 with deteriorated behaviour, what is the range within which we expect the correlation between the test results and deteriorated behaviour to lie?

The answer is somewhere between a correlation of zero and a correlation of 1.

If this sounds a bit far fetched to you, have a look at the growing literature on the ecological validity of tests of executive function and the like.

For example found that, in a group of people with Alzheimer's Disease, correlations between test scores and composites based on them had a median correlations with the two measures of 'real life' performance of .46 (range .35 - .57) with one, and with the other a median correlation of .55 (range .32 - .68).

> (Farias, S. T., Harrell, E., Neumann, C. and Houtz, A. (2003) *Archives of clinical Neuropsychology,18*. 655 - 672)

Similarly, a study investigating the relationship between tests used in the assessment of executive functioning bore very little relationship to real life measures of performance. Tests used were the Stroop Colour-Word score, the COWAT, perseverative errors from the WCST, and time on TMT (B) . the multiple correlation coefficient for this battery of tests was .45 (n.s.) against one everyday performance measure of relevant activities, and .42 (n.s.) for a different measure of very day performance.

> (Chaytor, N., Schmitter-Edgecombe, M., and Burr, R. (2006) Improving the ecological validity of executive functioning assessment. *Archives of clinical Neuropsychology, 21,* 217 – 227)

The error in predicting real life performance would therefore be high. Suppose we had someone who scored two standard deviations below the mean on this combination of executive functioning tests, our best bet would be that their performance on the real life measures would be about 0.9 standard deviations below the mean. However, of all of those who scored 2 standard deviations below the mean on the executive function tests, about 16 percent would show real life performance at or above average level on tasks supposedly involving executive functioning ability.

So you can see that this is an important problem.

## Chapter 13 The Assessment of individual results.

The main changes here are to be found in the **Module Differences**.

I think that Frank Grubbs Test for an outlier might be useful in some clinical situations. A good reference for this and some related tests is:

Grubbs, F. E. (1969) Procedures for detecting outlying observations in samples. *Technometrics*, **11**, 1 –21

There is also some explanation of some of the formulas of John Crawford and his associates. These formulas are useful in situations where normative and standardisation groups are small.

John Crawford has a well-stocked site at:

http://www.abdn.ac.uk/~psy086/dept/index.html

The site allows download of reprints and .exe program files. (Sorry Mac users)

The key references cited in the differences module are:

Crawford, J. R., Howell, D. C., and Garthwaite, P. H. (1998) Payne and Jones revisited: Estimating the abnormality of test score differences using a modified paired sample t-test. *Journal of clinical and experimental Neuropsychology*, **20**, 898-905

Crawford, J. R. and Howell, D. C. (1998) Regression equations in clinical neuropsychology: An evaluation of statistical methods for comparing predicted and obtained scores. *Journal of clinical and experimental Neuropsychology*, 20, 755-762

However, Crawford and his associates are a very productive team, so visit the site and look for updates from time to time. I suspect that any references to the work of Crawford and his colleagues will be quite quickly outdated by their development of a more refined technique.

## Chapter 14 Classification

I have further developed the discussion of Bayes Theorem in the Classification Module, and, I think, I have derived an original method for estimating local base rates. (Apologies to the true originator if I am wrong).

This method of establishing local base rates should make base rate considerations much more manageable.

In the Classification Module, I have also included some discussion of the use of base rate information in the selection of cut-off scores.

I am not sure that selection ratios have proved as important in clinical practice as I thought they might be, but I have to confess to a an important omission from the text of the book if they are.

This is that I was unaware of a paper which provides Taylor-Russell type tables for the case of the dichotomous criterion.

As most clinical classification problems essentially involve a dichotomous criterion this was a bad lapse.

However, the tables can be found in:

Abrahams, N. M., Alf, E. F. and Wolfe, J. H. (1971)
Taylor-Russell tables for dichotomous criterion variables.
*Journal of Applied Psychology, 55*, (5), 449 - 457

Philip Ley, February 2007