

arily in research publishing to dangle carrots to the half-hearted but I find the appetizers of the *British Medical Journal* inevitably lead me on to read the more erudite contents. Has the *Journal* not a serious obligation to attract and stimulate as broad a readership as it can and does this not therefore make the removal of the wrapper one of your major priorities?

Certain journals have that charismatic something which I call a soul. The *British Medical Journal* has it but so far our *Journal* has not quite found it.

Do we perhaps take ourselves too seriously in trying to achieve a heavy weight image and so lose our sense of proportion and humour? The RCGP is no longer an infant crying out to be heard but a mature adult, so the *Journal* should be mirroring this.

I believe it only requires a subtle change in presentation of papers into different categories with additional 'outsider' comments to increase the debate. In this way each issue has something of interest for everyone. At present the impression is that the *Journal* is being run primarily for the benefit of researchers with the needs of the readership some way behind. Editorial comment also leavens the content and this is where flair and imagination can transform any one issue.

The editorial team, however, is to be congratulated on its efforts and should take heart that we armchair critics can only further the cause for improvement with the odd observation and by trying to cross an occasional 't'.

F L P FOUIN

147 North Deeside Road
Milltimber
Aberdeen AB1 0JS

Leicester assessment package

Sir,

In mounting a detailed statistical challenge to our recommendation that the Leicester assessment package can be used for the assessment of consultation competence in general practice (letter, January *Journal*, p.51), Brauholtz overlooks the fact that assessment of professional competence, in whatever sphere, is not an exact science. In reviewing the complexities of assessing teaching competence, which is directly analogous to clinical competence, one of the foremost experts on assessment has stated '... in the end, validity is judgemental. At best it will be a form of content or construct validity, depending on the consensual judgement of "experts". To hunt for validity in statistical procedures is to pursue a will o' the wisp.'¹

A search for any assessment package

must balance what is reasonable and practical alongside acceptable statistical levels of validity and reliability. Brauholtz may not be aware of the work on assessment of clinical competence reported among other places in the proceedings of six Ottawa conferences. From many studies it appeared that over 30 hours of testing time with structured patient simulations would be required to arrive at reliable scores. In the United Kingdom situation, or elsewhere, it is highly unlikely that such methods could be applied, since both candidates and examiners are practising general practitioners. Indeed, other assessing authorities worldwide are faced with similar problems and we challenge Dr Brauholtz to nominate a single assessment process which achieves a coefficient of 0.95 or even 0.90.

Brauholtz criticizes our methodology for not testing the scores arrived at using the Leicester assessment package against a gold standard. He suggests that such a gold standard 'might be approximated by a number of experienced assessors (say four), each assessing a large number of consultations (say 20) several times over a period of months (say three)'. This is, however, merely a judgement, and some would say a reasonable judgement, but others might insist on six or eight assessors using 70 consultations over six months, just to make sure that all eventualities were taken into account.

Furthermore, since assessors in the real world get the opportunity to judge performance on one occasion rather than repeatedly, any experiment to test an assessment instrument should replicate reality. Furthermore, it is simply not feasible to test and retest candidates on 'several well-spaced occasions', as Brauholtz suggests.

While we accept that the numbers used in our reliability study were small, it was a fully crossed design so that there were no empty cells in the variance table, that is, each level of every facet was crossed with each level of every other facet. This was infinitely superior to nested designs in which true measurements of examiner variance cannot be made. True variance was deliberately introduced into the system by having candidates with a range of expertise. There are only two ways of introducing true variance; one is by having a large number of candidates at the same level, and the other is by having a small number of candidates at a wide variety of levels. Since the first was economically and practically impossible, and since we were investigating the measurement characteristics of the Leicester assessment package scale and the examiners, rather than the candidates, we concluded that this was a reasonable way to proceed. Our

study was almost unique in having a fully crossed design and the time involved for examiners, candidates and patients was a true reflection of what is possible in the real world of general practice.

We are concerned that Brauholtz seems to have conceptualized passing and failing in terms of a norm-referenced scale. It is surely axiomatic that a test of competence must be criterion referenced and those who have been working in the area do not suggest a pass/fail cut off in terms of standard deviations below the mean but in terms of basic failures in competence. In all the pilot studies and in our reported research, it has always been possible for assessors using the package to identify the small number of candidates whose competence gives cause for concern. This has also been true on the many occasions when the package has been used for regulatory assessment of doctors consulting with real patients in the examinations for the diploma in family practice (Royal College of General Practitioners/ Kuwait).

We are also disappointed that Brauholtz has interpreted our suggestion that 'all assessors should be trained and calibrated before being sanctioned to assess real candidates...' as being 'hardly a firm basis for recommending the Leicester package', when it was self-evidently meant to relate to the use of any assessment package.

In conclusion, the Leicester assessment package criteria have been shown to be valid by expert consensus² and the package as a whole capable of producing reliable results.³ Until Brauholtz, or anybody else, can cite an assessment package which has been demonstrated to be more valid, reliable and feasible in the real world, we shall continue to feel justified to recommend the use of the Leicester assessment package in formative and summative assessment of clinical competence in general practice.

ROBIN C FRASER
ROBERT K MCKINLEY
HELEN MULHOLLAND

Department of General Practice
University of Leicester
Leicester General Hospital
Leicester LE5 4PW

References

1. Stones E. Assessment of a complex skill: improving teacher education. *Assess Educ* 1994; 1: 235-251.
2. Fraser RC, McKinley RK, Mulholland H. Consultation competence in general practice: establishing the face validity of prioritized criteria in the Leicester assessment package. *Br J Gen Pract* 1994; 44: 109-113.
3. Fraser RC, McKinley RK, Mulholland H. Consultation competence in general practice: testing the reliability of the Leicester assessment package. *Br J Gen Pract* 1994; 44: 293-296.