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