

Assessment of eating disorders Comparison of interview and questionnaire data from a long-term follow-up study of bulimia nervosa

Pamela K. Keel*, Scott Crow, Traci L. Davis, James E. Mitchell

*Departments of Psychology and Psychiatry, Harvard University, Cambridge, University of Minnesota, Minneapolis,
and University of North Dakota and Neuropsychiatric Research Institute, Fargo, USA*

Abstract

Objective: This paper examines diagnostic agreement between interview and questionnaire assessments of women participating in a long-term follow-up study of bulimia nervosa. **Methods:** Women ($N=162$) completed follow-up evaluations comprising questionnaires and either face-to-face or telephone interviews. **Results:** Consistent with previous research, rates of eating disorders were higher when assessed by questionnaire than when assessed by interview; however, rates of full bulimia nervosa were similar. Overall diagnostic agreement was adequate for eating disorders

($\kappa=.64$) but poor for bulimia nervosa ($\kappa=.49$), with greater agreement between questionnaires and telephone interviews (κ 's range: .67–.71) than between questionnaires and face-to-face interviews (κ 's range: .35–.58). **Conclusion:** Findings support the possibility that increased rates of eating pathology on questionnaire assessments may be due, in part, to increased candor when participants feel more anonymous. Questionnaire assessments may not be inferior to interview assessments; they may reveal different aspects of disordered eating. © 2002 Elsevier Science Inc. All rights reserved.

Keywords: Eating disorders; Bulimia nervosa; Assessment

Introduction

Because eating disorders are associated with lifetime prevalence estimates between 0.5% and 3.0% of adolescent and young adult women [1], community-based studies of eating disorders require large samples in order to ensure adequate power for analyses. This is particularly true for studies wishing to evaluate eating disorders in males as well as females, as only 10% of anorexia and bulimia nervosa occur in men [1]. However, it is cumbersome and expensive to conduct individual clinical interviews with large samples. Thus, many large (e.g., >1000 participants) community-based studies of eating pathology in females and males have relied on questionnaire assessments [2–7], although there are some exceptions [8].

Questionnaire assessments of eating pathology have been designed to reflect the DSM diagnostic criteria for eating

disorders [9–13]. Comparison of such surveys and interview data suggests that these surveys demonstrate high sensitivity (few false negatives) but lower specificity (greater false positives) [10,12]. French and colleagues [10] utilized a survey of disordered eating and dieting in a school-based sample of adolescents and then conducted a semistructured expert interview of eating pathology in a subsample of 43 students. They used the percentage of all cases that were identified by interviews and detected by surveys (detection fraction) as a measure of survey sensitivity and they used the percentage of all cases identified by survey and confirmed by interviews (confirmation fraction) as a measure of survey specificity. The range for detection fractions was 69.2–100% across the following behaviors: dieting, vomiting, diet pills, laxatives, fasting and binge eating, indicating that a very high percentage of cases identified by interview were detected by the self-report survey. Conversely, the range for confirmation fractions was 13.6–66.6% for the following behaviors: dieting, vomiting, diet pills, fasting and binge eating, indicating only moderate confirmation of self-reported behaviors during the interview. In a study utilizing a similar survey

* Corresponding author. 1320 William James Hall, 33 Kirkland Street, Cambridge, MA 02138, USA. Tel.: +1-617-495-5592; fax: +1-617-495-4990.

E-mail address: pkeel@wjh.harvard.edu (P.K. Keel).

measure, Leon et al. [12] found that self-report surveys utilized to diagnose eating disorders (anorexia nervosa, bulimia nervosa and EDNOS) demonstrated a 95% detection fraction and a 53% confirmation fraction when compared to semistructured interviews. A combination of high sensitivity and lower specificity would explain the increased prevalence of eating disorders when assessed by questionnaire vs. interview [14].

Factors that may increase false positives on surveys include failure to ask questions concerning necessary diagnostic criteria, confusing questions or lay definitions that differ from clinical definitions and are over-inclusive. For example, women from the community described 42% of their subjective binge episodes (assessed by the Eating Disorders Examination clinical interview) as “binge episodes” even though these would not meet DSM-IV criteria for a binge episode [15]. Unlike questionnaires, interviews provide additional opportunities to clarify questions and ensure the application of clinical definitions. However, explanations of low questionnaire specificity are predicated on the assumption that clinical interviews are superior to questionnaire assessments. French et al. [10] posited that participants may provide more candid responses on questionnaires because they feel greater anonymity. This could explain the increased proportion of individuals reporting disordered eating on surveys than in interviews. If this were true, then data from previous studies [10,12] could be interpreted as representing poor sensitivity in interviews rather than poor specificity of questionnaires. French et al. [10] concluded that more research is required to understand whether respondents over-report disordered eating on questionnaires or whether they are “less willing to disclose such practices in a private interview” (p. 45).

The purpose of this study was to compare diagnoses of bulimia nervosa and eating disorders based on questionnaire and interview assessments in a large sample of women diagnosed with bulimia nervosa more than ten years previously. This sample provided a useful group to study because rates of eating pathology would be higher than in a community-based sample and greater balance in the distribution of the dependent variable increases statistical power. Further, interview assessments were conducted both in person and over the telephone for this sample. Mitchell et al. [16] suggested that participants may be more willing to report eating problems over the phone than in a face-to-face interviews. Thus, two levels of perceived anonymity (face-to-face vs. telephone) were present for interview assessments. All questionnaires were completed in private within the same month of interview assessment. If feelings of embarrassment or shame reduce reporting of eating disorder symptoms in an interview compared to a questionnaire, then we would predict a higher level of agreement between questionnaires and telephone interviews and than between questionnaires and face-to-face interviews.

Methods

Subjects

Subjects for the present study were initially evaluated in the University of Minnesota’s Eating Disorders Clinic and diagnosed with bulimia nervosa between 1981 and 1987 as part of one of two studies [16,17]. Follow-up assessments occurred more than a decade after presentation (mean [S.D.] length of follow-up = 11.5 [1.9] years). At baseline, all subjects were required to meet DSM-III criteria for bulimia, with the additional criterion of binge eating coupled with self-induced vomiting or laxative abuse at a minimum frequency of three times each week for 6 months before evaluation. Additional baseline inclusion and exclusion criteria for these subjects are presented in the original papers [16,17]. Of the 222 subjects sought for participation, 22 (9.9%) could not be located, 1 (0.5%) was deceased, 1 (0.5%) was severely disabled and 21 (9.5%) either declined participation or did not complete participation before data collection ended. Thus, 177 women participated in follow-up assessments representing 80.5% of the total sample excepting those individuals unable to participate due to death or disability. The ascertainment rate was 90.1%. Subjects who participated did not differ from subjects who did not participate on any baseline variables (specific results are reported elsewhere [18]). Mean age was 35.33 years (S.D. = 5.14). The sample was predominantly Caucasian (99%) with only two non-Caucasian participants. Due to missing questionnaire data, diagnoses could be generated from both interview and questionnaire data for only 162 women. Thus, for the purposes of comparison, all results are reported for these 162 women.

Procedure

Subjects completed questionnaires at home and participated in interviews that were conducted either in person or by telephone. Written informed consent was obtained from each subject prior to the interview at the time questionnaires were received. All interviews were conducted in private by the first author or research assistant trained using the DSM-IV SCID training tapes and supervision was available from a licensed clinical psychologist. Interviews were audiotaped to determine reliability. Among the 162 participants for the current study, face-to-face interviews were conducted with 92 (57%) subjects and 70 (43%) completed telephone interviews. Among individuals selecting telephone interviews, 74% lived either out of state or more than a 3-h drive from the research office. Analyses of eating disorder diagnostic status (EDDS) at follow-up, as determined from interviews, revealed no significant difference between face-to-face and telephone interviews ($\chi^2 = 1.51$, $df = 1$, $P = .22$). Additionally, no significant differences were found between face-to-face

and telephone interviews for current diagnoses of affective, anxiety, substance use or impulse control disorders (P -values ranged from .19 to .44).

Measures

At follow-up assessment, subjects participated in Structured Clinical Interview for DSM-IV, Axis I disorders (SCID-I) [19] and Hamilton Depression Rating Scale Interviews (HDRS) [20]. Subjects also completed a series of questionnaire assessments: Eating Disorders Questionnaire (EDQ) [21], Body Shape Questionnaire (BSQ) [22], Multi-dimensional Personality Questionnaire—Scale 8: Control/Impulsiveness (MPQ) [23] and Weissman's Social Adjustment Scale—Self-Report (SAS-SR) [24]. Reliability estimates across measures were generally high (Chronbach's α ranged from $r = .82$ to $r = .98$ for questionnaire assessments and κ 's ranged from .73 to 1.00 for interview assessments), with the exception of the SAS-SR subscales. In the current sample, Chronbach's α for the SAS-SR subscales were: work $r = .77$, social/leisure $r = .78$, extended family $r = .67$, marital $r = .74$, parental $r = .14$ and family unit $r = .60$.

For both interview and questionnaire data, eating disorders in the month of assessment were defined in two ways: presence vs. absence of full bulimia nervosa and presence vs. absence of any eating disorder (including anorexia nervosa, bulimia nervosa, binge eating disorder and other EDNOS) as used in analyses for previous papers from this study [18,25]. Thresholds for diagnoses were held constant between methods of assessment. DSM-IV criteria were employed for a diagnosis of full bulimia nervosa. As described previously [18], the minimum frequency of dis-

ordered eating behaviors for diagnosis of EDNOS was an average of once per month for 3 months.

Data were analyzed using SPSS for Macintosh. Two sets of analyses were planned. The first set of analyses compared diagnostic agreement between questionnaire and interview assessments using the κ statistic, first for full bulimia nervosa and then for any eating disorder. Then diagnostic agreement between questionnaire assessments and face-to-face interviews was compared to diagnostic agreement between questionnaire assessments and telephone interviews. The second set of analyses evaluated the validity of EDDS (defined as the presence vs. absence of any eating disorder during the month of follow-up) by comparing associations between EDDS based on questionnaires vs. interviews and concurrent measures of Axis I disorders, depression, body dissatisfaction, impulsivity and social adjustment. Responses to at least 80% of scale items were required for the subject's data to be included in analyses. Therefore, sample sizes for specific measures vary. When at least 80% but fewer than 100% of items were present, scale scores were prorated according to the following equation (prorated scale score = number of items \times old scale score \div number of responses). Thresholds for statistical significance were set at $\alpha < .05$.

Results

Rates of eating disorders and diagnostic agreement

Based on questionnaire assessment, 22 women (13.6%) met DSM-IV criteria for bulimia nervosa at follow-up and

Table 1
Associations between EDDS and other domains of outcome assessed by questionnaire

	EDDS		χ^2	P			χ^2	P
	Interview-based				Questionnaire-based			
	Remission	Disordered			Remission	Disordered		
Axis I disorder	No. (%)	No. (%)			No. (%)	No. (%)		
Affective	2 (1.8)	11 (22.4)	19.41	.00001	2 (2.4)	11 (14.7)	8.09	.004
Anxiety	19 (17.0)	6 (12.0)	0.65	.42	12 (14.1)	13 (16.9)	0.24	.63
Substance use	1 (0.9)	8 (16.0)	15.04	.0001	0 (0)	9 (11.7)	10.52	.001
Impulse control	2 (1.8)	7 (14.0)	9.83	.002	1 (1.1)	8 (10.4)	6.54	.01
	Mean (S.D.)	Mean (S.D.)	t (df)	P	Mean (S.D.)	Mean (S.D.)	t (df)	P
HDRS	4.18 (5.13)	7.82 (7.12)	3.26 (73) ^a	.002	3.73 (4.38)	7.04 (7.08)	3.53 (124) ^a	.001
BSQ	73.9 (29.9)	115.5 (34.2)	7.82 (159)	.000	67.6 (25.5)	108.3 (35.5)	8.25 (135) ^a	.000
MPQ	17.05 (5.49)	13.66 (6.86)	3.27 (153)	.001	17.38 (5.36)	14.49 (6.57)	2.97 (139) ^a	.003
SAS-SR	1.76 (0.33)	2.04 (0.55)	3.30 (65) ^a	.002	1.68 (0.29)	2.03 (0.49)	5.46 (122)	.000
Work	1.49 (0.38)	1.84 (0.62)	3.64 (64) ^a	.001	1.47 (0.38)	1.74 (0.56)	3.54 (127) ^a	.001
Social/leisure	1.95 (0.52)	2.25 (0.66)	2.83 (75) ^a	.006	1.84 (0.43)	2.26 (0.64)	4.70 (130) ^a	.000
Extend family	1.59 (0.43)	1.80 (0.54)	2.61 (153) ^a	.01	1.51 (0.39)	1.82 (0.51)	4.25 (131) ^a	.000
Family unit	1.86 (0.65)	2.06 (0.91)	1.29 (58) ^a	.20	1.73 (0.59)	2.15 (0.83)	3.45 (117) ^a	.001
Parental	1.72 (0.51)	1.88 (0.81)	0.76 (21) ^a	.46	1.64 (0.51)	1.91 (0.64)	2.17 (79)	.03
Marital	2.00 (0.60)	2.26 (0.86)	1.82 (61) ^a	.07	1.93 (0.60)	2.23 (0.77)	2.56 (126) ^a	.01

Because not all participants were married, had children or resided with a family unit, $N < 162$ for these analyses.

^a Variance differed significantly between groups and separate variance was used to calculate t -statistic resulting in decreased degrees of freedom.

77 women (47.5%) met study criteria for any eating disorder. Based on interview assessment, 18 women (11.1%) met DSM-IV criteria for bulimia nervosa at follow-up and 50 women (30.9%) met study criteria for any eating disorder. Diagnostic agreement for presence vs. absence of bulimia nervosa was $\kappa=.49$ between questionnaire and interview assessments and $\kappa=.64$ for eating disorders. Diagnostic agreement for bulimia nervosa was higher for questionnaires and telephone interviews ($\kappa=.67$) than for questionnaires and face-to-face interviews ($\kappa=.35$). Similarly, diagnostic agreement for eating disorders was higher between questionnaires and telephone interviews ($\kappa=.71$) compared to agreement between questionnaires and face-to-face interviews ($\kappa=.58$).

Associations between eating disorder outcome and other domains of outcome

Table 1 presents associations between EDDS assessed by questionnaire or interview and concurrent measures of Axis I disorders, depression, body dissatisfaction, impulsivity and social adjustment. For both questionnaire and interview assessments of EDDS, significant associations were found with the presence of mood, substance use and impulse control disorders and levels of depression, body dissatisfaction, impulsivity and social adjustment. Similarly, EDDS was not associated with presence of anxiety disorders for either interview or questionnaire assessment of eating disorders. Notably, EDDS was significantly associated with self-reported social adjustment in the domains of marital, parental and family functioning when eating disorders were assessed by questionnaire but not when eating disorders were assessed by interview. When both variables in concurrent associations were assessed by the same method (e.g., both eating disorders and social adjustment were assessed by questionnaire), test statistics and *P*-values suggested somewhat stronger associations than when variables were assessed by different methods (e.g., eating disorders were assessed by interview and social adjustment was assessed by questionnaire).

Discussion

Similar to previous studies [14], questionnaires, when compared with interviews, produced a higher rate of eating disorders in a given sample. Reasons for increased false positives on questionnaire assessments could have included failure to ask required diagnostic criteria, misunderstood questions and differences in lay vs. clinical definitions of symptoms. However, none of these factors would explain the greater diagnostic agreement we found between questionnaires and telephone interviews than between questionnaires and face-to-face interviews. Using cut-offs of $\kappa \geq .60$ as indicating fair or adequate agreement and $\kappa \geq .70$ as indicating good agreement [26], agreement between ques-

tionnaires and telephone interviews could be characterized as “fair to good,” and agreement between questionnaires and face-to-face interviews could be characterized as “poor to fair.” One factor that might explain this difference is an increased willingness to disclose potentially shameful behaviors when not directly faced with the person asking about these behaviors. Thus, despite the limitations associated with self-report surveys, questionnaire assessments may not be inferior to interview assessments. Findings support the possibility that increased rates of eating pathology on questionnaire assessments may be due, in part, to increased candor when participants feel more anonymous.

Associations with concurrent measures of Axis I disorders, depression, body dissatisfaction, impulsivity and social adjustment, supported the validity of eating disorders diagnosed by either questionnaires or interviews. The differences in association strengths observed when the same method was used to measure different constructs vs. when different methods were used to measure different constructs has been observed before [27] and does not necessarily reflect the superiority of one method of assessment over the other.

Findings from the present study challenge the position that structured clinical interviews represent the “gold standard” in the assessment of eating disorders [14]. Although researchers may feel greater confidence in diagnoses based on direct interactions with research participants, validity of these diagnoses may not differ significantly from those generated by questionnaires. Noted drawbacks of questionnaire assessments remain valid; however, interview assessments are not free from methodological limitations. These include inter-rater variability, misunderstood responses, and decreased perceived anonymity. Further, some of these drawbacks are not encountered with questionnaires. Thus, rather than concluding that studies that employ questionnaires to assess eating pathology are inferior to those that employ interviews [14], researchers may more accurately acknowledge the potential limitations inherent to each assessment method as well as limitations associated with the specific instruments utilized.

Limitations of the present study should be acknowledged. First, questionnaires and interviews were not completed at the same time. Although they were completed within a month of one another, lowered agreement between these methods could be due to differing time periods. Second, it is possible that telephone and face-to-face interviews showed differential agreement with questionnaires for reasons other than perceived anonymity. Such reasons could include differences between how interviewers conducted face-to-face vs. telephone interviews and differences between participants who selected face-to-face vs. telephone interviews. Both reasons seem unlikely because the same well-trained interviewers who demonstrated high inter-rater reliability conducted both types of interviews and there was no significant association between eating disorder outcome and interview type. A third limitation was that diagnostic agreement was assessed between questionnaires and only one type of semi-

structured clinical interview. Greater differences may have emerged if other interviews and questionnaires were compared. Finally, all women had confirmed diagnoses of bulimia nervosa at baseline for which they sought treatment. As such, these women would be better informed of what was meant by “binge eating” than women drawn from a community sample, thus reducing the likelihood of misinterpreted questions. This could limit generalizability of study findings to follow-up studies of individuals with a history of confirmed eating disorder diagnoses. However, recent questionnaires that include more precise questions for these difficult-to-assess symptoms have demonstrated success in community-based samples [28].

In addition to limitations, the present study also had several strengths. First, the study included comparisons of questionnaire and interview data from a large sample, and, within this large sample, approximately half had completed their interviews over the telephone. This allowed comparison of agreement between questionnaires and interviews with higher perceived anonymity (telephone interviews) vs. lower-perceived anonymity (face-to-face interviews). Second, the κ statistic was used to evaluate diagnostic agreement. Because this method corrects for chance agreement, it gives a more accurate assessment of true agreement. Third, both questionnaire and interview instruments demonstrated high levels of reliability in this sample. This is important because a measure cannot demonstrate good agreement with another measure if it cannot demonstrate good reliability with itself.

In conclusion, questionnaire assessments may not be inferior to interview assessments; they may reveal different aspects of disordered eating. Findings support the possibility that increased rates of eating pathology on questionnaire assessments compared to interview assessments may be due, in part, to increased candor when participants feel more anonymous. This may represent an advantage to using questionnaires to assess eating pathology. In addition, questionnaires provide an efficient means for collecting data within studies employing very large samples. Future work should continue efforts to develop convenient, reliable and valid assessments of eating pathology by determining whether these findings are replicated in a community-based sample.

References

- [1] American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4th ed. Washington (DC): American Psychiatric Association, 2000 (Text Revision).
- [2] Heatherton TF, Nichols P, Mahamedi F, Keel P. Body weight, dieting, and eating disorder symptoms among college students, 1982 to 1992. *Am J Psychiatry* 1995;152:1623–9.
- [3] Leon GR, Fulkerson JA, Perry CL, Early-Zald MB. Prospective analysis of personality and behavioral vulnerabilities and gender influences in the later development of disordered eating. *J Abnorm Psychology* 1995;104:140–9.
- [4] Lock J, Reisel B, Steiner H. Associated health risks of adolescents with disordered eating: how different are they from their peers? Results from a high school survey. *Child Psychiatry Hum Dev* 2001; 31:249–65.
- [5] Neumark-Sztainer D, Hannan PJ. Weight-related behaviors among adolescent girls and boys: results from a national survey. *Arch Pediatr Adolesc Med* 2000;154:569–77.
- [6] Pemberton AR, Vernon SW, Lee ES. Prevalence and correlates of bulimia nervosa and bulimic behaviors in a racially diverse sample of undergraduate students in two universities in southeast Texas. *Am J Epidemiol* 1996;144:450–5.
- [7] Westenhoefer J. Prevalence of eating disorders and weight control practices in Germany in 1990 and 1997. *Int J Eat Disord* 2001;29: 477–81.
- [8] Hay P. The epidemiology of eating disorder behaviors: an Australian community-based survey. *Int J Eat Disord* 1998;23:371–82.
- [9] Drenowski A, Yee DK, Krahn DD. Bulimia in college women: incidence and recovery rates. *Am J Psychiatry* 1988;145:753–5.
- [10] French S, Peterson CB, Story M, Anderson N, Mussell MP, Mitchell JE. Agreement between survey and interview measures of weight control practices in adolescents. *Int J Eat Disord* 1998;23:45–56.
- [11] Halmi K, Falk JR, Schwartz E. Binge-eating and vomiting: a survey of a college population. *Psychol Med* 1981;11:697–706.
- [12] Leon GR, Fulkerson JA, Perry CL, Keel PK, Klump K. Three to four year prospective evaluation of personality and behavioral risk factors for later disordered eating in adolescent girls and boys. *J Youth Adolesc* 1999;28:181–96.
- [13] Pyle R, Halvorson PA, Neuman PA, Mitchell JE. The increasing prevalence of bulimia in freshman college students. *Int J Eat Disord* 1986; 5:631–47.
- [14] Fairburn C, Beglin SJ. Studies of the epidemiology of bulimia nervosa. *Am J Psychiatry* 1990;147:401–8.
- [15] Beglin S, Fairburn CG. What is meant by the term “binge”? *Am J Psychiatry* 1992;149:123–4.
- [16] Mitchell JE, Pyle RL, Hatsukami D, Goff G, Glotter D, Harper J. A 2–5 year follow-up of patients treated for bulimia. *Int J Eat Disord* 1988;8:157–65.
- [17] Mitchell JE, Pyle RL, Eckert ED, Hatsukami D, Pomeroy C, Zimmerman R. A comparison study of antidepressants and structured intensive group psychotherapy in the treatment of bulimia nervosa. *Arch Gen Psychiatry* 1990;47:149–57.
- [18] Keel PK, Mitchell JE, Miller KB, Davis TL, Crow SJ. Long-term outcome of bulimia nervosa. *Arch Gen Psychiatry* 1999;56:63–9.
- [19] First MB, Spitzer RL, Gibbon M, Williams JBW. Structured clinical interview for DSM-IV Axis I disorders — patient edition. New York (NY): New York State Psychiatric Institute, Biometrics Research Department, 1995.
- [20] Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry* 1960;23:56–62.
- [21] Mitchell JE, Hatsukami D, Eckert E, Pyle R. Eating disorders questionnaire. *Psychopharmacol Bull* 1985;21:1025–43.
- [22] Cooper P, Taylor MJ, Cooper Z, Fairburn CG. The development and validation of the body shape questionnaire. *Int J Eat Disord* 1987;6: 485–94.
- [23] Tellegen A. Brief manual for the Differential Personality Questionnaire. Minneapolis: University of Minnesota, 1982.
- [24] Weissman M, Bothwell S. Assessment of social adjustment by patient self-report. *Arch Gen Psychiatry* 1976;33:1111–5.
- [25] Keel PK, Mitchell JE, Miller KB, Davis TL, Crow SJ. Predictive validity of bulimia nervosa as a diagnostic category. *Am J Psychiatry* 2000;157:136–8.
- [26] Spitzer RL, Forman JBW, Nee J. DSM-III field trials: I. Initial interrater diagnostic reliability. *Am J Psychiatry* 1979;136:815–7.
- [27] Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychol Bull* 1959;56:81–105.
- [28] Stice E, Telch CF, Rizvi SL. Development and validation of the eating disorders diagnostic scale: a brief self-report measure of anorexia, bulimia, and binge-eating disorder. *Psychol Assess* 2000;12:123–31.