



Published in final edited form as:

J Abnorm Psychol. 2015 August ; 124(3): 764–769. doi:10.1037/abn0000069.

Method Matters: Understanding Diagnostic Reliability in *DSM-IV* and *DSM-5*

Michael Chmielewski,

Department of Psychology, Southern Methodist University

Lee Anna Clark,

Department of Psychology, University of Notre Dame

R. Michael Bagby, and

Department of Psychology, University of Toronto, Ontario, Canada

David Watson

Department of Psychology, University of Notre Dame

Abstract

Diagnostic reliability is essential for the science and practice of psychology, in part because reliability is necessary for validity. Recently, the *DSM-5* Field Trials documented lower diagnostic reliability than past field trials and the general research literature, resulting in substantial criticism of the *DSM-5* diagnostic criteria. Rather than indicating specific problems with *DSM-5*, however, the Field Trials may have revealed long-standing diagnostic issues that have been hidden due to a reliance on audio/video-recordings for estimating reliability. We estimated the reliability of *DSM-IV* diagnoses using both the standard audio-recording method and the test-retest method used in the *DSM-5* Field Trials, in which different clinicians conduct separate interviews. Psychiatric patients ($N = 339$) were diagnosed using the SCID-I/P; 218 were diagnosed a second time by an independent interviewer. Diagnostic reliability using the audio-recording method ($N = 49$) was “good” to “excellent” (M kappa = .80) and comparable to the *DSM-IV* Field Trials estimates. Reliability using the test-retest method ($N = 218$) was “poor” to “fair” (M kappa = .47) and similar to *DSM-5* Field-Trials’ estimates. Despite low test-retest diagnostic reliability, self-reported symptoms were highly stable. Moreover, there was no association between change in self-report and change in diagnostic status. These results demonstrate the influence of method on estimates of diagnostic reliability.

Introduction

Diagnostic reliability is essential for advancing the science and practice of psychology (Regier et al., 2013). Without reliable diagnoses, accurate identification of risk factors for psychopathology becomes nearly impossible. Diagnostic unreliability can lead to erroneous

Address correspondence to Michael Chmielewski, Southern Methodist University, Department of Psychology, P.O. Box 75275-0442, Dallas, TX, 75275. mchmielewski@smu.edu.

Results suggest that (1) reliability of psychological diagnoses obtained from the SCID may be lower than commonly believed and (2) the reliability of common *DSM-IV* and *DSM-5* diagnoses are actually quite similar.

interpretations regarding the structure of mental disorders, their natural course, the nature of symptom change, and treatment efficacy; moreover, it greatly increases the likelihood that research findings will not replicate. Finally, diagnostic reliability is essential for diagnostic validity (Nelson-Gray, 1991; Spitzer & Fleiss, 1974).

Prior to *DSM-III* (American Psychiatric Association, 1980), diagnostic reliability was poor, due in part to the lack of specific diagnostic criteria (Spitzer & Fleiss, 1974). *DSM-III*'s operationalized criterion sets improved diagnostic reliability, leading to the widespread belief that the manual solved this problem (Klerman, 1984; Spitzer, Forman, & Nee, 1979). This belief, combined with the resources required to obtain estimates of diagnostic reliability, has led to cursory attention being given to diagnostic judgments in the scientific literature. For example, researchers simply state that interviewers were thoroughly trained, or that the specific interview(s) used were shown to be reliable in the past. The end result is that researchers rarely provide specific estimates of diagnostic reliability derived from the studied sample. In 2013, the *Journal of Abnormal Psychology* published 67 articles that reported diagnostic data on specific *DSM* disorders; of these, only 18 (27%) included kappa reliability estimates derived from the study sample.

Diagnostic Reliability in *DSM-III*, *DSM-IV*, and *DSM-5*

Given this situation, it is not surprising that the *DSM-5* Field Trials—which resulted in lower kappa reliability estimates than past field trials and the general research literature—have generated considerable controversy and concern regarding the new manual's merits. Members of the *DSM-5* Task Force, using revised kappa guidelines (Kraemer, Kupfer, Clarke, Narrow, & Regier, 2012), interpreted the *DSM-5* Field Trials results as indicating “good to very good reliability” for most diagnoses (Regier et al., 2013). Others have been far more critical (Frances, 2012; Spitzer, Williams, & Endicott, 2012), arguing that the manual “flunked its reliability tests” (Frances, 2012) and that traditional kappa guidelines should be applied (Frances, 2012; Spitzer et al., 2012).

Many have blamed the *DSM-5* itself, arguing that specific wording in the *DSM-5* diagnostic-criterion sets led to lower reliabilities (Frances, 2012). However, this cannot explain why diagnoses that were essentially unchanged from *DSM-IV* (American Psychiatric Association, 2000), such as major depressive disorder (MDD), demonstrated substantially lower kappas in the *DSM-5* Field Trials compared to previous estimates. Others have suggested that (a) the lack of standardized interviews in the *DSM-5* Field Trials (Regier et al., 2013) or (b) sample differences between the *DSM-5* Field Trials (which used representative samples) and previous field trials (which did not) contributed to the lower reliabilities (Regier et al., 2013).

Audio/Video-Recording Versus Test-Retest Methods

Although all of the above could have contributed to lower kappa reliabilities in the *DSM-5* Field Trials, we believe that much of the difference is attributable to the methods used to assess diagnostic reliability. On the rare occasions that sample-specific estimates of diagnostic reliability are reported in the research literature, they are estimated almost exclusively using the audio/video-recording method. Of the 18 *Journal of Abnormal*

Psychology studies published in 2013 that reported sample-specific estimates of diagnostic reliability, 17 (94%) used the audio/video-recording method. In this method, one clinician conducts the interview and provides diagnoses; a second “blinded” clinician then provides an independent set of diagnoses based on recordings of the interview. Reliability estimates using this method typically are high, consistent with the view that diagnostic reliability is no longer a concern.

Unfortunately, the audio/video recording approach can be expected to yield higher kappa estimates than other methods for several reasons. First, once interviewing clinicians conclude that a patient does not meet diagnostic criteria for a disorder, they typically do not ask about the remaining symptoms; therefore, the second clinician does not have all the information necessary to confer a diagnosis independently and agreement is achieved by default. This problem is not remedied by semi-structured interviews because most interviews, such as the SCID-I/P, include “skip-outs.” Second, only the interviewing clinician can probe patient responses or obtain additional information regarding specific symptoms. Third, two clinicians may obtain different responses if separate interviews are conducted. This is not to say that patients are experiencing symptoms differently, but simply that they may volunteer different information to the two clinicians. As such, the audio/video-recording method, which constrains the information provided to the two diagnosticians to be identical, can be expected to generate higher kappa values compared to those obtained when separate interviews are conducted (Kraemer et al., 2012; Zimmerman, 1994).

If diagnostic reliability is defined as the extent to which a patient would receive the same diagnosis at different hospitals or clinics, or the extent to which different studies are recruiting similar patients, then the test-retest method provides a more meaningful estimate of diagnostic reliability (Kraemer et al., 2012; Williams et al., 1992). In the test-retest method, two different interviewers independently conduct separate interviews. Because true change in clinical status could occur over the test-retest interval, artificially lowering diagnostic reliability (Brown, Di Nardo, Lehman, & Campbell, 2001), it is essential that the test-retest time frame is short enough that true change in is highly unlikely. Blashfield and Livesley (1991, p. 265) argued that test-retest diagnostic reliability is especially important for diagnostic validity, stating that “short-term stability must be expected” and that “failure to demonstrate stability when different assessments are used raises questions about validity.”

The *DSM-IV* Field Trials for the mood, anxiety, and substance use disorders either exclusively used the audio-recording method or did not assess reliability at all. The *DSM-III* Field Trials used both the joint interview ($N = 150$)—which is similar to the audio/video-recording method—and test-retest methods ($N = 131$); however, individual diagnoses were not examined, making it difficult to compare results across studies (see Kraemer et al., 2012; Spitzer et al., 1979). In contrast, the *DSM-5* Field Trials exclusively used the test-retest method (median test-retest interval = 1 week). Therefore, the *DSM-5* Field Trial estimates may be more accurate representations of *DSM* diagnostic reliability in typical settings. Put differently, apparent differences in diagnostic reliability across *DSM* editions may largely reflect the different methods that were used to assess them (Kraemer et al., 2012).

In this study, conducted prior to the *DSM-5* Field Trials, we estimated the reliability of *DSM-IV* diagnoses using both the audio-recording and test-retest methods. We used a large unselected patient sample to represent the typical clinical setting and to ensure that reliability was not inflated by the use of a highly selected sample (see Kraemer et al., 2012). All diagnoses were made by thoroughly trained interviewers using the SCID-I/P. Additionally, self-report data were collected during the same sessions to assess whether patients' experience of their symptoms changed over the 1-week test-retest interval.

Method

Participants and Procedure

Psychiatric patients ($N = 339$; age range = 18–83 years, $M = 42.4$ years; 229 female, 109 male, 1 unreported) were recruited from the outpatient Adult Psychiatry Clinic at the University of Iowa Hospital and Clinics, and other outpatient and residential psychology clinics in Iowa. Participants were at least 18 years of age and fluent in English, with no other exclusion criteria. Participants completed self-report measures in small-group sessions; they were taken individually to a private room for audio-recorded SCID-I/P interviews. They then returned to the small-group session to complete the measures. All participants were invited to return for a second session held 1 week later; 218 (64%) returned and completed the full protocol a second time. The 1-week interval was chosen to decrease the likelihood that true diagnostic change would occur, while being long enough to reduce memory effects; it is equivalent to the median test-retest interval in the *DSM-5* Field Trials. In the vast majority of cases (86%), the two interviews occurred exactly 7 days apart ($M = 7.2$, $SD = 1.44$, range = 2 to 17 days). The prevalence rate of *DSM-IV* diagnoses (Table 1) was very consistent across assessments. Participants who completed only the first session were more likely to be male and diagnosed with a substance use disorder than participants who completed both sessions ($p = .004$); there were no other differences in diagnoses, in ethnicity, or in self-report scores.

To assess reliability using the audio-recording method we followed the convention from the most stringent studies in the *Journal of Abnormal Psychology*: audio-recording reliability for 10–15% of participants. This resulted in 49 audiotapes being selected randomly, mostly from Time 1, and scored independently by a second interviewer. To assess reliability using the test-retest method, different “blinded” interviewers conducted the interviews at Time 1 and Time 2. The proportion of interviews conducted and audiotapes rated by any single interviewer was consistent across all conditions. Results were very similar when restricting analyses to cases ($N=31$) with estimates from both methods.

Interviews and Measures

Interviewers—Interviewers were at least masters' level and had previous training and experience with semi-structured diagnostic interviews. Additionally, all interviewers underwent formal training on the SCID-I/P; this included training videos, 1 month of training from established interviewers, and joint ratings of audio-recordings from previous studies. Once interviewers were trained to agreement with the SCID trainers, based on joint ratings of previous audio-recordings and role-plays, 7 weeks of additional training

interviews were conducted in a college-student sample prior to their starting patient interviews. During these 7 weeks, interviewers met weekly with SCID-I/P trainers to discuss interview questions, develop consensus, and listen to recorded interviews to ensure that diagnostic protocol was followed. These meetings continued for the first month of patient interviews and then as necessary for the remainder of the study. Weekly meetings were also held throughout the course of the study with Ph.D.-level clinical faculty.

Interviews—Participants were diagnosed using the mood-disorders, anxiety-disorders, psychotic-disorder, and substance-use-disorders modules of the SCID-I/P (First, Spitzer, Gibbon, & Williams, 2002). We report results for 9 *DSM-IV* diagnoses: MDD, panic disorder, posttraumatic stress disorder (PTSD), social phobia, dysthymic disorder, obsessive-compulsive disorder (OCD), specific phobia, bipolar-I disorder, and generalized anxiety disorder (GAD). We also report results for three broader diagnostic groupings, two of which (substance-use disorder and “other” bipolar disorder) were created due to individual diagnoses containing an inadequate number of cases; the third, “any psychotic disorder,” was decided *a priori*. Hierarchical exclusion rules for GAD were relaxed to permit comorbid diagnoses.

Self-Report Measures—Participants completed the *Inventory of Depression and Anxiety Symptoms* (IDAS; Watson et al., 2007). The IDAS scales show strong psychometric properties compared to commonly used depression and anxiety measures (Watson et al., 2007, 2008). We present data for the five IDAS scales that have the strongest links to specific *DSM-IV* diagnoses (Watson et al., 2008): General Depression, Social Anxiety, Panic, Traumatic Intrusions, and Anxious Mood.

Results

Diagnostic Reliability

Estimates of diagnostic reliability assessed by the audio-recording method are shown in the left column of Table 2, along with bootstrapped 95% confidence intervals (samples = 1000). The mean kappa of .80, as well as those of the majority of diagnoses, would be considered “excellent” by traditional standards (Fleiss, 1981; Spitzer et al., 1979). Diagnostic reliability using the test-retest method is shown in the right column of Table 2. The mean kappa of .47 would be considered only “fair” by traditional standards and only a single diagnosis demonstrated “good” reliability. Moreover, approximately 25% of diagnoses would be considered “poor” by traditional standards.

When comparing individual kappas across methods, it is important to note that examining bootstrapped confidence intervals provides a more conservative test of whether two kappas differ in magnitude than does null hypothesis testing (Samuel et al., 2011). However comparing confidence intervals is the only method available for kappas derived from dependent samples (McKenzie et al., 1996; Samuel et al., 2011). It is noteworthy that the confidence intervals for four diagnoses—MDD, OCD, Social Phobia, and Dysthymia—do not overlap, clearly indicating a significant difference across method.

Despite the test-retest diagnostic disagreement between interviewers (see Table 3), patients' self-reports of their symptoms on the IDAS showed little change across the 1-week retest interval (test-retest $r_s = .75$ to $.84$; mean = $.80$). To ensure that diagnostic disagreement was not due to a true change in symptom presentation or severity, we created ordinal diagnostic change scores (i.e., -1 , 0 , $+1$) from Time 1 to Time 2 for each diagnosis and correlated these scores with changes on the corresponding IDAS scale (i.e., Time 1 score minus Time 2 score). Change on the IDAS scales was unrelated to change in this metric of diagnostic status ($M r = .06$, range = $-.04$ to $.14$).

One complication in comparing these results to those of the *DSM-IV* and *DSM-5* Field Trials is that different sets of diagnoses were examined across studies. However, this same general pattern emerged even when restricting analysis to the exact same diagnoses (mean kappas: current audio-recording = $.78$, *DSM-IV* Field Trial audio-recording = $.60$, current test-retest = $.53$, *DSM-5* test-retest = $.43$). Interestingly, kappas from the current study appear to be as high as or higher than their counterparts in the field trials.

Discussion

The results of our study strongly suggest that apparent differences in diagnostic reliability between the *DSM-IV* and *DSM-5* Field Trials largely reflect the methods that were used to assess reliability, rather than actual differences in the diagnoses themselves. Additionally, the study results suggest patients may not receive the same diagnosis across clinics or studies. In our data, the audio-recording method resulted in estimates of diagnostic reliability that would be considered "excellent" by traditional standards (M kappa = $.80$). However, the test-retest method resulted in estimates of diagnostic reliability (M kappa = $.47$) that would be considered only "fair" by traditional standards. Moreover, approximately $\frac{1}{4}$ of the test-retest estimates would be considered "poor". It is important to note that (1) the SCID-I/P was used; (2) patients' self-reported symptoms were very stable (M test-retest $r = .80$); and (3) change in self-report was unassociated ($M r = .06$) with change in diagnostic status. We also note that Zanarini et al. (2000) found a similar reduction in diagnostic reliability using a small number of audio-recordings ($N = 27$) and test-retests ($N = 52$) in a non-representative patient sample.

Three previous studies examined the test-retest reliability of current diagnoses in large patient samples using *DSM-III-R* ($N = 267$: Di Nardo, Moras, Barlow, Rapee, & Brown, 1993; $N = 390$: Williams et al., 1992) or *DSM-IV* ($N = 362$: Brown et al., 2001) criteria. Comparing identical diagnoses, reliability in the current study appears slightly lower (kappa = $.61$ vs. $.66$) than in Williams et al. (1992) and lower than in Di Nardo et al. (1993) and Brown et al. (2001) (kappa = $.45$ vs. $.60$ and $.65$, respectively). However, in those studies, case conferences were held after every set of interviews to identify causes of diagnostic disagreement and to reach a consensus diagnosis, which likely raised kappa values. In the current study, diagnostic issues were discussed only when an interviewer had a question about an interview they had conducted, which more closely resembles the typical research/clinical setting. In addition, Brown et al. (2001) and Di Nardo et al. (1993) used several inclusion and exclusion criteria that may have reduced diagnostic "noise." Williams et al. (1992) also provided interviewers with summaries of hospital admission records that are not

available in most studies or practices outside of a hospital setting. As such, these three studies likely represent the upper limit of test-retest diagnostic reliability under specialized conditions. In contrast, the current-study results may be more representative of diagnostic reliability in the typical research study that uses well-trained interviewers to conduct semi-structured interviews. Finally, it is worth noting that Brown et al. (2001) and Di Nardo et al. (1994) used the Anxiety Disorder Interview Schedule (ADIS; Di Nardo, Brown, & Barlow, 1994; Di Nardo, Moras, Barlow, Rapee, & Brown, 1993) whereas the current study and Williams et al. (1992) used the SCID, which may have affected the diagnostic reliabilities obtained.

Implications for DSM-5

There has been considerable criticism of the *DSM-5* Field Trial results, with many arguing that changes to diagnostic criteria in *DSM-5* are to blame for the apparent reduction in reliability (Frances, 2012). The current study suggests that this criticism may not be warranted. Instead, it appears that the *DSM-5* Field Trials' test-retest design may have revealed longstanding diagnostic issues. When assessed by the standard audio-recording method, the reliability of *DSM-IV* diagnoses in this study (M kappa = .80) was equivalent or superior to corresponding values from the *DSM-IV* (M kappa = .65) and *DSM-III* (M kappa = .78) Field Trials, which also used audio/video-recording and joint-interview methods, respectively. However, diagnostic reliability for common *DSM-IV* diagnoses using the test-retest method (M kappa = .47) was very similar to the level of reliability observed in the *DSM-5* Field Trials (M kappa = .44), which also used the test-retest method. This general finding held even when restricting analysis to include the exact same diagnoses across studies.

It is noteworthy that 1) audio-recording-based kappas in our representative sample were similar to those from previously reported non-representative samples and 2) the test-retest reliability in the current study using the SCID-I/P was similar to that of the *DSM-5* Field Trials, which did not use semi-structured interviews, but instead used a systematic and semi-structured method to explore and rate sets of symptoms. Interestingly, Williams et al. (1992) reported that the use of the SCID in their study did not result in higher reliability estimates compared to the *DSM-III* Field Trials, which did not use structured interviews. These findings suggest that sample differences and the lack of standardized interviews in the *DSM-5* Field Trials likely do not explain the bulk of the observed difference in diagnostic reliability.

How Reliable is Reliable Enough?

Prior to the publication of the *DSM-5* Field Trials' results, Spitzer and colleagues (2012) argued that a kappa below .60 would be concerning, even considering the *DSM-5*'s test-retest methodology. Given this viewpoint, the current results regarding the test-retest reliability of *DSM-IV* diagnoses (M kappa = .47) would be a cause for concern as well. Although some have argued that reliability levels in this range are adequate for clinical care (Regier et al., 2013), we would argue that reliability in this low range is insufficient to facilitate the advancement of psychological research. It has long been argued that reliability sets an upper limit on validity (Nelson-Gray, 1991; Spitzer & Fleiss, 1974); however, the

current test-retest analyses are arguably more strongly linked to diagnostic validity than are results obtained via audio/video-recording or joint-interview methods. As Blashfield & Livesley (1991) noted, high test-retest reliability over short timeframes is essential for diagnostic validity. From this perspective, the current results—together with those of the *DSM-5* Field Trials—raise questions about the reliability and validity of *DSM* diagnoses, at least as assessed by the SCID, the most widely used semi-structured diagnostic interview.

Limitations

We could not replicate the large complex stratified-sample design of the *DSM-5* Field Trials, which contributed to the size of our confidence intervals. Relatedly, although we analyzed more audio-tapes ($n = 49$) than most studies, it would have been preferable to obtain audio-tape-based ratings for all interviews to further reduce confidence intervals. It is unclear whether our results generalize to disorders not included in the current study. The included disorders, while more prevalent, contain overlapping features which may lead to diagnostic disagreements (i.e., two interviewers consider the same symptom to reflect different diagnoses); studies with diagnoses that are more easily distinguishable may find higher levels of reliability. It also is possible that relaxing hierarchical exclusion rules for GAD lowered its reliability. Although interviewers were not doctoral-level clinicians, they achieved equivalent or greater reliability than doctoral-level clinicians in previous field trials. Even if doctoral-level clinicians would have achieved higher reliability, our general findings regarding differences between the audio-recording and test-retest method likely would stand. The current study was conducted prior to creation of the *DSM-5*, and thus did not assess *DSM-5* diagnoses. Finally, we did not examine potential causes of disagreement between interviewers (see Brown et al., 2001).

Conclusions

Although psychiatric diagnoses have become more reliable and valid since the publication of *DSM-III* (Klerman, 1984; Spitzer et al., 1979), the current results—together with those from the *DSM-5* Field Trials—suggest that the reliability of psychological diagnosis may be lower than commonly believed. From this perspective, the *DSM-5* Field Trials appear to have brought to light important issues regarding diagnostic reliability that have existed for some time, but were obfuscated by common methods of assessing reliability. In many ways, the controversy regarding the *DSM-5* can be interpreted as “blaming the messenger,” as the current results, combined with those of the *DSM-5* Field Trials, suggest that the diagnostic reliability of the *DSM-IV* and *DSM-5* are likely quite similar. Our results add to the large body of literature documenting the limitations of categorical diagnoses (Markon, Chmielewski, & Miller, 2011) and indicate there is significant room for improvement in diagnostic reliability. At the very least, our results indicate that psychopathology researchers should give the issue of diagnostic reliability more than cursory attention.

Acknowledgments

This research was supported by National Institute of Mental Health Grant R01-MH068472 to Dr. Watson.

References

- American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 3. Washington, DC: Author; 1980.
- American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4. Washington, DC: Author; 2000. text revision
- Blashfield RK, Livesley WJ. Metaphorical analysis of psychiatric classification as a psychological test. *Journal of Abnormal Psychology*. 1991; 100(3):262–270. [PubMed: 1918603]
- Brown TA, Di Nardo PA, Lehman CL, Campbell LA. Reliability of DSM-IV anxiety and mood disorders: Implications for the classification of emotional disorders. *Journal of Abnormal Psychology*. 2001; 110(1):49–58. <http://doi.org/10.1037//0021-843X.110.1.49>. [PubMed: 11261399]
- Di Nardo, PA.; Brown, TA.; Barlow, DH. Anxiety Disorders Interview Schedule for DSM–IV: Lifetime version (ADIS–IV–L). San Antonio, TX: Psychological Corporation; 1994.
- Di Nardo P, Moras K, Barlow DH, Rapee RM, Brown TA. Reliability of DSM-III-R anxiety disorder categories. Using the Anxiety Disorders Interview Schedule-Revised (ADIS-R). *Archives of General Psychiatry*. 1993; 50(4):251–256. [PubMed: 8466385]
- First, M.; Spitzer, RL.; Gibbon, M.; Williams, JBW. Structured clinical interview for DSM-IV-TR axis I disorders, research version, patient edition (SCID-I/P). New York: Biometrics Research, New York State Psychiatric Institute; 2002.
- Frances, A. Newsflash from APA meeting: DSM-5 has flunked its reliability tests. 2012 May 8. Retrieved March 21, 2013, from http://www.huffingtonpost.com/allen-frances/dsm-5-reliability-tests_b_1490857.html
- Fleiss, Joseph L. Statistical methods for rates and proportions. 2. Hoboken, New Jersey: Wiley; 1981.
- Klerman GL. A debate on DSM-III: The advantages of DSM-III. *The American Journal of Psychiatry*. 1984; 141(4):539–542. [PubMed: 6703133]
- Kraemer, HC.; Kupfer, DJ.; Clarke, DE.; Narrow, WE.; Regier, DA. DSM-5: How reliable is reliable enough?; *The American Journal of Psychiatry*. 2012. p. 13-15.<http://doi.org/10.1176/appi.ajp.2011.11010050>
- Markon KE, Chmielewski M, Miller CJ. The reliability and validity of discrete and continuous measures of psychopathology: A quantitative review. *Psychological Bulletin*. 2011; 137(5):856–879. <http://doi.org/10.1037/a0023678>. [PubMed: 21574681]
- McKenzie DP, Mackinnon AJ, Péladeau N, Onghena P, Bruce PC, Clarke DM, McGorry PD. Comparing correlated kappas by resampling: Is one level of agreement significantly different from another? *Journal of Psychiatric Research*. 1996; 30(6):483–492. [http://doi.org/10.1016/S0022-3956\(96\)00033-7](http://doi.org/10.1016/S0022-3956(96)00033-7). [PubMed: 9023792]
- Nelson-Gray RO. DSM-IV: empirical guidelines from psychometrics. *Journal of Abnormal Psychology*. 1991; 100(3):308–315. [PubMed: 1918610]
- Regier DA, Narrow WE, Clarke DE, Kraemer HC, Kuramoto SJ, Kuhl EA, Kupfer DJ. DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical diagnoses. *The American Journal of Psychiatry*. 2013; 170(1):59–70. <http://doi.org/10.1176/appi.ajp.2012.12070999>. [PubMed: 23111466]
- Samuel DB, Hopwood CJ, Ansell EB, Morey LC, Sanislow CA, Markowitz JC, Grilo CM. Comparing the temporal stability of self-report and interview assessed personality disorder. *Journal of Abnormal Psychology*. 2011; 120(3):670–680. <http://doi.org/10.1037/a0022647>. [PubMed: 21443287]
- Spitzer RL, Fleiss JL. A re-analysis of the reliability of psychiatric diagnosis. *The British Journal of Psychiatry*. 1974; 125:341–347. <http://doi.org/10.1192/bjp.125.4.341>. [PubMed: 4425771]
- Spitzer RL, Forman JB, Nee J. DSM-III field trials: I. Initial interrater diagnostic reliability. *The American Journal of Psychiatry*. 1979; 136(6):815–817. [PubMed: 443467]
- Spitzer RL, Williams JBW, Endicott J. Standards for DSM-5 reliability. *The American Journal of Psychiatry*. 2012; 169(5):537. author reply 537–538. <http://doi.org/10.1176/appi.ajp.2012.12010083>. [PubMed: 22549210]

- Watson D, O'Hara MW, Chmielewski M, McDade-Montez EA, Koffel E, Naragon K, Stuart S. Further validation of the IDAS: Evidence of convergent, discriminant, criterion, and incremental validity. *Psychological Assessment*. 2008; 20(3):248–259. <http://doi.org/10.1037/a0012570>. [PubMed: 18778161]
- Watson D, O'Hara MW, Simms LJ, Kotov R, Chmielewski M, McDade-Montez EA, Stuart S. Development and validation of the Inventory of Depression and Anxiety Symptoms (IDAS). *Psychological Assessment*. 2007; 19(3):253–268. <http://doi.org/10.1037/1040-3590.19.3.253>. [PubMed: 17845118]
- Williams JBW, Gibbon M, First M, Spitzer RL, Davies M, Borus J, Wittchen HU. The structured clinical interview for DSM-III-R (SCID): II. multisite test-retest reliability. *Archives of General Psychiatry*. 1992; 49(8):630. <http://doi.org/10.1001/archpsyc.1992.01820080038006>. [PubMed: 1637253]
- Zanarini MC, Skodol AE, Bender D, Dolan R, Sanislow C, Schaefer E, Gunderson JG. The Collaborative Longitudinal Personality Disorders Study: reliability of axis I and II diagnoses. *Journal of Personality Disorders*. 2000; 14(4):291–299. [PubMed: 11213787]
- Zimmerman M. Diagnosing personality disorders: A review of issues and research methods. *Archives of General Psychiatry*. 1994; 51(3):225. <http://doi.org/10.1001/archpsyc.1994.03950030061006>. [PubMed: 8122959]

Table 1

Prevalence of Diagnosed DSM-IV Disorders

Diagnosis	Full sample		Test-retest subsample					
	Time 1		Time 1		Time 2		Time 2	
	N	%	N	%	N	%	N	%
Major depressive disorder	145	42.8	89	40.8	89	40.8	89	40.8
Generalized anxiety disorder	79	23.3	50	23.0	54	24.8	54	24.8
Psychotic disorder	72	21.2	47	21.6	36	16.5	36	16.5
Bipolar I disorder	46	13.6	32	14.7	25	11.5	25	11.5
Dysthymic disorder	46	13.6	29	13.3	23	10.6	23	10.6
Posttraumatic stress disorder	46	13.6	26	11.9	20	9.2	20	9.2
Specific phobia	37	10.9	31	14.2	28	12.9	28	12.9
Social phobia	35	10.3	22	10.1	26	12.0	26	12.0
Panic disorder	33	9.7	17	7.8	24	11.0	24	11.0
Obsessive-compulsive disorder	27	8.0	20	9.2	23	10.6	23	10.6
Substance use disorder	26	7.7	10	4.6	12	5.5	12	5.5
Other bipolar (II or NOS)	19	5.6	12	5.5	15	6.9	15	6.9

Note. Full N = 339. Test-retest N = 218. NOS = Not otherwise specified.

Table 2

Interrater Reliabilities (Kappa) of SCID Diagnoses

Diagnosis	Audio-recording	1-week test-retest
Obsessive-compulsive disorder	1.00 (<i>1.00–1.00</i>)	.41 (<i>.20–.60</i>)
Major depressive disorder	.92 (<i>.80–1.00</i>)	.60 (<i>.49–.71</i>)
Social phobia	.91 (<i>.66–1.00</i>)	.25 (<i>.07–.43</i>)
Posttraumatic stress disorder	.90 (<i>.64–1.00</i>)	.52 (<i>.33–.69</i>)
Panic disorder	.85 (<i>.38–1.00</i>)	.60 (<i>.38–.78</i>)
Psychotic disorder	.82 (<i>.56–1.00</i>)	.60 (<i>.46–.72</i>)
Substance use disorder	.81 (<i>.55–1.00</i>)	.62 (<i>.31–.83</i>)
Dysthymic disorder	.75 (<i>.43–.94</i>)	.22 (<i>.03–.39</i>)
Bipolar I disorder	.73 (<i>.38–1.00</i>)	.58 (<i>.40–.73</i>)
Specific phobia	.73 (<i>.37–1.00</i>)	.54 (<i>.38–.69</i>)
Other bipolar (II or NOS)	.64 (<i>.19–1.00</i>)	.25 (<i>.01–.48</i>)
Generalized anxiety disorder	.55 (<i>.16–.84</i>)	.45 (<i>.29–.58</i>)
Mean	.80	.47

Note. $N = 47–49$ (audio-recording), 217–218 (test-retest). Bootstrapped confidence intervals ($N = 1000$, $CI = 95\%$) in italics. NOS = Not otherwise specified.

Table 3

Diagnostic Agreement/Disagreement (Percents)

Diagnosis	Audio-recording			Test-retest		
	% Both Absent	% Disagree	% Both Present	% Both Absent	% Disagree	% Both Present
Obsessive-compulsive disorder	92	0	8	85	11	5
Major depressive disorder	45	4	51	50	19	31
Social phobia	85	2	13	82	15	4
Posttraumatic stress disorder	88	2	10	85	9	6
Panic disorder	92	2	6	87	7	6
Psychotic disorder	75	6	19	75	12	13
Substance use disorder	86	4	10	93	4	3
Dysthymic disorder	76	8	16	80	17	4
Bipolar I disorder	90	4	6	99	.5	.5
Specific phobia	83	6	10	81	11	8
Other bipolar (II or NOS)	92	4	4	97	3	0
GAD	77	13	11	66	20	14

Note. $N = 47-49$ (audio-recording), 217-218 (test-retest) methods. NOS = Not otherwise specified.