Proceedings of the Institute of Acoustics

PROPOSAL FOR A REVISED SUBJECTIVE SPEECH INTELLIGIBILITY TESTING METHOD

P.W. Barnett AMS Acoustics, Rayleigh House, Bush Hill Park, EN1 2QB.

1. INTRODUCTION

Speech intelligibility measurement by subjective means, whilst undeniably the closest and most representative of the quantity under investigation, is not without its difficulties.

The simplest form of subjective testing - Word Scores - whether from open or closed sets, has associated with it the following:

- Lengthy
- Difficult to control
- Subject to variation

and it naturally follows therefore that it can be expensive. Furthermore a ready extrapolation is that in view of the factors outlined above, subjective methods are low on the list as a means to measure or quantify speech intelligibility.

It may be therefore that if some of the problems are overcome, then subjective methods with their obvious advantages might experience a small revival.

Word Score measurement has formed part of AMS Acoustics main activity over the past 10 years and our thoughts on a revised method result from both our experience and on anomalies of the results obtained over that period of time.

2. WORD SCORE TESTING

The subject of Word Score testing has been well and adequately covered at this Conference* and for the sake of brevity the details of subjective testing will not be discussed here.

Suffice it to say that a reader voices a list of Words (typically 50) or syllables and a listener writes down what he or she heard.

The list of words or syllables are generally Phonetically-Balanced, that is each test contains the same (approximate) distribution of phonetics, thereby rendering them of equal difficulty and hence interchangeable.

From our observations we have noted that for a certain condition there is not an equal probability that a word will be misheard.

This situation, which does give some cause for concern, forms the basis of our present thoughts on a revised method.

*H.M. Goddard - Subjective Intelligibility Testing in Practice

It is worth giving additional consideration to the probability of the mishearing issue.

Firstly we need to understand why a word is misheard and we can divide this into two different categories.

Category 1 - Those influences which are associated with the communication system under test.

Category 2 - Those influences which are not associated with the communication system under test.

Category 1 is, of course, the measurement at hand. It is the process under audit or measure. We can appreciate this process by considering the situation where a communication path is initially set at 100% i.e. all words are correctly heard. If then the communication means is gradually degraded in infinitesimal steps the measurement method would return some representative relationship and indeed we would not expect any quantisation of the results beyond the resolution of the measurement system.

Category 2 influences which cause mishearing might be quite different and would include talker-listener proficiency and misunderstanding which results from other influences outside of category 1.

With the exception of talker-listener proficiency, we can think of these other influences as statistical noise.

Insofar as is possible it would be a primary measurement endeavour to reduce the effects of category 2 influences and to an extent this is accomplished by training and practice for both the talker and listener.

3. RESULTS OF OBSERVATIONS

Graph 1 shows the results of scores obtained for List 11 PB Words out of 1000 population. The percentage scored correctly for each word is derived from:

$$\%C_t = \frac{\sum_1^n \left[\sum_1^x \left(C_{j1} + C_{j2} \ldots C_{jx} \right)_1 + \sum_1^x \left(C_{j1} + C_{j2} \ldots C_{jx} \right)_2 + \sum_1^X \left(C_{j1} + C_{j2} \ldots C_{jx} \right)_X \right]}{\sum_1^n \left[x_1 P_1 + x_2 P_2 \ldots . x_n P_n \right]}$$

where: C_t = Total percentage of tests 1 to n
C_{jx} = Number scored correctly by juror
x_n = Number of jurors per test
P_n = Number of words in test
n = Number of tests in sequence.

It can be seen from Graph 1 that the difference between words is considerable.

Hence we are able to deduce that for the test conditions under review, individual words do not have an equal probability of being scored either correctly or incorrectly.

To emphasise the point, Graph 1 is represented as Graph 2 rank ordered.

Hence we can infer that some words are more robust than others and indeed some are more vulnerable than others. Graphs 3 to 6 show the results for robust and vulnerable words against the actual average recorded score.

The effect is even more pronounced in the case of CVC nonsense syllables (see Graphs 7 and 8).

The difference clearly results from the recognition aspects in words which is less prevalent in CVC nonsense syllables.

4. DISCUSSION

Clearly therefore we are able to predict within reasonable certainty that for a given test condition, a word will be heard either correctly or incorrectly.

The simplest case to consider is a word that under the test conditions will be heard correctly.

The inclusion of such a word involves the following:

(i) Increased time to read the word.
(ii) Increased time to score the word (for live tests (i) and (ii) are the same).
(iii) Increased probability that a category 2 influence effect will affect the results.
(iv) Reduce the useful life of a listener**
(v) Increased costs associated with (i) through (iv) above.

In its simplest form therefore it makes some sense to omit those words from the list of 50 that will, with reasonable confidence, be scored correctly.

The same argument can be applied therefore to those words whose vulnerability ensures that they will be scored incorrectly.

The inclusion of either a certain tick (✓) or cross (x) on the answer sheet merely introduces an unnecessary uncertainty into the results. Furthermore, those who have experience of marking Word Scores are well aware of the frustration of the processing time associated with either always correct or always incorrect answers.

5. THE EFFECT OF THE SPACE

Our experience is based on some 200 individual spaces (and up to 20 x positions within the space) whose size ranges from 150 to 600m^3 and RT in the range 0.5 to 3.5 secs.

We rejected spaces with volumes greater than 1000m^3 since we know that the masking mechanisms of these spaces differ significantly from those with a mean free path of less than 6m.

Successive syllable masking and word masking are quite different in mechanism and effect.

**Listeners have a restricted life - see ISO TR 4870

Our proposals in this Paper relate only to successive syllable masking i.e. when the mfp is less than 6m.

In addition, our experience encompasses signal-to-noise ratios in the range 0dB to >30dB hence we would not recommend or suggest that the method to be proposed is used if the S/N is less than unity.

Accordingly we advise that the method proposed is not used outside of these limits:

S/N <0dB.
mfp > 6m
D/R <5dB $\geq$-20dB.

As an aside, we believe that the risk involved with values outside of the D/R limits is minimal.

6. PROPOSAL

At present our proposal for a shortened method deals with intelligibility in the range 60% to 100% and is based on reducing the normal 50 words per list to 20 words.

The 20 words are chosen from the list of 50 as those words which are the most vulnerable i.e. those words which are likely to be scored incorrectly even when the acoustic or noise degradation is small.

The format of the equation is as follows:

$$\%C = 2x + 60 \qquad \text{[i]}$$

where x = number scored correctly out of 20. Note the remaining 30 words are assumed to be scored correctly.

Since the cut-off between the 20 words chosen as vulnerable and the 30 words chosen as robust will be poorly defined a small correction needs to be applied and our revised formula takes the form:

$$\%C = 2x + 60 - \frac{(100 - C)}{y} \qquad \text{[ii]}$$

where: y is a factor derived empirically.

Equation [ii] reduces to:

$$\%C = \frac{2xy + 60y - 100}{y - 1} \qquad \text{[iii]}$$

We have deduced by empirical means, from around 20,000 individual scores of 50 words, that y is 6 for PB Word Scores and 3 for CVC nonsense syllables.

Hence: $$\%C_{PB} = \frac{12x}{5} + 52$$

$$\%_{cvc} = 3x + 40$$

where: x = score out of 20.

7. VALIDATION

From data collection by AMS Acoustics we have deduced that the error is small (see Graph 9) and that this error is generally negative i.e. the statistical version provides a slightly better result.

From Graph 10 it can be seen that the error is less than +/-5% in the range 70% to 100%.

It may be possible to slightly reduce the error in the vicinity 60% to 70% by further consideration of the factor $\frac{(100 - C)}{y}$ which in our model is simple.

8. CONCLUSIONS

From the foregoing I believe it has been demonstrated that a shortened version of Word Score testing does not significantly reduce accuracy for results in the range 70% to 100%.

Further, since the statistical noise is clearly reduced, the actual results obtained may be proven to be more accurate. From our observations the statistical noise associated with the measurements is in the range 3% - 4% with a standard deviation of 3% which is of the same order as the apparent error.

Finally, the following conclusions have been reached:

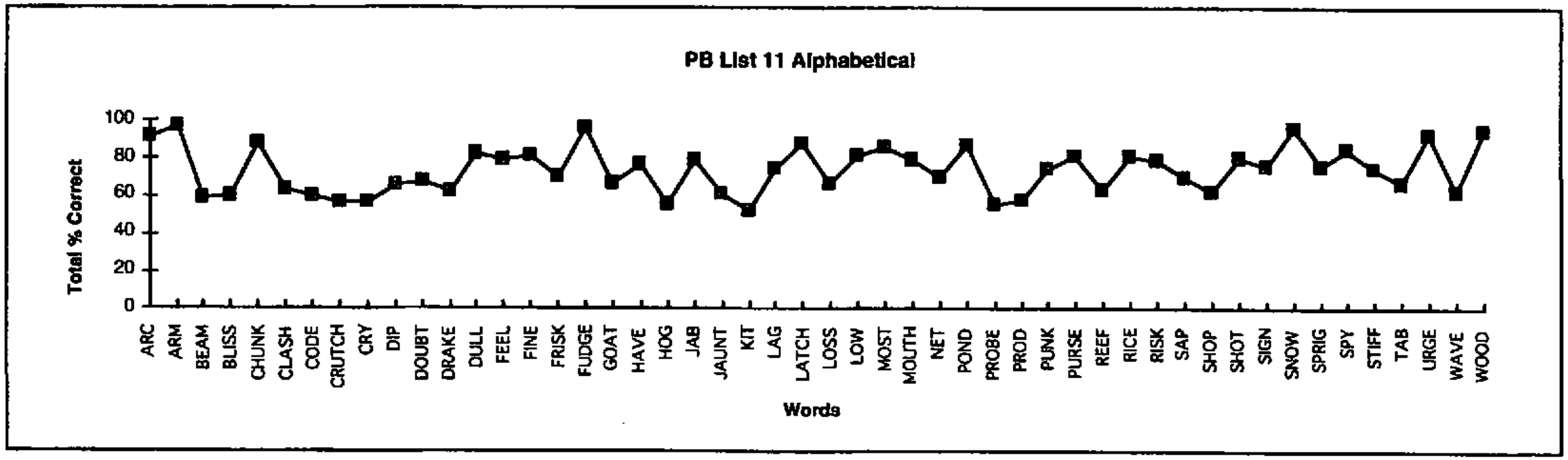
1. The truncated version has merit since it does not significantly reduce accuracy.
2. The truncated version will reduce statistical noise which again must have merit.
3. The truncated version is quick and hence less expensive to carry out.

Proposals for a Revised Subjective Testing Method - P.W. Barnett

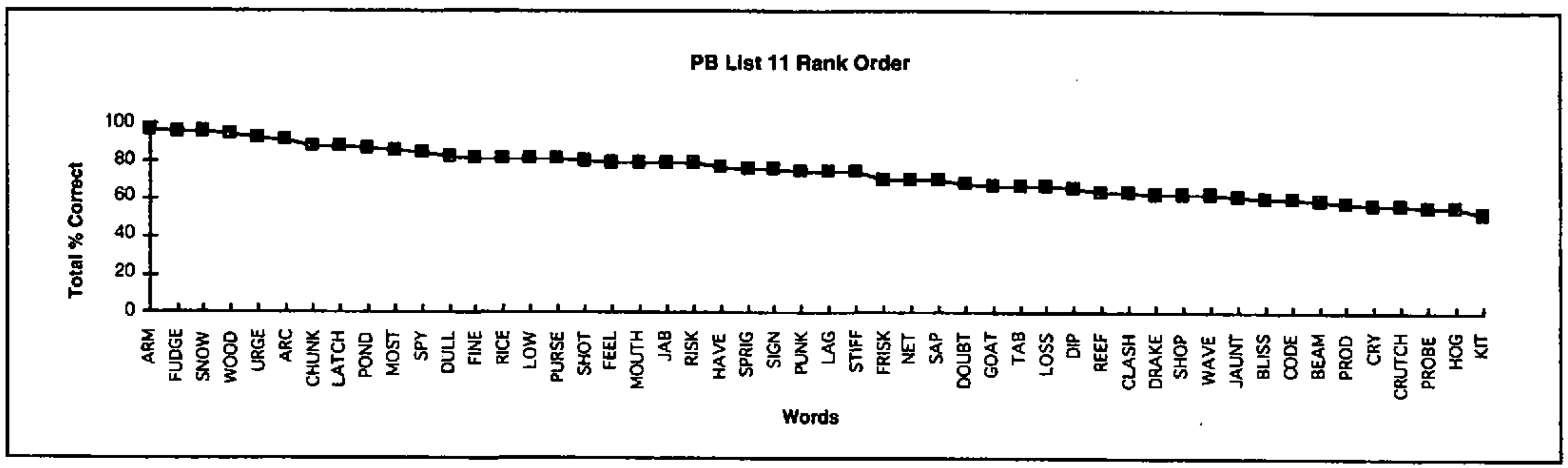
The list below shows the truncated lists for PB Word Lists 1, 2, 3, 5 and 11. it is our intention to publish the complete list after further analysis.

Truncated Word List

PB LIST 1	PB LIST 2	PB LIST 3	PB LIST 5	PB LIST 11
then	moose	rate	gape	fudge
wheat	rap	sit	inch	snow
hive	bean	dig	sick	wood
death	job	jam	nose	urge
hunt	scythe	bead	roe	arc
hid	rib	please	greek	chunk
creed	hit	sped	shove	latch
fraud	corpse	leave	browse	pond
ford	dab	wharf	love	most
bask	bought	barb	feed	spy
rise	pick	sob	punt	dull
cleanse	ways	crave	scare	fine
grove	need	cape	lend	rice
nook	hock	fame	rind	low
heap	hire	who	thick	purse
pan	frog	path	flap	shot
rub	bud	hurl	puff	feel
cane	mute	vow	kid	mouth
clove	pit	bald	bathe	jab
plush	quart	why	wrath	risk

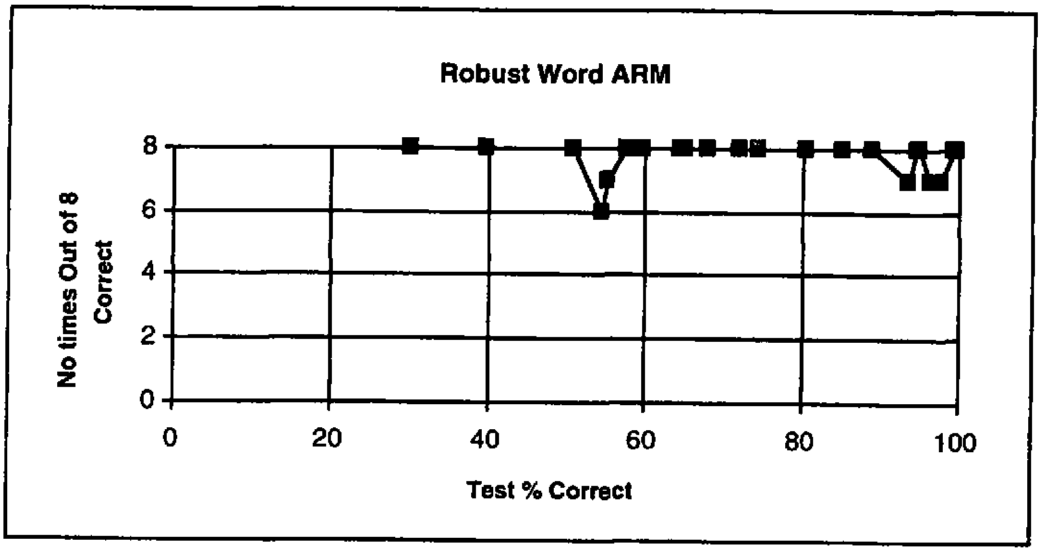


Graph 1

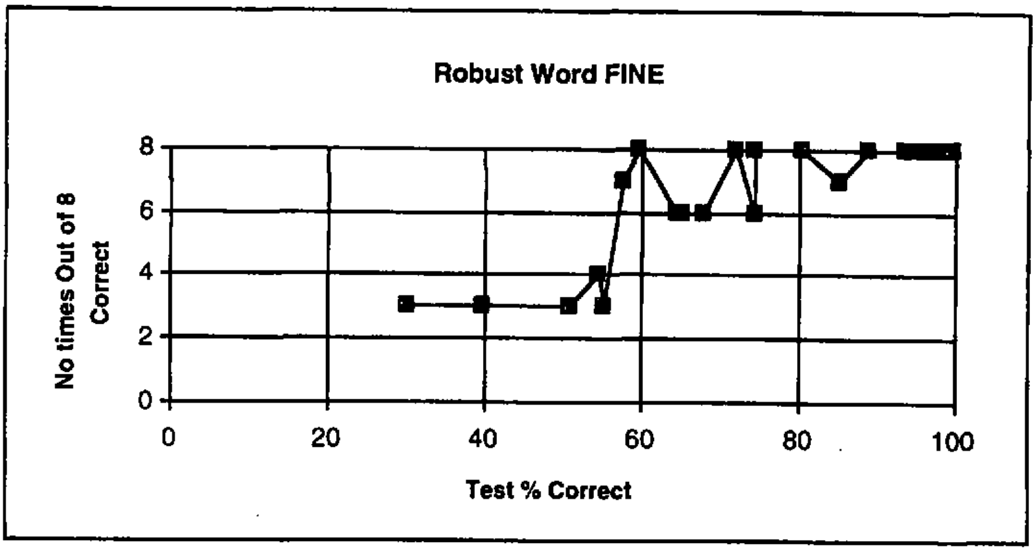


Graph 2

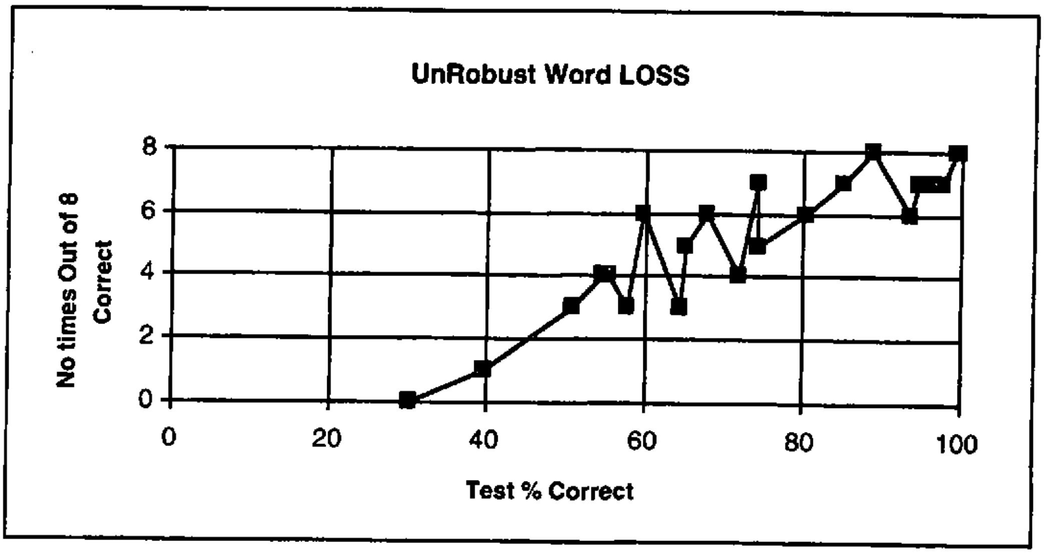
Proposal for a Revised Subjective Testing Method - P.W. Barnett



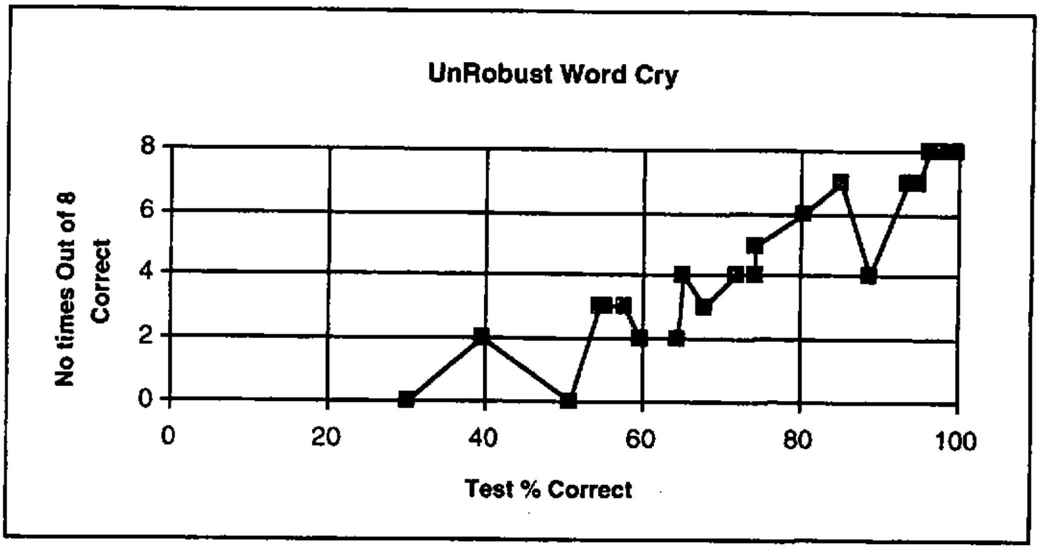
Graph 3



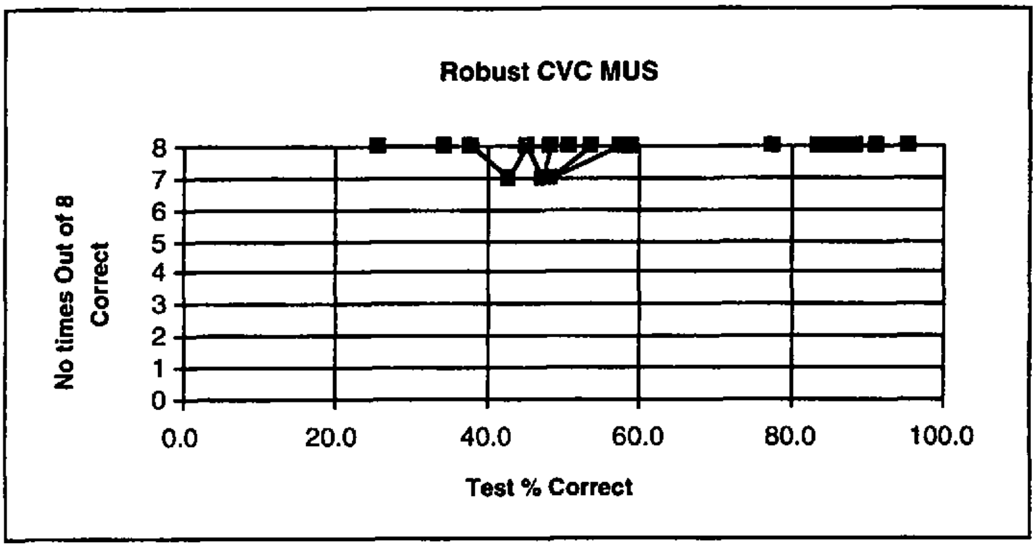
Graph 4



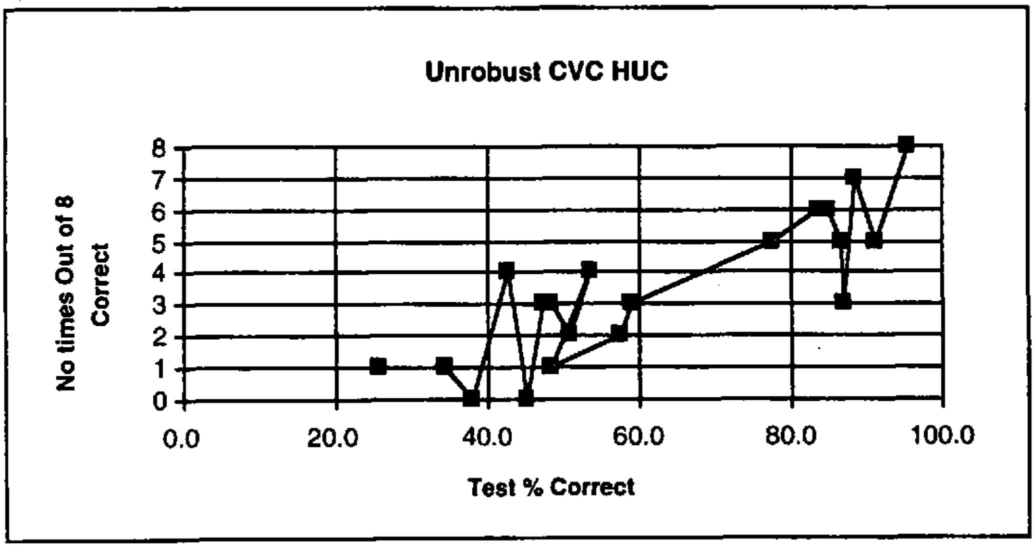
Graph 5



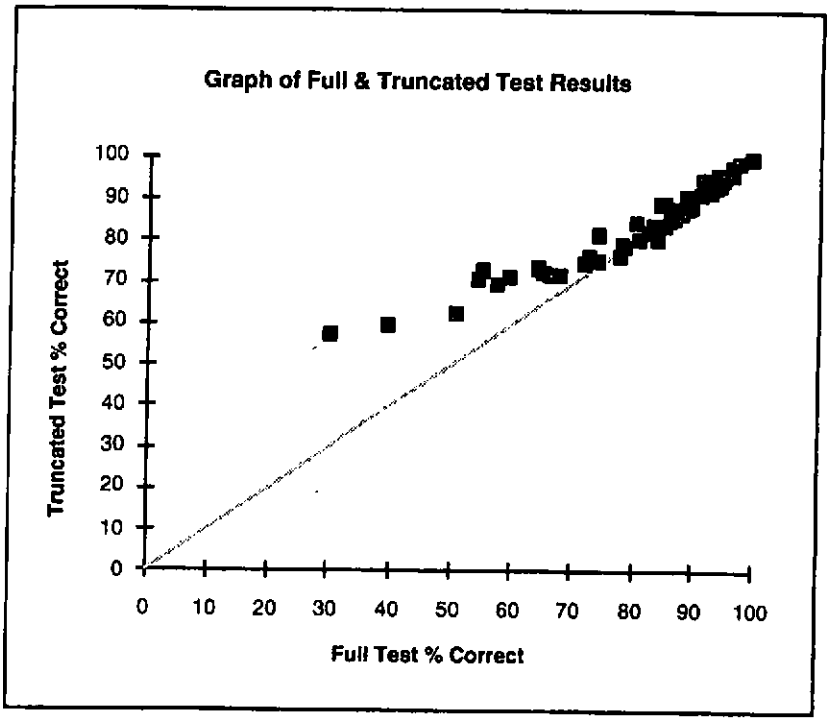
Graph 6



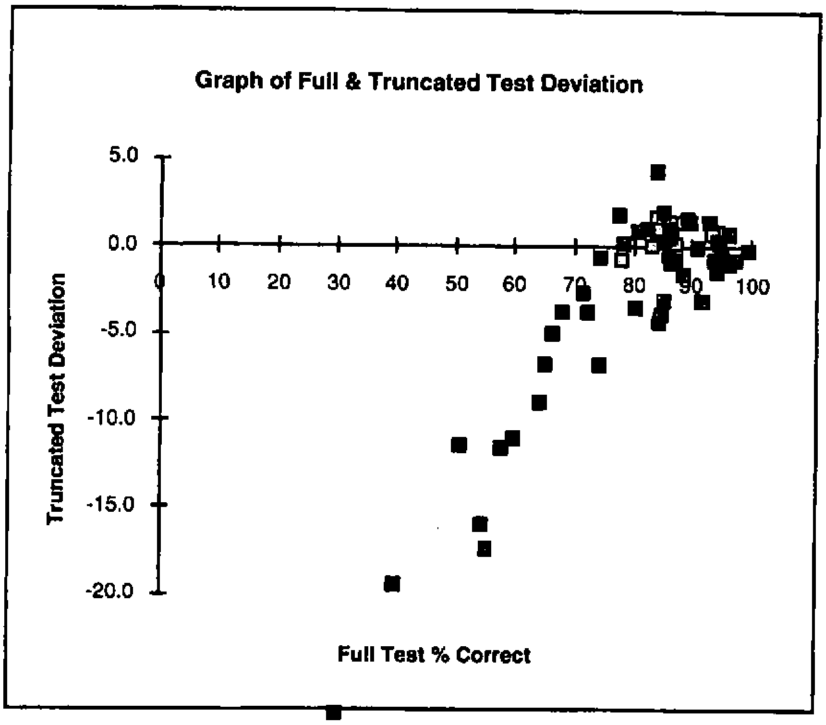
Graph 7



Graph 8



Graph 9



Graph 10