Standards for DSM-5 Reliability

To the Editor: In the January issue of the *Journal*, Helena Chmura Kraemer, Ph.D., and colleagues (1) ask, in anticipation of the results of the DSM-5 field trial reliability study, how much reliability is reasonable to expect. They argue that standards for interpreting kappa reliability, which have been widely accepted by psychiatric researchers, are unrealistically high. Historically, psychiatric reliability studies have adopted the Fleiss standard, in which kappas below 0.4 have been considered poor (2). Kraemer and colleagues propose that kappas from 0.2 to 0.4 be considered "acceptable." After reviewing the results of three test-retest studies in different areas of medicine (diagnosis of anemia based on conjunctival inspection, diagnosis of pediatric skin and soft tissue infections, and bimanual pelvic examinations) in which kappas fall within ranges of 0.36–0.60, 0.39–0.43, and 0.07–0.26, respectively, Kraemer et al. conclude that "to see κ_I for a DSM-5 diagnosis above 0.8 would be almost miraculous; to see κ_I between 0.6 and 0.8 would be cause for celebration." Therefore, they note that for psychiatric diagnoses, "a realistic goal is κ_I between 0.4 and 0.6, while κ_I between 0.2 and 0.4 would be acceptable."

When we (R.L.S., J.B.W.W.) conducted the DSM-III field trial, following the Fleiss standard, we considered kappas above 0.7 to be "good agreement as to whether or not the patient has a disorder within that diagnostic class" (3). According to the Kraemer et al. commentary, the DSM-III field trial results should be cause for celebration: the overall kappa for axis I disorders in the test-retest cohort (the one most comparable methodologically to the DSM-5 sample) was 0.66 (3). Therefore, test-retest diagnostic reliability of at least 0.6 is achievable by clinicians in a real-world practice setting, and any results below that standard are a cause for concern.

Kraemer and colleagues' central argument for these diagnostic reliability standards is to ensure that "our expectations of DSM-5 diagnoses … not be set unrealistically high, exceeding the standards that pertain to the rest of medicine." Although the few cited test-retest studies have kappas averaging around 0.4, it is misleading to depict these as the "standards" of what is acceptable reliability in medicine. For example, the authors of the pediatric skin lesion study (4) characterized their measured test-retest reliability of 0.39–0.43 as "poor." Calling for psychiatry to accept kappa values that are characterized as unreliable in other fields of medicine is taking a step backward. One hopes that the DSM-5 reliability results are at least as good as the DSM-III results, if not better.

References

1. Kraemer HC, Kupfer DJ, Clarke DE, Narrow WE, Regier DA: DSM-5: how reliable is reliable enough? Am J Psychiatry 2012; 169:13–15
2. Fleiss J: Statistical Methods for Rates and Proportions, 2nd ed. New York, Wiley, 1981
3. Spitzer R, Forman J, Nee J: DSM-III field trials, I: initial interrater diagnostic reliability. Am J Psychiatry 1979; 136:815–817
4. Marin JR, Bilker W, Lautenbach E, Alpern ER: Reliability of clinical examinations for pediatric skin and soft-tissue infections. Pediatrics 2010; 126:925–930

ROBERT L. SPITZER, M.D.
JANET B.W. WILLIAMS, Ph.D.
Princeton, N.J.
JEAN ENDICOTT, Ph.D.
New York City

Dr. Spitzer reports no financial relationships with commercial interests. Dr. Williams works for MedAvante, a pharmaceutical services company. Dr. Endicott has received research support from Cyberonics, the New York State Office of Mental Hygiene, and NIH and has served as a consultant or advisory board member for AstraZeneca, Bayer Schering, Berlex, Cyberonics, Eli Lilly, Forest Laboratories, GlaxoSmithKline, Otsuka, Shire, and Wyeth-Ayerst.

This letter (doi: 10.1176/appi.ajp.2012.12010083) was accepted for publication in March 2012.

Response to Spitzer et al. Letter

To the Editor: Homage must be paid to the DSM-III field trials (1) that strongly influenced the design of the DSM-5 field trials. It could hardly be otherwise, since methods for evaluating categorical diagnoses were developed for DSM-III by Dr. Spitzer and his colleagues, Drs. Fleiss and Cohen. However, in the 30 years after 1979, the methodology and the understanding of kappa have advanced (2), and DSM-5 reflects that as well.

Like DSM-III, DSM-5 field trials sampled typical clinic patients. However, in the DSM-III field trials, participating clinicians were allowed to select the patients to evaluate and were trusted to report all results. In the DSM-5 field trials, symptomatic patients at each site were referred to a research associate for consent, assigned to an appropriate stratum, and randomly assigned to two participating clinicians for evaluation, with electronic data entry. In DSM-III field trials, the necessary independence of the two clinicians evaluating each patient was taken on trust. Stronger blinding protections were implemented in the DSM-5 field trials. Selection bias and lack of blindness tend to inflate kappas.

The sample sizes used in DSM-III, by current standards, were small. There appear to be only three diagnoses for which 25 or more cases were seen: any axis II personality disorder (kappa=0.54), all affective disorders (kappa=0.59), and the subcategory of major affective disorders (kappa=0.65). Four kappas of 1.00 were reported, each based on three or fewer cases; two kappas below zero were also reported based on 0–1 cases. In the absence of confidence intervals, other kappas may have been badly under- or overestimated. Since the kappas differ from one diagnosis to another, the overall kappa cited is uninterpretable (1).

Standards reflect not what we hope ideally to achieve but what the reliabilities are of diagnoses that are actually useful in practice. Recognizing the possible inflation in DSM-III and DSM-IV results, DSM-5 did not base its standards for kappa entirely on their findings. Fleiss articulated his standards before 1979 when there was little experience using kappa. Are the experience-based standards (3) we proposed unreasonable? There seems to be major disagreement only about