AUTHOR'S REPLY: The curiously dogmatic tone of Lee & Chan's criticism might be easier to accept if they had read our paper more carefully, since many of the points they raise are discussed in our paper.

The rationale for our design is simple. Conventional cognitive–behavioural therapy (CBT) (19 sessions over 18 weeks) is undoubtedly an intensive form of treatment. The resources to supply this form of therapy to the 1% of young women who meet the diagnostic criteria for bulimia nervosa is unlikely to be available in most areas. The long-term outcome of this form of therapy is not nearly as optimistic as Lee & Chan suggest. Fewer than 40% of patients are in remission after one year (Fairburn *et al*, 1993). Our brief CBT (eight sessions) requires less therapist time, and can easily be taught to non-specialists. Drug treatment alone is associated with a high rate of non-compliance in bulimia nervosa. Our design aimed to test a model of treatment which would be applicable in ordinary clinical practice, reduce therapist time, improve compliance, and optimise response.

Of course there is a risk of reaching a ceiling effect with a combination of psychotherapy and pharmacotherapy. However, it stands to reason that a ceiling effect is less likely to occur with a less intensive, rather than more intensive form of therapy. This point is dealt with in detail in our paper.

The comparison of our design with a study comparing the combination of antidepressant or placebo with electroconvulsive therapy (ECT) is entirely inappropriate. The correct comparison would replace ECT with CBT. Such a study would be entirely justified, and when done, has shown that the combination of antidepressant and CBT is probably superior to each form of treatment alone (e.g. Hollan *et al*, 1992).

The authors suggest that, before performing this study, we should have compared d-fenfluramine with placebo. This study has already been done, and is discussed at length in our paper (Russell *et al*, 1988). In fact, the high drop-out rate from this study is one of the factors which led us to the study design, which succeeded in having an exceptionally low drop-out rate.

Lee & Chan's description of CBT does not do justice to the model used in our study (derived from Fairburn's model (Fairburn, 1985)). The educational component, outlining the interaction between attitudes, eating behaviour, and biology, was heavily emphasised. The effect of medication on biological processes is easily incorporated within this model, and does not, Lee & Chan assert without supportive evidence, negate the effects of one or other treatment.

Lee & Chan suggest that d-fenfluramine may have a role in the treatment of obese bulimics. It is unlikely that they will be able to test this hypothesis without including a psychological package in a treatment trial, since a trial of d-fenfluramine versus placebo will almost certainly be undermined by high drop-out rates. I look forward to reading how they will be able to learn by our mistakes.

FAIRBURN, C. (1985) Cognitive behavioural treatment for bulimia. In *Handbook of Psychotherapy for Anorexia Nervosa and Bulimia* (eds D. M. Garner & P. E. Garfinkel), pp. 160–192. New York: Guilford Press.

FAIRBURN, C. G., JONES, R., PEVELER, R. C., *et al* (1993) Psychotherapy and bulimia nervosa. Longer-term effects of interpersonal psychotherapy, behaviour therapy and cognitive behavior therapy. *Archives of General Psychiatry*, **50**, 419–428.

HOLLAN, S. D., DERUBEIS, R. J., EVANS, M. D., *et al* (1992) Cognitive therapy and pharmacotherapy for depression singly and in combination. *Archives of General Psychiatry*, **49**, 774–781.

RUSSELL, G. F. M., CHECKLEY, S. A., FELDMAN, J., *et al* (1988) A controlled trial of d-fenfluramine in bulimia nervosa. *Clinical Neuropharmacology*, **11**, S146–159.

THOMAS A. FAHY

Institute of Psychiatry
Denmark Hill
London SE5 8AF

Diagnostic agreement in psychiatry

SIR: The important study by Okasha *et al* (*Journal*, May 1993, **162**, 621–626) compared diagnostic reliability for ICD–9, ICD–10 and DSM–III–R. However, readers would have liked to have known whether the differences in overall reliability between the diagnostic systems reached statistical significance. The original description of kappa (Cohen, 1960) gave the simple arithmetic for testing for the significance of the difference between two independent kappas. Unfortunately, the reader cannot do this testing on the basis of Tables 1 and 2 in the Okasha *et al* paper because for inter-rater reliability one requires cross-tabulation of the two clinicians' allocation of cases to calculate both observed and chance agreement. Without significance testing the intriguing finding of higher overall inter-rater reliability for ICD–10 compared with both ICD–9 and DSM–III–R might be explained by chance.

Secondly, the authors correctly stated that kappa is base-rate dependent so that interpretation of results for disorders comprising less than 5% of the sample should be treated cautiously. However, in stating that this is "one of the main criticisms of kappa" they are, perhaps, unaware of the alternative view that this base-rate dependence is indeed one of kappa's strengths. Shrout *et al* (1987) have comprehensively rebutted the argument of Spitznagel &

Helzer (1985), the source Okasha *et al* quote. Briefly, the lower the base rate the more important chance agreement becomes, that is, when prevalence is low, chance agreement about the many negative cases is disproportionately large in comparison with possible disagreement about the few positive cases. Thus, the lower kappa values associated with low base rates represent "valid quantification of chance-corrected diagnostic agreement" (Shrout *et al*, 1987).

COHEN, J. (1960) A coefficient of agreement for nominal scales. *Educational & Psychological Measurement*, **20**, 37–46.

SHROUT, P. E., SPITZER, R. L. & FLEISS, J. L. (1987) Quantification of agreement in psychiatric diagnosis revisited. *Archives of General Psychiatry*, **44**, 172–177.

SPITZNAGEL, E. L. & HELZER, J. E. (1985) A proposed solution to the base rate problem in the kappa statistic. *Archives of General Psychiatry*, **42**, 725–728.

M. W. BERNADT
J. S. EMMANUEL

Farnborough Hospital
Kent BR6 8ND

SIR: Okasha *et al*, in the discussion of their comparative reliability study of several operational diagnostic systems (*Journal*, May 1993, **162**, 621–626) write that "it [reliability] establishes the ceiling for validity, the lower it is, the lower validity necessarily becomes". The first half of their statement is correct, but the second half unfortunately represents a misunderstanding which might be common among some psychiatrists who read or write about reliability and validity of their diagnoses.

In the first place it is important to remember that it is not the reliability coefficient itself but the square root thereof that sets the upper bound of the validity coefficient (Carmines & Zeller, 1979), and therefore validity can theoretically be larger than reliability. The crucial point here, however, is that, to quote Meehl (a renowned psychometrician),

> "usually the operative validity (net attenuated construct validity) runs far below that upper bound.... Hence, alterations in the format of assessment or in the content sampled, which might under some circumstances reduce reliability, could nevertheless increase the net attenuated construct validity. Similarly, changes in content or format that increase reliability may theoretically decrease validity" (Meehl, 1986).

The same author cites the modified Rorschach test, which attempted during World War II to test large numbers of people and to increase the reliability by altering the original open-ended, unstructured format, as an example of the latter paradox because "it seemed to eliminate whatever slight validity the instrument had as usually administered." (Meehl, 1986)

CARMINES, E. G. & ZELLER, R. A. (1979) *Reliability and Validity Assessment*. London: Sage Publications.

MEEHL, P. E. (1986) Diagnostic taxa as open concepts: metatheoretical and statistical questions about reliability and construct validity in the grand strategy of nosological revision. In *Contemporary Directions in Psychopathology* (eds T. Millon & G. L. Klerman), pp. 215–231. New York: Guilford.

TOSHIAKI FURUKAWA

Nagoya City University School of Medicine
Mizuho-cho, Mizuho-ku, Nagoya 467 Japan

AUTHORS' REPLY: Bernadt & Emmanuel raise the question of whether the difference in kappa values between ICD–10, ICD–9, and DSM–III–R has reached statistical significance. We are unaware of any special statistical measure to do that.

As in many reliability studies, we used the guidelines laid down by Landis & Koch (1977). Accordingly, a kappa value of 0.6–0.80 is considered good or substantial agreement, and a kappa value above 0.80 is taken to indicate very good or almost perfect agreement. On this basis we were able to reach the conclusion that:

(a) for inter-rater reliability at three-digit level, both ICD–10 and ICD–9 proved to be generally superior to DSM–III–R (kappa values of +0.823, +0.787, and +0.636 respectively)
(b) for inter-rater reliability at four-digit level, ICD–10 was clearly superior to both DSM–III–R and ICD–9 (kappa values of +0.80, +0.63, and +0.62 respectively)
(c) for all systems, inter-rater reliability at three- and four-digit levels was above +0.80, thus it was difficult to reach any conclusion out of those figures.

As for the requested tabulation, this would be impossible to construct as the ratings were made for each system separately. We accept the comment made on the kappa being base-rate dependent as we had no access to Shrout *et al*'s (1987) paper (see above).

We value the comments made by Dr Furukawa that clarifies an area of misunderstanding. However, we believe that our statement stands true as there is no contradiction with Dr Furukawa's comment and it does not imply that reliability has to be greater than validity, but only indicates strong positive correlation between both.

LANDIS, J. R. & KOCH, G. G. (1977) The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.

AHMED OKASHA

Ain Shams University
3 Shawarby Street, Cairo, Egypt