Fact Sheet 24

How accurate is my screening device?

Comparing the efficacy of educational screening systems

This document draws attention to some of the scientific publications that report accuracy rates for various forms of screening for the prediction or identification of dyslexia, literacy or learning difficulties. The information provided is only intended to alert you to some of the issues and to provide you with a starting point of enquiry or consideration. We recommend that you read the original publications and/or carry out your own literature review in order to examine the area in any detail.

It is a great shame that very few studies relating to educational assessment and prediction report rates of true positives, true negatives, and especially rates of false positives and false negatives (definitions of these terms are provided at end of the document). When attempting to draw conclusions about the efficacy of any educational screening device the importance of considering and quantifying rates of misclassifications cannot be overemphasised. The lack of reporting of false positives and false negatives may have arisen because of 1) limitations or inadequacies of the design of the research or study; 2) limitations of the statistical analyses possible or used; or 3) because the author(s) choose not to publish the rates for whatever reason. Fortunately, there have been a number of studies that have included such important measures and in the table below we report some of these published rates.

Note that even when accuracy rates for screening devices are published certain considerations should be made, and therefore we draw your attention to the caveats for interpreting published accuracy rates described below.

Caveats for interpreting published accuracy rates

1.) Consider more than just one accuracy figure

When comparing screening devices it is important to know more than just the overall accuracy figure. The number and type of mis-classifications or misses are vital pieces of information. Some test publishers do not even publish accuracy figures at all, or if they do they only publish the overall accuracy figure. Even this overall accuracy rate can be very misleading especially when you do not know the rates of false negatives and false positives. The overall predictive accuracy figure is influenced by the comparative sizes of the classification groups. Consider the fictitious screening device illustrated in Table 1. (The second and third columns in this table show the *predicted* group membership [as predicted by the screening device] and the second and third rows of the table show the *actual* group membership. By comparing the actual with the predicted you can determine how accurate the device is.) Table 1 shows a seemingly impressive overall accuracy rate of **98.7%**, but when you look further you can see that it **misses over half** of the group of individuals that it is attempting to predict!

Table 1 Example of screening accuracy rates (target group or interest group is 'D')

	Predicted D Gp	**Predicted ~D Gp**	**Total**
Actual D Gp	7	8	15
Actual ~D Gp	5	980	985
Total	12	988	1000

Table 1 shows an **overall predication rate of 98.7%** ([7+980]/1000). The **true positive rate** or 'sensitivity rate' (i.e. those the screening device accurately identifies as belonging to the target group 'D') is **46.6%** (7 / 15) and the **true negative rate** or 'specificity rate' (i.e. those

accurately predicted as belonging to the '~D' group) is **99.5%** (980/985). The highest accuracy rate of this device comes from predicting the relatively large non-interest group i.e. the ~D group.

This screening device actually misses over half of the true Ds (8 out of the 15 true Ds). This is a **false negative rate of 53.3%** which has very important implications since these are the individuals that need help and never get it because they have not been identified. In contrast the rate of **false positives at 0.5%** (5/985) is very low (i.e. those inaccurately classified as being in the D group). However, even though this rate of false positives is very low it may still have serious educational consequences. If you are providing specialist teaching to the D group (as determined by the screening device) nearly half of that group (5 out of 12) may not require or benefit from this teaching because they are not in fact true Ds but rather belong to the ~D group.

The implications and consequences to both schools and individuals of missing the educational needs of individuals or wrongly or spuriously 'remediating' individuals are very serious. It must be said that no educational screening device can be expected to perform without error, but proper and due consideration should be given when selecting which method of assessment should be adopted.

2.) Check how the individuals were selected

Another caveat when interpreting screening accuracy figures relates to how the individuals that form the various groups where chosen or selected and what criteria was used to make the group categorisations. Any screening device will give extremely high accuracy rates if the formation of the groups was based to some extent upon the results from the screening device itself. It would thereby be self-referential to some extent and would produce high accuracy figures for accurately predicting what it predicts! For example, if dyslexic referrals have been made to you and your assessment device confirms that they are dyslexic and you use this group to form your dyslexic group; and then you compare this group to a group of non-dyslexics, then this self-referential bias will occur and distort the accuracy rates in your favour.

Measurement of accuracy of any screening system should be based on data obtained from independent groups.

3.) Beware of the trade-off between false positives and false negatives

Defining the criteria for group classification will affect the resulting rates of false positives and false negatives. Altering the criteria for group membership can be used to improve the rate of either false negatives or false positives; but altering the criteria to improve the rate of one invariably worsens the rate of the other. This is another reason why both false positives and false negatives should be reported and considered together. Bear in mind that if only one rate is reported (e.g. as low) then the other non-reported rate is likely to be significantly higher.

4.) Statistical analysis assumptions and criteria should not be violated

Without getting bogged down in the minutiae of any study and its statistical analyses it is likely that most, if not all, readers of scientific papers will not examine this sort of information. (For the most part it will not even be available in a published scientific paper.) However, the reader can take a view regarding the professionalism, quality and transparency of the science that is used and reported by the authors and thereby place their trust in the analyses accordingly.

5.) Further information

For further discussion of the accuracy of educational screening devices in the field of literacy see Singleton (1997).

Accuracy rates of educational assessment devices compared

The findings from the studies indicated in Table 2 provide some food for thought! For example, the Kingslake paper shows that the Aston Index has a false positive rate of 47% and the Fletcher and Satz paper shows that teacher based assessment missed 87% of those individuals who were in fact at risk. On the other hand Lucid is obviously very pleased to include its own published accuracy rates since they compare very favourably (overall accuracy of 96% with false positives of 2.3% and false negatives of 16.7%). But it must be said that this is exactly why the CoPS system was published in the first place. Our team established the long term *predictive validity* of the tests that we use **before** any product was published. This type of research is rarely carried out! We do not know of any other test developer that has carried out such an extensive **prospective** longitudinal study – our main study followed up over 400 children for over 4 years. At the start of the study we devised and created 27 dyslexia sensitive computerised tests and measured how well each of these tests predicted (both individually and in combination) literacy development (and also numeracy development and intelligence) over the following 4 years. We also measured how well conventional standardised tests of reading, verbal and non-verbal intelligence (i.e. BAS, BPVS, MAT, Neale, Macmillan, Edinburgh Reading, WISC) predicted literacy development over the same period. Furthermore, we compared the predictive efficacy of the conventional measures with our own 27 computerised tests. Our computerised tests proved to be significantly better predictors than the conventional measures. From the original 27 tests, we selected 8 tests that were the most predictive and robust measures to form the CoPS program which went on general release in 1996.

Table 2 Comparison of accuracy rates for educational screening

Screening System	Publication reference¥	Overall prediction accuracy	False positive rate	False negative rate
Aston Index	Kingslake (1982) and Netwon et. al. (1979)	Not available	47%	21%
Infant Rating Scale	Kingslake (1982) and Lindsay (1980)	Not available	64%	69%
Swansea Evaluation Profiles	Kingslake (1982)	Not available	41%	49%
Teacher based	Fletcher & Satz (1984)	74%	14%	87%
Test battery†	Fletcher & Satz (1984)	77%	54%	74%
Teacher rating scale	Feshbach et. al. (1974)	Not available	Not available	70%
de Hirsch Predictive Index	Feshbach et. al. (1974)	Not available	Not available	74%
Phonological & other measures±	Catts (1991)	75.6%	Not available	Not available
Lucid CoPS‡	Singleton et. al. (1996)	96%	2.3%	16.7%

¥ A reference list of these publications is provided below.

† The test battery included the Peabody Picture Vocabulary Test, the Beery Test of Visual-Motor Integration, a perceptual matching test and an Alphabet Recitation task.

‡ See also Singleton et. al. (2000) for other comparisons between the CoPS measures and BAS Word recognition, BPVS, MAT and Macmillan. Please see the Related Scientific Publications page on our web site for other scientific papers that relate to our work and systems.
± This study did have a rather small sample size.

Definitions

1. **Discriminant analysis**
   'Discriminant analysis', or 'discriminant function analysis' is a form of 'regression analysis' designed for classification. It is one of the main statistical techniques to quantify rates of true positives, true negatives, false positives and false negatives. It allows two or more continuous independent variables, or predictor variables, (such as scores from reading tests, phonic skills tests, memory tests, and so on) to be used to place individuals, or cases, into the categories of a categorical dependent variable (such as 'dyslexic' and 'non-dyslexic' groups). The discriminant function analysis then allows you to quantify the accuracy of classifications with measures of true positives, true negatives, false positives and false negatives.

2. **True positive**
   A 'true positive' (also known as the Model Sensitivity) is when an individual, or case, is accurately classified into the group that you are attempting to identify or predict.

3. **True negative**
   A 'true negative' (also known as the Model Specificity) is when an individual, or case, is accurately classified into the group that you are not attempting to predict or identify.

4. **False positive**
   A 'false positive' is a misclassification where an individual (or case) is wrongly classified into the group that you are attempting to predict or identify. For example, if you are interested in identifying or predicting dyslexics, then a false positive is where you wrongly classify an individual as belonging to the dyslexic group when in actual fact they are not dyslexic.

5. **False negative**
   A 'false negative' is a misclassification where an individual (or case) is wrongly classified into the group that you are not trying to predict or identify. For example, if you are interested in identifying or predicting dyslexics, then a false negative is where you wrongly classify an individual as belonging to the non-dyslexic group when in fact they are actually dyslexic.

References

Catts, H.W. (1991). Early identification of dyslexia – evidence from a follow-up study of speech-language impaired children. *Annals of Dyslexia*, 41, 167-177

Feshbach, S., Adelman, H., and Fuller, W. W. (1974). Early identification of children with high risk of reading failure. *Journal of Learning Disabilities*, 7, 639-644.

Fletcher, J.M., and Satz, P. (1984). Test-based versus teacher-based predictions of academic achievement: A three-year longitudinal follow-up. *Journal of Pediatric Psychology*, 9 (2), 193-203.

Kingslake, B. (1982). The predictive (In) Accuracy of On-entry to school screening procedures when used to anticipate learning difficulties. *Special Education*, 10 (4), 23-26.

Lindsay, G.A., (1980). The infant rating scale. *British Journal of Educational Psychology*, 50 (2), 97-104.

Singleton, C.H., Thomas, K.V. and Leedale, R.C. (1996) *CoPS 1 Cognitive Profiling System Manual*. Lucid Research Ltd.

Singleton, C. H. (1997) Screening early literacy. In J.R.Beech and C.H.Singleton (Eds.) The Psychological Assessment of Reading. London: Routledge, pp. 67-101.

Singleton, C.H., Thomas, K.V. and Horne, J.K. (2000) Computerised cognitive profiling and the development of reading. *Journal of Research in Reading*, 23(2), 158-180.

For more information about Lucid or the developments or research please visit the Lucid web site www.lucid-research.com. The Lucid staff can be contacted by email info@lucid-research.com, telephone +44 (0)1482 862121 or fax +44 (0)1482 882911.

Please note that the information contained in this document is correct at time of going to press.