Published in final edited form as:
Psychiatry Res. 2012 April 30; 196(2-3): 302–308. doi:10.1016/j.psychres.2011.12.021.

Test-retest reliability of the proposed DSM-5 eating disorder diagnostic criteria

Robyn Sysko[a,*], **Christina A. Roberto**[a,b,c,d], **Rachel D. Barnes**[c], **Carlos M. Grilo**[b,c], **Evelyn Attia**[a], and **B. Timothy Walsh**[a]

[a]Division of Clinical Therapeutics, New York State Psychiatric Institute and the Department of Psychiatry, College of Physicians and Surgeons of Columbia University, New York, NY, USA

[b]Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA

[c]Department of Psychology, Yale University, New Haven, CT, USA

[d]School of Epidemiology & Public Health, Yale University, New Haven, CT, USA

Abstract

The proposed DSM-5 classification scheme for eating disorders includes both major and minor changes to the existing DSM-IV diagnostic criteria. It is not known what effect these modifications will have on the ability to make reliable diagnoses. Two studies were conducted to evaluate the short-term test-retest reliability of the proposed DSM-5 eating disorder diagnoses: anorexia nervosa, bulimia nervosa, binge eating disorder, and feeding and eating conditions not elsewhere classified. Participants completed two independent telephone interviews with research assessors (n=70 Study 1; n=55 Study 2). Fair to substantial agreements (κ= 0.80 and 0.54) were observed across eating disorder diagnoses in Study 1 and Study 2, respectively. Acceptable rates of agreement were identified for the individual eating disorder diagnoses, including DSM-5 anorexia nervosa (κ's of 0.81 to 0.97), bulimia nervosa (κ=0.84), binge eating disorder (κ's of 0.75 and 0.61), and feeding and eating disorders not elsewhere classified (κ's of 0.70 and 0.46). Further, improved short-term test-retest reliability was noted when using the DSM-5, in comparison to DSM-IV, criteria for binge eating disorder. Thus, these studies found that trained interviewers can reliably diagnose eating disorders using the proposed DSM-5 criteria; however, additional data from general practice settings and community samples are needed.

Keywords

diagnostic reliability; anorexia nervosa; bulimia nervosa; binge eating disorder; feeding and eating disorders not elsewhere classified

1. Introduction

The process of revising the diagnostic criteria for psychiatric disorders is well underway with the publication of the fifth edition of the Diagnostic and Statistical Manual of Mental

Corresponding author: Columbia Center for Eating Disorders, New York State Psychiatric Institute, 1051 Riverside Drive, Unit 98, New York, NY, 10032. Fax=212-543-5607, Telephone=212-543-4296, syskor@nyspi.columbia.edu.

Disorders (DSM-5; www.dsm5.org) scheduled for 2013. On the basis of extensive literature reviews, secondary data analyses, and feedback from mental health professionals, the proposed diagnostic criteria include modifications that attempt to remedy limitations of the previous edition of the DSM (DSM-IV; American Psychiatric Association, 1994). For eating disorders, a significant problem with the DSM-IV classification scheme is the prevalence and the heterogeneity of the eating disorder not otherwise specified category (Fairburn and Bohn, 2005). Although it is a residual diagnosis, rates of eating disorder not otherwise specified can be as high as 50% to 90% of all individuals with eating disorders seeking treatment in routine clinical settings (Ricca et al., 2001; Turner and Bryant-Waugh, 2004; Zimmerman et al., 2008). A range of solutions is offered by the DSM-5 criteria for eating disorders to decrease use of the eating disorder not otherwise specified category and address other diagnostic issues identified by the Eating Disorders Workgroup (Attia and Roberto, 2009; Becker et al., 2009a; Becker et al., 2009b; Keel and Striegel-Moore, 2009; Marcus and Wildes, 2009; Peat et al., 2009; Striegel-Moore et al., 2009; van Hoeken et al., 2009; Walsh and Sysko, 2009; Wilson and Sysko, 2009; Wolfe et al., 2009; Wonderlich et al., 2009).

One major change to the existing DSM-IV eating disorders is recommended: the designation of binge eating disorder as a formal diagnosis. Several modest changes are also proposed, including a change in the frequency of binge eating and/or purging behaviors for the diagnosis of bulimia nervosa, and the elimination of an example provided in DSM-IV to guide clinicians in diagnosing anorexia nervosa (e.g., weight loss leading to maintenance of body weight less than 85% of expected). In addition, the category of eating disorder not otherwise specified is to be replaced by “feeding and eating conditions not elsewhere classified,” with six clinically significant conditions described for individuals who fail to meet criteria for other DSM-5 eating disorders. These conditions include atypical anorexia nervosa, subthreshold bulimia nervosa, subthreshold binge eating disorder, purging disorder, night eating syndrome, and other feeding or eating condition not elsewhere classified; however, due to a lack of available data, these conditions are not designated as disorders and detailed criteria are not provided. A number of other small changes in the language of the diagnostic criteria are recommended for the sake of clarification. At this time, it is not known whether these changes will affect the ability to make reliable eating disorder diagnoses.

Two studies were conducted to examine the short-term test-retest reliability of the proposed DSM-5 eating disorder diagnoses over the telephone with two independent raters using either a semi-structured interview or a structured clinical checklist of criteria and diagnoses. A secondary aim of the first study described below was to evaluate the frequency of DSM-5 eating disorders in comparison to other diagnostic schemes, as it is also not clear what effect alterations to the diagnostic criteria will have on the frequency of individuals receiving an eating disorder diagnosis. We hypothesized that, similar to previous research examining DSM-IV eating disorder diagnoses (e.g., Zanarini et al., 2000; Zanarini and Frankenburg, 2001; Thomas et al., 2010; Lobbestael et al., 2011), trained interviewers (research assistants) and other research staff could reliably diagnose eating disorders using the proposed DSM-5 criteria. Further, we hypothesized that the changes proposed in the DSM-5 would reduce the number of individuals in Study 1 receiving a residual eating disorder diagnosis.

2. Method

Study 1

2.1. Participants—As described previously (Sysko and Walsh, 2011), all individuals calling the Columbia Center for Eating Disorders (CCED) complete a brief interview over the telephone. Participants for the current study included any callers aged 18 or older

expressing interest in receiving clinical treatment for an eating disorder. On the basis of an earlier telephone interview study (Sysko and Walsh, 2011), we anticipated that the diversity of callers to the CCED would allow for an evaluation of reliability across several DSM-5 diagnoses (anorexia nervosa, bulimia nervosa, binge eating disorder, feeding and eating conditions not elsewhere classified). The New York State Psychiatric Institute Institutional Review Board reviewed and approved this study.

2.2. Diagnostic Procedure—Following the completion of the initial telephone interview, a research assistant asked the caller whether he or she would be willing to participate in the current study. If the individual agreed, verbal consent was documented and the research assistant assigned a DSM-5 eating disorder diagnosis from information collected during the call. Between three and seven days later, a different research assistant planned to call the participant to complete the telephone interview again. If the second telephone interview was completed successfully, a second DSM-5 diagnosis was assigned. For both telephone interviews, diagnoses were recorded on forms that required individual criteria for anorexia nervosa, bulimia nervosa, and binge eating disorder to be endorsed along with an overall DSM-5 diagnosis. When callers were given a diagnosis of feeding and eating conditions not elsewhere classified, the research assistants provided a description of the rationale for this decision. Research assistants conducting the second telephone interview completed forms and assigned diagnoses without reviewing any information collected during the first phone interview. Participants who were successfully contacted for both interviews received a $20 gift card by mail.

2.2.1. Telephone Interview: All telephone interviews are conducted by bachelor's level research assistants. Prior to the start of data collection for this study, the research assistants received an orientation to the DSM-5 criteria for eating disorders and the telephone interview format from two of the authors (RS, BTW), and subsequently piloted the research procedures by conducting initial telephone interviews over a two week period (n=16). Following these pilot interviews, the research assistants met with one of the authors (BTW), who provided additional clarification about issues that arose while conducting the interviews, and established guidelines for determining the "current" eating disorder diagnosis (i.e., patients meeting criteria for anorexia nervosa in the prior three months should not be given a bulimia nervosa diagnosis) and evaluating "markedly low weight" for anorexia nervosa (i.e., for adults, body mass index < 18.5 kg/m^2). Common diagnostic questions were also discussed (e.g., when inappropriate compensatory behavior was considered to be "recurrent") on two occasions for approximately 30 minutes while data collection was ongoing.

The following items were assessed as part of the telephone interview: self-reported height and weight; restriction of food intake or other behaviors that might affect body weight; fear of gaining weight; attempts to avoid weight gain; feeling fat; concern about low weight (if applicable); out-of-control eating; purging behaviors (vomiting, laxatives, diuretics, other); concern about shape and weight; and distress and functional impairment related to the eating problem. Additional information regarding body weight over the three months prior to the interview was obtained if the caller reported symptoms consistent with a diagnosis of anorexia nervosa at a body mass index at or above 18.5 kg/m^2. Questions regarding out-of-control eating and concern about shape and weight were modeled on the assessment of objective and subjective bulimic episodes and the importance of shape and weight items, respectively, from the Eating Disorder Examination (EDE-12; Fairburn and Cooper, 1993), a well-established semi-structured investigator-based interview for assessing eating disorder psychopathology (Grilo et al., 2001). An objective bulimic episode is the consumption of an amount of food considered to be large with a sense of loss of control, which is consistent with the DSM-IV and DSM-5 definition of binge eating. For callers endorsing episodes of

out-of-control eating, information was collected about the features associated with binge eating episodes described in the Appendix of DSM-IV (e.g., eating rapidly, eating large amounts of food when not physically hungry, etc.). Distress related to binge eating episodes was evaluated, and an estimate was obtained for the number of times per week objective or subjective bulimic episodes occurred in the month prior to the screening and whether this pattern was consistent in the two prior months. In addition, distress about eating symptoms, and not just current weight, and functional impairment (e.g., the eating problem making it hard for the individual to do their work, take care of things at home, or get along with other people) were also assessed.

2.3. Data Analysis

2.3.1. Statistical Analyses: Means and standard deviations were calculated for continuous demographic measures from the telephone interviews. Independent samples *t*-tests were used to compare callers who completed and failed to complete both telephone interviews on demographic variables and the DSM-5 diagnosis assigned during the first telephone interview. The kappa statistic (κ; Cohen, 1960) and percent agreement were calculated for overall agreement across all eating disorder diagnoses and for DSM-5 anorexia nervosa, including restricting and binge eating/purging subtypes, bulimia nervosa, binge eating disorder, and eating disorder not otherwise specified. Percent agreement for individual diagnoses included only ratings with the reference diagnosis (e.g., for binge eating disorder: telephone interview 1 diagnosis= binge eating disorder, telephone interview 2 diagnosis= bulimia nervosa) and excluded pairs that did not include the reference diagnosis (e.g., telephone interview 1 diagnosis=bulimia nervosa, telephone interview 2 diagnosis=feeding and eating condition not elsewhere classified). In addition, Scott's Pi, the algebraic equivalent to intraclass kappa (Banerjee et al., 1999), was used to further evaluate the reliability of the DSM-5 diagnoses.

To facilitate comparisons across studies, we interpreted κ using two standards from previous research on diagnostic reliability for eating disorders. Studies evaluating diagnostic agreement from structured interviews (Zanarini et al., 2000; Zanarini and Frankenberg, 2001; Lobbestael et al., 2011) used standards described by Fleiss (1981), which considers $\kappa < 0.40$ to be poor, κ of 0.40–0.75 to be fair, and $\kappa > 0.75$ to be excellent. Thomas and colleagues (2010), in comparing diagnoses assigned by clinician and research assessors, applied the standards of Landis and Koch (1977) where κ of 0–0.20 is poor, κ of 0.21–0.40 is fair, κ of 0.41–0.60 is moderate, κ of 0.61–0.80 is substantial, and κ of 0.81 to 1.00 is almost perfect. All analyses were performed in SPSS (SPSS Inc., Chicago, IL), with the exception of Scott's Pi, which was calculated using ReCal2 (http://dfreelon.org/utils/recalfront/recal2/#doc).

2.3.2. Qualitative differences between the first and second telephone interviews: In addition to the aforementioned statistical evaluation of agreement between diagnoses assigned in the first and second telephone interviews, qualitative differences in short-term test-retest reliability were also examined. Reasons for discrepant DSM-5 diagnoses (n=13 if considering anorexia nervosa subtypes; 8 excluding subtypes) were generated by reviewing notes taken by the research assistants during the telephone interviews.

2.3.3. Comparison of the frequency of diagnoses made using DSM-IV, DSM-5, and the Broad Categories for the Diagnosis of Eating Disorders: Comparisons were made between the frequencies of callers classified by one of three diagnostic systems. A range of solutions have been considered to address limitations of the DSM-IV criteria for eating disorders, including most notably the high percentage of individuals receiving a residual diagnosis of eating disorder not otherwise specified. Thus, comparisons of eating disorder

diagnostic frequency were performed across DSM-IV, DSM-5 and a more radical proposal for eating disorders classification (the Broad Categories for the Diagnosis of Eating Disorders; Walsh and Sysko, 2009) to assess the ability of these schemes to reduce the proportion of individuals receiving a residual eating disorder diagnosis. Table 1 summarizes the eating disorder diagnostic categories of DSM-IV, DSM-5, and the Broad Categories for the Diagnosis of Eating Disorders scheme. DSM-5 diagnoses from both telephone interviews were summed for anorexia nervosa, bulimia nervosa, binge eating disorder, or feeding and eating conditions not elsewhere classified. For DSM-IV diagnoses and the Broad Categories for the Diagnosis of Eating disorders scheme (Walsh and Sysko, 2009), notes from only the second phone interview were examined by one of the authors (RS), and callers were classified on the basis of this information alone.

3. Study 1 Results

3.1. Participant Characteristics—A total of 125 initial telephone interviews were conducted between September 2010 and February 2011 at the CCED. Seventy callers (56%) completed a second telephone interview, which was on average 5.1 ± 1.6 days after the first interview (range of 2–9 days). No significant differences were observed for individuals who did (n=70) or did not (n=55) complete a second telephone interview for age [$t(123)=-1.85$, $P=0.07$], gender [$\chi^2(1, n=125)=1.25$, $P=0.26$], body mass index [$t(123)=0.02$, $P=0.99$], or DSM-5 diagnosis assigned during the first telephone interview [$\chi^2(4, n=125)=3.50$, $P=0.48$]. The sample included 64 females (91.4%) and 6 males (8.6%), with a mean age of 31.33 ± 11.11 years (range= 18–73 years) and an average body mass index of 21.58 ± 7.52 kg/m^2 (range= 9.46–52.81 kg/m^2).

3.2. Reliability of overall DSM-5 diagnoses and DSM-5 anorexia nervosa, bulimia nervosa, binge eating disorder, and feeding and eating conditions not elsewhere classified—Across all eating disorder diagnoses, agreement between the first and second telephone interviews was ‘excellent’ or ‘substantial’ ($\kappa= 0.80$; Scott’s Pi= 0.80; 81.4% agreement; see Table 2). As displayed in Table 3, ‘excellent’ or ‘almost perfect’ agreement was also observed for DSM-5 anorexia nervosa restricting type, bulimia nervosa, anorexia nervosa-binge eating purging type, and cases of anorexia nervosa where subtypes were disregarded. ‘Fair’ or ‘substantial’ agreement was noted for DSM-5 binge eating disorder and feeding and eating conditions not elsewhere classified. Agreement, as measured by Scott’s Pi, identical to Cohen’s kappa when rounded to two decimals in all cases.

3.3. Discrepancies between the first and second telephone interviews—A total of six discrepancies involving the diagnosis of anorexia nervosa were noted, with five occurring between DSM-5 anorexia nervosa restricting type and binge purge type. Two cases were inconsistencies in coding, in which the research assistant endorsed the individual criterion of binge eating and/or purging on the caller’s form, but did not subsequently assign a diagnosis that reflected these symptoms. Three other discrepancies were due to confusion over what behaviors are consistent with the binge eating and purging type of anorexia nervosa. Callers reporting loss of control over eating after consuming a small amount of food along with either excessive exercise, fasting, laxative use that was not recurrent or above the recommended dose, or use of Adderral (dextroamphetamine and amphetamine), were incorrectly classified as anorexia nervosa, binge eating-purging type. Finally, one caller was at a normal weight at the time of the interviews (body mass index= 19.6 kg/m^2) as the result of participation in a partial hospitalization program, but in the three months prior to the telephone interviews, her body mass index was 16.3 kg/m^2. On the basis of this information, one rater correctly diagnosed her with anorexia nervosa, restricting type using the guidelines established at the beginning of the study (see Telephone Interview, section

2.2.1. above) and the other rater assigned a diagnosis of feeding and eating conditions not elsewhere classified.

Seven discrepancies related to DSM-5 bulimia nervosa or binge eating disorder were identified. Three callers were diagnosed with DSM-5 bulimia nervosa during one telephone interview and a feeding and eating condition not elsewhere classified in the other. In two of these cases, different frequencies of eating disordered behaviors were reported to the research assistants, such that in the number of episodes of binge eating were less than once per week in one of the interviews, and in the third case, the patient was not able provide enough information in one interview about the amount of food consumed in a typical binge eating episode to determine if the episodes were objectively large. Discrepancies were also observed in two cases where the diagnosis of bulimia nervosa was assigned in one telephone interview and binge eating disorder in the other, which related to one patient reporting her use of laxatives as related only to constipation to one research assistant but not the other, and another patient describing use of one diuretic pill three times weekly, which one research assistant correctly considered to be inappropriate compensatory behavior, but the other did not. Finally, two interviews were differently assigned the diagnoses of DSM-5 binge eating disorder and feeding and eating conditions not elsewhere classified, discrepancies that are explained by different descriptions of eating disordered symptoms from the callers. One patient indicated that she experienced a sense of loss of control over her eating while consuming an objectively large amount of food in one telephone interview, but only while consuming smaller amounts of food in the other interview, and the second caller reported fasting once every other week in one call but not the other, which led one interviewer to describe her symptoms as “subthreshold bulimia nervosa.”

3.4. Frequency of DSM-IV, DSM-5, and the Broad Categories for the Diagnosis of Eating Disorders diagnoses—To identify the frequency of DSM-5 diagnoses among individuals in this sample, six of the aforementioned discrepant interviews were excluded. For these interviews (n=3 discrepant for bulimia nervosa and feeding and eating conditions not elsewhere classified; n=1 discrepant for bulimia nervosa and binge eating disorder; n=2 discrepant for binge eating disorder and feeding and eating conditions not elsewhere classified), it was not possible to evaluate the most accurate DSM-5 diagnosis due to differences in the information reported by the patient in the two calls. Thus, for the remaining 64 callers, a total of 14 (21.9%) were diagnosed with anorexia nervosa, restricting type, 12 (18.8%) with anorexia nervosa binge-eating/purging type, 21 (32.8%) with bulimia nervosa, 8 (12.5%) with binge eating disorder, and 9 (14.1%) with feeding and eating conditions not elsewhere classified.

Among the data obtained from the second telephone interview, a total of 23 (32.9%) patients were diagnosed with DSM-IV anorexia nervosa, including 10 with anorexia nervosa, restricting type (14.3%), and 13 with anorexia nervosa, binge eating-purging type (18.6%). Nineteen callers (27.1%) were assigned a diagnosis of DSM-IV bulimia nervosa, and 28 (40.0%) an eating disorder not otherwise specified. Applying the Broad Categories for the Diagnosis of Eating Disorders scheme (Walsh and Sysko, 2009), a total of 29 individuals (41.4%) were classified in the category of Anorexia Nervosa and Behaviorally Similar Disorders (AN-BSD), 28 (40.0%) as Bulimia Nervosa and Behaviorally Similar Disorders (BN-BSD), 12 (17.1%) as Binge Eating Disorder and Behaviorally Similar Disorders (BED-BSD), and one individual was considered to have an eating disorder not otherwise specified (1.4%).

Within the broad category of anorexia nervosa, 23 (79.3%) individuals were designated as having Typical Anorexia Nervosa, three (10.3%) with Anorexia Nervosa, without Evidence of Distortions Related to Body Shape and Weight, and three (10.3%) as AN-BSD with

Significant Weight Loss at or above a Minimally Acceptable Weight. For the BN-BSD category, 21 (75%) were classified as Typical Bulimia Nervosa, three as Purging Disorder (10.7%), three with Disorders Behaviorally Similar to BN Not Otherwise Classified (10.7%), and one as Bulimia Nervosa, Low Frequency (3.6%). Within the BED-BSD category, all of whom had DSM-IV EDNOS, 10 (83.3%) were classified as Typical Binge Eating Disorder, and two (16.7%) Binge Eating Disorder, Low Frequency. The individual with an eating disorder not otherwise specified diagnosis using the BCD-ED classification reported symptoms consistent with rumination syndrome.

Study 2

4. Method

4.1. Participants—Individuals calling the Program for Obesity, Weight and Eating Research (POWER) at Yale University completed a two-step telephone interview process as part of the present study in addition to also determining their potential eligibility to participate in research studies at the clinic. The telephone screen and intake assessments had IRB approval. Participants for the current study included any callers aged 18 or older expressing interest in receiving clinical treatment for binge eating. The POWER primarily recruits individuals for studies of binge eating and weight loss, and as a result, few individuals with anorexia nervosa or bulimia nervosa contact the clinic. Thus, Study 2 focused primarily on the short-term test-retest reliability of binge eating disorder, a category that is designated as a formal diagnosis for the first time in the proposed DSM-5 classification scheme.

4.2. Diagnostic procedures & telephone interview—The diagnostic procedures for Study 2 were similar to the procedures for Study 1 with modifications detailed below. Unlike Study 1, at the end of each telephone interview, both DSM-IV and DSM-5 diagnoses were recorded. In addition, research staff with a range of training experiences conducted the telephone interviews (11% of the interviews were conducted by an individual with a Master's degree, 44% by doctoral students, and 45% by PhDs). Staff were briefly oriented to the interview and DSM-5 diagnostic criteria, but did not receive additional training on completing the structured clinical checklist interview as in Study 1. Similarly, staff members were provided a checklist of the diagnostic criteria rather than a semi-structured interview as used in Study 1. Participants in this study were not further classified using the Broad Categories for the Diagnosis of Eating Disorders scheme because few people contacted the clinic with symptoms of the other possible ED diagnoses. Finally, participants were not financially compensated for their participation in this study. The same statistical approach was used in Study 1 and Study 2.

5. Study 2 Results

5.1. Participant characteristics—A total of 227 initial telephone inquiries were received between August 2010 and September 2010. Fifty-five of these callers (24%) completed both telephone interviews, which was on average 8.09 ± 7.67 days after the first interview (range of 1–31 days). The sample included 41 females (74.5%), with a mean age of 44.20 ± 11.71 years (range= 18–64 years) and an average body mass index of 38.34 ± 6.76 kg/m^2 (range= 26.0 – 60.0 kg/m^2).

5.2. Reliability of diagnoses

5.2.1. Overall DSM-4 and DSM-5 diagnoses: Across all eating disorder diagnoses, agreement between the first and second telephone interviews for DSM-4 diagnoses was 'poor' or 'fair' (κ= 0.39; Scott's Pi= 0.39; 61.8% agreement; see Table 4). Agreement

improved to ‘fair’ to ‘moderate’ for DSM-5 diagnoses (κ= 0.54; Scott’s Pi= 0.54; 74.5% agreement).

5.2.2. Binge eating disorder, eating disorder not otherwise specified and feeding and eating conditions not elsewhere classified: As displayed in Table 5, ‘poor’ to ‘moderate’ agreement was observed for DSM-IV binge eating disorder, which improved to ‘fair’ to ‘substantial’ when the DSM-5 diagnostic criteria were applied. DSM-IV EDNOS had ‘poor’ reliability, which improved to ‘fair’ to ‘moderate’ reliability when DSM-5 criteria for feeding and eating conditions not elsewhere classified were applied. Agreement, as measured by Scott’s Pi, was identical to Cohen’s kappa when rounded to two decimals in all cases. The agreement for bulimia nervosa was not calculated given the small number of cases identified.

5.3. Discrepancies between the first and second telephone interviews—Sixteen discrepancies related to DSM-IV binge eating disorder were identified. Of the 40 callers diagnosed with DSM-IV binge eating disorder in one of the two calls, three were diagnosed with DSM-IV bulimia nervosa, eleven were diagnosed with a DSM-IV eating disorder not otherwise specified, and two were given no diagnosis in the other interview. In eight cases, there was disagreement when determining whether the frequency or duration of reported objective bulimic episodes met the appropriate threshold for a diagnosis of binge eating disorder. In five cases, an individual reported engaging in inappropriate compensatory behaviors to one interviewer, but not the other. In one case, there was disagreement when determining whether the amount of food was large enough to constitute a binge. In this instance, the first rater incorrectly classified the amount of food as “unusually large.” In two cases there was a discrepancy in rating whether the individual experienced a loss of control. In both instances, the individual denied a loss of control during the second phone call, after having previously described experiencing a loss of control in the first interview.

Ten discrepancies related to DSM-5 binge eating disorder were identified. Of the callers diagnosed with DSM-5 binge eating disorder in one interview, four were diagnosed with DSM-5 bulimia nervosa, three were diagnosed with DSM-5 feeding and eating conditions not elsewhere classified, and three were not given a diagnosis in the other interview. In six cases, an individual reported engaging in inappropriate compensatory behaviors to one interviewer, but not the other. In two cases, there was a discrepancy in considering the amount of food large enough to constitute a binge. In one instance, the first rater incorrectly classified an amount of food as “unusually large.” In the second instance, the individual reported a smaller amount of food when describing a binge episode during the second call. In two cases there was a discrepancy in rating whether the individual experienced a loss of control. In both instances, the patient denied experiencing a loss of control during the second phone interview, after having previously endorsed feeling out of control in the first interview.

6. General Discussion

Two studies examined the short-term test-retest reliability of the proposed DSM-5 eating disorder categories utilizing brief semi-structured or structured telephone interviews performed by trained research assistant interviewers. Across eating disorder diagnoses, ‘fair’ to ‘substantial’ or ‘excellent’ agreement was observed. When considering individual DSM-5 proposed eating disorders, ‘fair’ to ‘almost perfect’ diagnostic agreement was noted. Thus, the modified DSM-5 criteria for eating disorders produce acceptable short-term test-retest reliability over the telephone. Despite notable differences in study design, telephone interview format, and classification scheme, we found similar overall rates of test-retest reliability across eating disorder diagnoses (κ of 0.80 and 0.54, respectively) as a test-retest

reliability study of two semi-structured interviews for DSM-IV (κ= 0.64; Zanarini et al., 2000). Our results are also comparable to the average kappa for eating disorders (κ = 0.70) reported in a meta-analysis of studies evaluating diagnostic agreement between clinical and structured interviews (Rettew et al., 2009).

Among specific eating disorder diagnoses, a parallel pattern emerged. The range of short-term test-retest reliabilities identified in Study 1 for anorexia nervosa, bulimia nervosa, and feeding and eating conditions not elsewhere classified (κ's of 0.46 to 0.97) are similar to studies examining concordance (test-retest or inter-rater reliability) between eating disorder diagnoses with two face-to-face diagnostic interviews (structured or clinical; κ's of 0.58 to 1.0, Zanarini and Frankenburg, 2001; Andreas et al., 2009; Thomas et al., 2010). While patients with binge eating disorder may have been included in previous research among groups with a DSM-IV eating disorder not otherwise specified (e.g., Thomas et al., 2010; Zanarini & Frankenberg, 2001), specific kappas for the test-retest reliability of this provisional DSM-IV diagnosis are not available. Variability in kappa values was noted in our two studies across DSM-5 eating disorder diagnoses; however, despite the notable range, extant data suggests that differences in reliability of a similar magnitude to those found in the current studies are likely to be expected. For studies of the diagnostic reliability of clinical or structured interviews, Rettew and colleagues (2009) reported a range of kappas from −0.01 to 1.00 for bulimia nervosa. In addition, studies examining test-retest or inter-rater reliabilities of other psychiatric conditions such as major depression using two structured interviews, or one structured and one clinical interview, also report significant variability (κ's ranging from 0.27 to 0.90; Zanarini et al., 2000; Zanarini and Frankenburg, 2001; Andreas et al., 2009; Lobbestael et al., 2011).

While previous studies have not examined the test-retest reliability of DSM-IV or DSM-5 eating disorder diagnoses using telephone interview methods, telephone and face-to-face diagnostic interviews appear to be equivalent in the assessment of at least some psychiatric disorders (e.g., anxiety disorders, major depressive disorder, Rohde et al., 1997; psychotic disorders, Hajebi et al., in press). However, caution is needed when comparing the agreement observed in Studies 1 and 2 to previous research using face-to-face assessments. Rates of diagnostic reliability can differ depending on the type of face-to-face diagnostic interview used (clinical versus structured interview; Rettew et al., 2009; Zimmerman and Mattia, 1999), which further complicates direct evaluations between these studies and extant data. Lastly, studies of semi-structured diagnostic interviews for other forms of psychopathology (Zanarini et al., 2000) and for eating disorder psychopathology (Grilo et al., 2004) report both less than perfect stability over time and test-retest reliability.

The qualitative review of discrepancies between telephone interviews suggests that for the purpose of research studies more specific guidance for certain criteria may be useful to increase the reliability of ED diagnoses. In particular, the time course of low weight (e.g., at any point in the last three months) and a frequency of "recurrent" binge eating and purging behaviors could help to distinguish between subtypes of anorexia nervosa and bulimia nervosa and binge eating disorder. In studies of DSM-IV eating disorders, variable frequencies of binge eating and/or purging have been used to differentiate the subtypes of anorexia nervosa, from weekly to monthly episodes (Thomas et al., 2010), which could lead to conflicting diagnostic assignments. Therefore, future research on the DSM-5 categories should aim to identify a clinically meaningful threshold for these behaviors.

In addition, Study 2 suggests that improved short-term test-retest reliability may be achieved by using the DSM-5, in comparison to DSM-IV, criteria for binge eating disorder. The Appendix of DSM-IV specifies that for the diagnosis of binge eating disorder, an individual should experience a minimum average of two binge days, not binge episodes, per week over

the prior six months. In DSM-5, the criteria specify that binge eating episodes are to occur at least once weekly over a three month period. In Study 2, the DSM-IV criterion for the frequency of binge eating yielded the greatest number of discrepancies. In contrast, when DSM-5 criteria were used, there were no discrepancies based on binge eating frequency, suggesting that requiring a minimum of once weekly episodes improves the reliability of the diagnosis. This finding is consistent with a review by Wilson and Sysko (2009), which identified limited evidence of the validity or utility of the twice-weekly DSM-IV frequency criterion for binge eating disorder.

Study 1 also supported our hypothesis that the proposed DSM-5 eating disorder categories reduce the proportion of individuals receiving a residual diagnosis, which was a significant problem with the existing DSM-IV classification scheme. In particular, 40% of the Study 1 participants were assigned a DSM-IV diagnosis of eating disorder not otherwise classified, but only 14% were considered to be in the feeding and eating conditions not elsewhere classified by DSM-5. However, a more substantial reduction in residual eating disorder diagnosis, to 1.4%, was achieved by applying the Broad Categories for the Diagnosis of Eating Disorders scheme (Walsh and Sysko, 2009). As observed in previous research (Sysko and Walsh, 2011), applying this scheme has the potential to virtually eliminate the use of residual eating disorder diagnoses. However, there are potential disadvantages to adopting this form of classification, including limited extant data, especially with regard to using the scheme with community samples, and problems with interpreting existing information on course, outcomes, or treatment response because of greater heterogeneity introduced to the eating disorder categories by this scheme.

There are limitations to this research. Although these studies suggest that the DSM-5 classification scheme can be applied reliably, including by non-clinician interviewers, the research was conducted in two tertiary care centers with adult participants. It is therefore possible that achieving concordance between raters will be more difficult when the revised eating disorder criteria are used with community samples, individuals who seek treatment outside of specialist programs, or younger populations. In addition, the staff of specialty programs are quite familiar with the assessment of eating disorder symptoms and the existing diagnostic criteria in comparison to general practice settings or primary care clinics, factors that could influence the likelihood of concordant diagnoses (Zimmerman et al., 2008). Further, our interview was administered by phone, which limits the ability to compare these studies with prior research. It is possible that the rates of reliability obtained in these studies are different from what would have been observed with two face-to-face interviews, or if the interviews were conducted over a longer time period. The short lag between interviews and the potential for symptom fluctuation during this time, changes in patient report over two interviews, and differences between raters could account, in part, for the observed reliability estimates. Our telephone interviews (semi-structured or structured clinical checklist) assessed eating disorder symptoms among individuals interested in receiving treatment specifically for eating/weight concerns. In addition, only one quarter of the participants in Study 2 could be re-contacted for a second interview. This low response rate may relate to the lack of compensation for the second interview and suggests a potential selection bias by including only participants willing to complete a second interview. Overall, the assessment and recruitment methodology might yield different reliability estimates than observed with unstructured clinical interviews in general practice settings or semi-structured diagnostic interviews in specialty eating disorder clinics. Different assessment methods for eating disorders each vary with regard to their psychometric properties and advantages and disadvantages (Grilo et al., 2001). Moreover, neither study examined the reliability of the proposed DSM-5 categories of Pica, Rumination Disorder, or Avoidant/Restrictive Food Intake Disorder, and only Study 2 distinguished between case and non-case status, and only a small number of individuals without an eating disorder diagnosis (n=5) were identified.

Finally, a range of reliability estimates were observed in these studies and it is not possible to determine whether this variability reflected interpretation of the DSM-5 criteria across centers or methodological factors. For example, the CCED receives calls from a diverse group of individuals experiencing a wide range of problems with eating, whereas the POWER primarily recruits for studies of binge eating and/or weight loss. Thus, the lower diagnostic agreement at POWER may reflect difficulty distinguishing between episodes of binge eating and other forms of overeating within a group composed primarily of obese individuals. Alternatively, observed differences in reliability might be due to site-specific variation in interviewer education, training and study supervision, or assessment methods.

In conclusion, these studies provide important initial evidence that the proposed DSM-5 eating disorder diagnoses have good reliability for anorexia and bulimia nervosa and acceptable reliability for binge eating disorder and feeding and eating disorders not elsewhere classified. Future research should evaluate the reliability of DSM-5 diagnoses in community samples of adults, children and adolescents as well as those presenting to primary care settings. Although an increased focus on diagnostic issues in the literature has resulted from the impending publication of DSM-5, it is important that research evaluating the revised eating disorders criteria continue after the release of the DSM-5 to assess the reliability and validity (Grilo and White, 2011; White and Grilo, 2011; Wolfe et al., 2009) of the new eating disorder criteria sets and diagnoses in addition to the clinical utility of the revised classification scheme.

Acknowledgments

We would like to acknowledge the assistance of the Columbia Center for Eating Disorders research assistants, including: Staci Berkowitz, Stephanie Brewer, Leora David, Zoe Grunebaum, Margaret Martinez, Rachel Ojserkis, and Molly Siegel, and the Program for Obesity, Weight, and Eating Research staff, including: Kerstin Blomquist, Ph.D., Abbe Boeke, Ph.D., Stacey Fruman, M.S., Sylvia Herbozo, Ph.D., Siddhi Shah, Jennifer Weinstein, M.A. Dr. Sysko is supported by National Institute of Diabetes and Digestive and Kidney Diseases grant K23 DK088532 and Dr. Grilo is supported by National Institute of Diabetes and Digestive and Kidney Diseases grant K24 DK070052.

References

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. 4. Washington, DC: American Psychiatric Association; 1994. Revised

Andreas S, Theisen P, Mestel R, Koch U, Schulz H. Validity of routine clinical DSM-IV diagnoses in inpatients with mental disorders. Psychiatry Research. 2009; 170:252–255. [PubMed: 19896721]

Attia E, Roberto CA. Should amenorrhea be a diagnostic criterion for anorexia nervosa? International Journal of Eating Disorders. 2009; 42:581–589. [PubMed: 19621464]

Banerjee M, Capozzoli M, McSweeney L, Sinha D. Beyond kappa: A review of inter-rater agreement measures. The Canadian Journal of Statistics. 1999; 27:3–23.

Becker AE, Eddy KT, Perloe A. Clarifying criteria for cognitive signs and symptoms for eating disorders in DSM-V. International Journal of Eating Disorders. 2009a; 42:611–619. [PubMed: 19650082]

Becker AE, Thomas JJ, Pike KM. Should non-fat-phobic anorexia nervosa be included in DSM-V? International Journal of Eating Disorders. 2009b; 42:620–635. [PubMed: 19655370]

Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960; 20:37–46.

Fairburn CG, Bohn K. Eating disorder NOS (EDNOS): an example of the troublesome 'Not Otherwise Specified' (NOS) category in DSM-IV. Behaviour Research and Therapy. 2005; 43:691–701. [PubMed: 15890163]

Fairburn, CG.; Cooper, PJ. The Eating Disorder Examination. In: Fairburn, CG.; Wilson, GT., editors. Binge eating: Nature, assessment, and treatment. The Guilford Press; New York: 1993. p. 317-360.

Fleiss, JL. Statistical methods for rates and proportions. Wiley; New York: 1981.

Grilo CM, Masheb RM, Lozano-Blanco C, Barry DT. Reliability of the eating disorder examination in patients with binge eating disorder. International Journal of Eating Disorders. 2004; 35:80–85. [PubMed: 14705160]

Grilo CM, Masheb RM, Wilson GT. A comparison of different methods for assessing the features of eating disorders in patients with binge eating disorder. Journal of Consulting and Clinical Psychology. 2001; 69:317–322. [PubMed: 11393608]

Grilo CM, White MA. A controlled evaluation of the distress criterion for binge eating disorder. Journal of Consulting and Clinical Psychology. 2011; 79:509–514. [PubMed: 21707133]

Hajebi A, Motevalian A, Amin-Esmaeili M, Hefazi M, Radgoodarzi R, Rahimi-Movaghar A, Sharifi V. Telephone versus face-to-face administration of the Structured Clinical Interview for Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, for diagnosis of psychotic disorders. Comprehensive Psychiatry. in press.

Keel PK, Striegel-Moore RH. The validity and clinical utility of purging disorder. International Journal of Eating Disorders. 2009; 42:706–719. [PubMed: 19642215]

Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33:159–174. [PubMed: 843571]

Lobbestael J, Leurgans M, Arntz A. Inter-rater reliability of the Structured Clinical Interviwe for DSM-IV Axis I Disorders (SCID I) and Axis II Disorders (SCID II). Clinical Psychology and Psychotherapy. 2011; 18:75–79. [PubMed: 20309842]

Marcus MD, Wildes JE. Obesity: Is it a mental disorder? International Journal of Eating Disorders. 2009; 42:739–753. [PubMed: 19610015]

Peat C, Mitchell JE, Hoek HW, Wonderlich SA. Validity and utility of subtyping anorexia nervosa. International Journal of Eating Disorders. 2009; 42:590–594. [PubMed: 19598270]

Rettew DC, Lynch AD, Achenbach TM, Dumenci L, Ivanova MY. Meta-analysis of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. International Journal of Methods in Psychiatric Research. 2009; 18:169–184. [PubMed: 19701924]

Rohde P, Lewinsohn PM, Seeley JR. Comparability of telephone and face-to-face interviews in assessing Axis I and II disorders. American Journal of Psychiatry. 1997; 154:1593–1598. [PubMed: 9356570]

Ricca V, Mannucci E, Mezzani B, Di Bernardo M, Zucchi T, Paionni A, Placidi GP, Rotella CM, Faravelli C. Psychopathological and clinical features of outpatients with an eating disorder not otherwise specified. Eating and Weight Disorders. 2001; 6:157–165. [PubMed: 11589418]

Striegel-Moore RH, Franko DL, Garcia J. The validity and clinical utility of night eating syndrome. International Journal of Eating Disorders. 2009; 42:720–738. [PubMed: 19621465]

Sysko R, Walsh BT. Does the broad categories for the diagnosis of eating disorders (BCD-ED) scheme reduce the frequency of eating disorder not otherwise specified? International Journal of Eating Disorders. 2011; 44:625–629. [PubMed: 21997426]

Thomas JJ, Delinsky SS, St Germain SA, Weigel TJ, Tangren CM, Levendusky PG, Becker AE. How do eating disorder specialist clinicians apply DSM-IV diagnostic criteria in routine clinical practice? Implications for enhancing clinical utility in DSM-5. Psychiatry Research. 2010; 178:511–517. [PubMed: 20591498]

Turner H, Bryant Waugh R. Eating disorder not otherwise specified (EDNOS): Profiles of clients presenting at a community eating disorder service. European Eating Disorders Review. 2004; 12:18–26.

van Hoeken D, Veling W, Sinke S, Mitchell JE, Hoek HW. The validity and utility of subtyping bulimia nervosa. International Journal of Eating Disorders. 2009; 42:595–602. [PubMed: 19621467]

Walsh BT, Sysko R. Broad categories for the diagnosis of eating disorders (BCD-ED): an alternative system for classification. International Journal of Eating Disorders. 2009; 42:754–764. [PubMed: 19650083]

White MA, Grilo CM. Diagnostic efficiency of DSM-IV indicators for binge eating episodes. Journal of Consulting and Clinical Psychology. 2011; 79:75–83. [PubMed: 21261436]

Wilson GT, Sysko R. Frequency of binge eating episodes in bulimia nervosa and binge eating disorder: Diagnostic considerations. International Journal of Eating Disorders. 2009; 42:603–610. [PubMed: 19610014]

Wolfe BE, Baker CW, Smith AT, Kelly-Weeder S. Validity and utility of the current definition of binge eating. International Journal of Eating Disorders. 2009; 42:674–686. [PubMed: 19610126]

Wonderlich SA, Gordon KH, Mitchell JE, Crosby RD, Engel SG. The validity and clinical utility of binge eating disorder. International Journal of Eating Disorders. 2009; 42:687–705. [PubMed: 19621466]

Zanarini MC, Frankenburg FR. Attainment and maintenance of reliability of Axis I and II disorders over the course of a longitudinal study. Comprehensive Psychiatry. 2001; 42:369–374. [PubMed: 11559863]

Zanarini MC, Skodol AE, Bender D, Dolan R, Sanislow C, Schaefer E, Morey LC, Grilo CM, Shea MT, McGlashan TH, Gunderson JG. The collaborative longitudinal personality disorders study: Reliability of Axis I and II diagnoses. Journal of Personality Disorders. 2000; 14:291–299. [PubMed: 11213787]

Zimmerman M, Francione-Witt C, Chelminski I, Young D, Tortolani C. Problems applying the DSM-IV eating disorders diagnostic criteria in a general psychiatric outpatient practice. Journal of Clinical Psychiatry. 2008; 69:381–384. [PubMed: 18348598]

Zimmerman M, Mattia JI. Differences between clinical and research practices in diagnosing borderline personality disorder. American Journal of Psychiatry. 1999; 156:1570–1574. [PubMed: 10518168]

Table 1

Summary of the diagnostic and proposed criteria for eating disorders from the fourth and fifth editions of the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV; DSM-5), respectively, and the Broad Categories for the Diagnosis of Eating Disorders (BCD-ED) scheme

	DSM-IV	DSM-5	BCD-ED
Anorexia Nervosa[1]			
A. Low body weight	✓	✓	✓
B. Fear of gaining weight or becoming fat	✓	✓	
C. Disturbance in experience of body weight or shape	✓	✓	
D. Amenorrhea	✓		
Evidence of behavioral resistance to gaining weight			✓
Distress or functional impairment related to the eating disorder			✓
Bulimia Nervosa			
A. Recurrent episodes of binge eating	✓	✓	✓
B. Recurrent inappropriate compensatory behavior (e.g., self-induced vomiting, laxatives, fasting) to prevent weight gain	✓	✓	✓
C. Frequency of binge eating and inappropriate compensatory behavior over a three-month period	Twice-weekly	Once-weekly	Recurrent
D. Undue influence of body shape and weight on self-evaluation	✓	✓	
E. Does not meet criteria for anorexia nervosa	✓	✓	✓
Distress or functional impairment related to the eating disorder			✓
Binge Eating Disorder[2]			
A. Recurrent episodes of binge eating	✓	✓	✓
B. Binge-eating episodes associated at least three of the following:			
1. eating much more rapidly than normal			
2. eating until feeling uncomfortably full	✓	✓	
3. eating large amounts of food when not physically hungry			
4. eating alone because of being embarrassment			
5. feeling disgusted, depressed, or very guilty after overeating			
C. Marked distress regarding binge eating	✓	✓	
D. Frequency of binge eating	Twice-weekly for six months	Once-weekly for three months	Recurrent for three months
E. Does not meet criteria for anorexia nervosa or bulimia nervosa	✓	✓	✓
Distress or functional impairment related to the eating disorder			✓

Notes.

[1] Text describing the DSM-5 criteria for anorexia nervosa have been revised for additional clarification; see www.dsm5.org for additional detail, and descriptions of a low body weight differ across the classification schemes.

[2] Criteria for binge eating disorder are listed in the Appendix of DSM-IV as an example of an eating disorder not otherwise specified and a provisional category in need of additional study, but in the proposed DSM-5 scheme, binge eating disorder is classified as a separate eating disorder category.

Table 2

DSM-5 eating disorder diagnoses assigned in the first and second phone interviews (Study 1)

		DSM-5 diagnosis assigned in the second phone interview				
		AN-R	**AN-BEP**	**BN**	**BED**	**FECNEC**
DSM-5 diagnosis assigned in the first phone interview	AN-R	11	1			1
	AN-BEP	2	11			
	BN			21		2
	BED			2	7	2
	FECNEC			1		9

Note. Shaded boxes identify diagnostic agreement between the first and second calls. AN-R= anorexia nervosa, restricting subtype; AN-BEP= anorexia nervosa, binge-eating/purging type; BN=bulimia nervosa; BED=binge eating disorder; DSM-5 = Diagnostic and Statistical Manual of Mental Disorders, 5th edition; FECNEC= feeding and eating conditions not elsewhere classified

Table 3

Test-retest reliability within DSM-5 eating disorder categories (Study 1)

DSM-5 diagnosis	Percent agreement	Cohen's kappa
Anorexia nervosa, restricting subtype	62.5%	0.81
Anorexia nervosa, binge-eating/purging subtype	66.7%	0.85
Anorexia nervosa (disregarding subtypes)	96.1%	0.97
Bulimia nervosa	80.8%	0.84
Binge eating disorder	63.6%	0.75
Feeding and eating conditions not elsewhere classified	60.0%	0.70

Table 4

DSM-4 and DSM-5 eating disorder diagnoses assigned in the first and second phone interviews (Study 2)

		DSM-4 diagnosis assigned in the second phone interview			
		BED	**BN**	**EDNOS**	**No Dx**
DSM-4 diagnosis assigned in the first phone interview	BED	21		7	2
	BN	3	1		
	EDNOS	4	1	5	2
	No Dx			2	5

		DSM-5 diagnosis assigned in the second phone interview			
		BED	**BN**	**FECNEC**	**No Dx**
DSM-5 diagnosis assigned in the first phone interview	BED	30	1	2	3
	BN	3	2	1	
	FECNEC	1		4	1
	No Dx			2	5

Note. Shaded boxes identify diagnostic agreement between the first and second calls. BN=bulimia nervosa; BED=binge eating disorder; DSM-5 = Diagnostic and Statistical Manual of Mental Disorders, 5th edition; FECNEC= feeding and eating conditions not elsewhere classified; No Dx= no eating disorder diagnosis assigned

Table 5

Inter-Rater Reliability for Binge Eating Disorder, Bulimia Nervosa and Eating Disorder Not Otherwise Specified (Study 2)

Diagnosis	DSM-IV percent agreement	DSM-IV Cohen's kappa	DSM-5 percent agreement	DSM-5 Cohen's kappa
Binge eating disorder	70.9%	0.42	81.8%	0.61
Residual eating disorder diagnosis[†]	70.9%	0.27	87.3%	0.46

[†]DSM-IV=eating disorder not otherwise specified; DSM-5=feeding and eating conditions not elsewhere classified