Psychological Methods
2003, Vol. 8, No. 1, 3–15

1082-989X/03/$12.00 DOI: 10.1037/1082-989X.8.1.3

Pseudotaxonicity in MAMBAC and MAXCOV Analyses of Rating-Scale Data: Turning Continua Into Classes by Manipulating Observer's Expectations

Theodore P. Beauchaine
University of Washington

Everett Waters
State University of New York at Stony Brook

Taxometric procedures such as mean above minus below a cut and maximum covariance can determine whether a trait is distributed as a discrete latent class. These methods have been used to infer taxonic structure in several personality and psychopathology constructs, often from analyses of rating scale data. This is problematic given (a) well established biases in ratings, (b) the human tendency to think categorically, and (c) implicit typological models of personality and psychopathology among expert raters. Using an experimental method in which the cognitive sets of raters were manipulated as dimensional versus categorical, it is demonstrated that pseudotaxonicity can be created readily with rating scale measures. This suggests that researchers avoid an exclusive reliance on rating scales when conducting taxometrics investigations.

Whether or not specific personality and psychopathology traits are distributed discretely or continuously is a question of central interest to many behavioral scientists. Are there really two types of people in the world, those with Type A and Type B behavior patterns (e.g., Strube, 1989)? Does psychopathy reflect behaviors falling at one extreme of an antisocial continuum, or does it mark a qualitative difference in the structure of personality (e.g., Harris, Rice, & Quinsey, 1994)? Answers to these and similar questions have both philosophic and pragmatic implications (e.g., Gangestad & Snyder, 1985; Meehl 1995a, 2001; Waldman & Lilienfeld, 2001). Indeed, given evidence that a personality characteristic or psychiatric disorder is distributed discretely, we are in a position to establish nonarbitrary cutoffs that truly "carve nature at its joints" (Gangestad & Snyder, 1985). Moreover, solid evidence of discreteness provides strong support for the validity of a behavioral syndrome or diagnostic entity (Beauchaine & Beauchaine, 2002).

Efforts to distinguish types from continua are not new. Perhaps the simplest approach to this problem is to plot the univariate distributions of traits or symptoms and search for evidence of bimodality (see Kendell, 1989). This approach is not particularly powerful, however, as two distinct but overlapping distributions often appear to be unimodal when mixed, even at effect sizes as large as 2.0 (see Grayson, 1987; Murphy, 1964). Thus, more sophisticated approaches to differentiating types from continua have emerged. These efforts began with the development of cluster analysis, which was initially described as a method for identifying biological taxa. At present, there are literally hundreds of clustering algorithms (see Blashfield & Aldenderfer, 1988), most of which were created with the objective of identifying groups of observations (e.g., species, people) based on two or more manifest characteristics. Other approaches to this problem include latent class analysis, and recently described methods for identifying groups based on latent growth trajectories (e.g., Nagin, 1999). Although these approaches may be useful in describing cross-sectional and longitudinal patterns of functioning among subgroups of individuals within a sample, none can establish that the identified latent classes are truly discrete (see Beauchaine, in press;

Theodore P. Beauchaine, Department of Psychology, University of Washington; Everett Waters, Department of Psychology, State University of New York at Stony Brook.

We thank Robert J. Beauchaine III for his helpful contribution to this work. We also thank Chris Fraley and Niels Waller for their constructive suggestions.

Correspondence concerning this article should be addressed to Theodore P. Beauchaine, Department of Psychology, University of Washington, Box 351525, Seattle, Washington 98195-1525. E-mail: tbeaucha@u.washington.edu

Haertel, 1990; Klein & Riso, 1993; Masters, 1985; Nagin, 1999; Nagin & Tremblay, 2001). Thus, each may yield results that are consistent with discrete latent structure, but such findings cannot falsify taxonic conjectures because similar results are often observed in the presence of continuous latent distributions.[1] In other words, false positives are common.

In an effort to address this problem, Meehl and colleagues invented a set of algorithms referred to as *Coherent Cut Kinetics* (CCKs; Grove & Meehl, 1993; Meehl, 1973, 1995a; Meehl & Yonce, 1994, 1996; Waller & Meehl, 1998). These techniques were designed to identify taxon groups when they exist and to reject the taxon hypothesis when they do not. The best known CCK algorithms are mean above minus below a cut (MAMBAC) and maximum covariance (MAXCOV). MAMBAC and MAXCOV are typically used in a two-stage fashion toward testing taxonic conjectures (see Meehl, 1995a, 2001). Because MAMBAC is less intensive computationally, it is often used first to screen candidate indicators of taxon group membership. If MAMBAC results are consistent with a taxonic hypothesis, selected indicators are then subjected to the MAXCOV procedure.[2] As with MAMBAC, MAXCOV yields results that are either consistent or inconsistent with a two-group latent structure hypothesis. Given the former, MAXCOV produces Bayesian estimates of each individual's probability of belonging to the taxon group (see Beauchaine & Beauchaine, 2002; Waller & Meehl, 1998). Converging evidence for discrete groups from both MAMBAC and MAXCOV analyses provides for increased confidence in the validity of taxonic results. Such corroborative analyses across procedures are often referred to as consistency tests and are considered essential in evaluating taxonic hypotheses (e.g., Meehl, 1995a).

Meehl and others have presented extensive Monte Carlo simulations assessing the performance of MAMBAC and MAXCOV under a variety of conditions (e.g., Beauchaine & Beauchaine, 2002; Cleland & Haslam, 1996; Meehl, 1995b; Meehl & Yonce, 1994, 1996). Results from these studies suggest that the algorithms are unlikely to produce false positives and therefore can falsify taxonic conjectures (see also Meehl, 1996). Moreover, they are robust in the face of considerable distributional overlap, significant nuisance (within-groups) covariances, and moderate distributional skew. They are also effective across a wide range of taxon base rates. These operating characteristics have led to an increasing use of MAMBAC and MAXCOV in recent years. MAXCOV, either alone or in conjunction with other CCK procedures, has been used to infer latent taxonic structure in antisocial behavior (Skilling, Quinsey, & Craig, 2001), dissociative experiences (Waller, Putnam, & Carlson, 1996; Waller & Ross, 1997), eating disorders (Williamson et al., 2002), endogenous depression (Haslam & Beck, 1994), hedonic capacity (Blanchard, Gangestad, Brown, & Horan, 2000), infant reactivity (Woodward, Lenzenweger, Kagan, Snidman, & Arcus, 2000), psychopathy (Harris et al., 1994), schizotypy (Erlenmeyer-Kimling, Golden, & Cornblatt; 1989; Golden & Meehl, 1979; Korfine & Lenzenweger, 1995; Lenzenweger, 1999; Lenzenweger & Korfine, 1992; Meyer & Keller, 2001; Tyrka et al., 1995), sexual orientation (Gangestad, Bailey, & Martin, 2000), and Type A behavior patterns (Strube, 1989).

Although it is important to test taxonic hypotheses among these constructs, it is of concern that some of the above referenced works were conducted using strictly rating-scale data, either by means of self-report or expert opinion. This practice is potentially problematic for at least two reasons. First, it is well known that rating scales can elicit systematic response tendencies including halo effects and positive or negative response biases (e.g., Cooper, 1981; Macmillan & Creelman, 1990; Rajendar, 1996; Saal, Downey, & Lahey, 1980). Either of these could conceivably produce spurious taxonic structure (i.e., pseudotaxonicity), a possibility that is of particular concern for ratings that require subjective inference on the part of

[1] Despite this limitation, cluster analysis, latent class analysis, and latent-growth trajectory approaches offer the advantage of allowing for models with $n > 2$ groups. In contrast, the taxometric methods described in sections to follow are limited to analyses involving a single taxon group and its complement class ($n = 2$). Thus, there are advantages and disadvantages to using either set of approaches.

[2] Because MAMBAC operates on variable pairs, there are $2\binom{k}{2}$ possible combinations of indicators available for analysis. In contrast, MAXCOV operates on variable triads, with $k[(k - 1)!/(k - 3)!2!]$ possible combinations. As the number of candidate indicators increases, MAXCOV combinations therefore accelerate much faster than MAMBAC combinations. Performing MAMBAC analyses first allows for screening of variables and elimination of those that do not mark the putative latent taxon. A more reasonable number of remaining variables can then be subjected to the MAXCOV algorithm.

judges. This is because as much as 50% of the variance in such measures may be attributable to rater bias (Hoyt & Kerns, 1999).

Second, a corpus of literature suggests that human beings are prone toward categorical thinking (see Malt, 1993; Rosch & Lloyd, 1978; Smith, 1995), which includes a natural tendency to classify people on the basis of preexisting beliefs and prior experiences (see Cantor & Genero, 1986; Cantor & Mischel, 1979; Flanagan & Blashfield, 2002; Semin & Rosch, 1981). Moreover, when making categorical decisions, raters become more confident with increasing practice (Dawes, Faust, & Meehl, 1989), and their beliefs become more divergent over time, as assessed by Likert-scale measures (Simon, Pham, Le, & Holyoak, 2001).

These findings suggest that a sole reliance on rating-scale data may be ill-advised when performing taxometrics research, particularly when there is reason to suspect that raters hold categorical beliefs about the latent structure of a construct. In attachment research, for example, infants are classified as secure or insecure, in part on the basis of expert-coded ratings of their behaviors during separation from and reunion with their mothers (see Ainsworth, Blehar, Waters, & Wall, 1978; Sroufe & Waters, 1977). Similarly, adult attachment classifications are often determined from coded ratings of Adult Attachment Interview transcripts (George, Kaplan, & Main, 1985; see also Crowell & Treboux, 1995). In both cases, the constructs of interest are derived in large part from scale scores. Because many researchers endorse categorical models of attachment classifications (see Fraley & Waller, 1998), subtle biases could impact their ratings of behaviors, producing latent distributions that appear to be discrete in taxometric analyses. Our own unpublished Strange Situation data have yielded inconsistent outcomes from one MAMBAC/MAXCOV trial to the next, some suggesting taxonicity. Moreover, we have obtained similar results using Adult Attachment Interview scores. Although these findings could be interpreted as consistent with a categorical model, it is also possible that shared expectations among raters are producing pseudotaxonic outcomes. The only published taxometric analysis of attachment status also yielded mixed results (Fraley & Spieker, in press), some of which were suggestive of taxonic structure despite the use of a conservative analytic strategy that may have reduced the likelihood of identifying latent taxa (Waters & Beauchaine, in press). Before findings such as these are interpreted, it is important to investigate the susceptibility of rating-scale data to artifactual taxonicity induced by the preexisting beliefs of raters.

Thus, the purpose of the present investigation was to determine whether manipulating the cognitive sets of raters can induce pseudotaxonicity when using MAMBAC and MAXCOV. Undergraduates were recruited and trained to rate graduate school admissions statements of applicants to a clinical psychology program, across several dimensions. Half of the raters were told that the statements reflected a full-range student ability, as assessed by later performance in graduate school. The other half were informed that the statements were selected from students who either struggled or excelled in graduate school. Ratings from both groups were subjected to MAMBAC and MAXCOV analyses.

Method

Participants

Thirty undergraduates at the University of Washington took part in the study and were assigned randomly to either an experimental group ($n = 15$) or a control group ($n = 15$). Because participation required a significant time investment on the part of raters (see below), each was paid $100 for completing the experiment.

Materials

Materials included 75 personal statements from prospective graduate students applying to the Clinical Psychology program at the State University of New York at Stony Brook during a recent year of admission. Statements were scanned into a word processing program, and all identifying information was extracted, including names of individuals, institutions, and geographical locales. Statements authored by applicants who spoke a first language other than English were omitted from the pool. All other statements were accepted serially from a total of roughly 200. Because the statements were not ordered systematically, this procedure is likely to have captured a representative sample. All statements were between one and two pages in length.

Procedure

Before beginning the experiment, approval to use human participants was obtained from the institutional review boards at both the State University of New York at Stony Brook and University of Washington. Each participant rated the 75 statements on six dimensions designed to assess either (a) the quality of the essay itself or (b) the likelihood of the applicant succeeding in graduate school and beyond (see Table 1). Ratings for each dimension were rendered on a 10-point scale with anchors ranging from *very poor* to *extremely good.*

Prior to rating the essays, all participants attended a half-hour meeting in which the study procedures were described. Two separate meetings were conducted, one with the control group and one with the experimental group. In both meetings, participants were told that personal statements are highly predictive of success in graduate school as assessed by academic records and that the objective of the study was to determine whether naive raters could rate statements as accurately as faculty members. Only one instruction differed across groups: Members of the control group were told that the essays reflected a full range of graduate school performance, whereas members of the experimental group were told that the statements were written by students who either struggled or excelled in graduate school. Both groups were encouraged explicitly to use the entire 10-point scale when making their ratings.

In addition, participants were told that (a) the statements were collected over several years, (b) all of the applicants had been admitted to graduate school, and (c) the performance outcomes of applicants in terms of graduate school success were known. To enhance participants' investment in the task, we told them that they would be paid $75 for rating the statements, with a $25 bonus for those whose ratings predicted graduate school success better than the average rater. At the end of the study, all participants were paid the full $100.

To complete the ratings, participants logged onto a password-protected Web site that presented them with each essay, one after the other. The order of essay presentation was randomized. Essays appeared in a large frame that occupied the right two thirds of the Web page, and ratings were entered in a frame that occupied the left third of the page. Included in the latter frame were brief descriptions of each rating category. To maximize attention to the task, we allowed participants to rate no more than five essays per session. They were also required to complete their ratings within 30 days of their first login.

Data Analyses

Descriptive statistics. Data analyses proceeded in two phases. First, descriptive statistics were calculated and group comparisons conducted for all rating dimensions. Although mean differences in ratings would not necessarily affect taxonic outcomes, between-groups comparisons were assessed for descriptive purposes using hierarchical linear modeling

Table 1
Rating Dimensions

Dimension	Operational definition
Believability	Does the author communicate in a straightforward manner or appear to be selling himself or herself too strongly? Is the author trying to impress readers with unnecessary technical terms and jargon?
Clarity/succinctness	Overall, is the essay well organized and easy to follow or overly verbose or awkward? Could the author communicate the same information in less space? Did the author write with purpose or meander through seemingly unrelated topics?
Persuasiveness	Overall, is the essay compelling? Does it convince you that the applicant is a reasonable candidate for graduate school? Are solid qualifications presented in a persuasive manner, or does the author appear to be either unsure of himself or herself or to be out of touch with the requirements for candidacy to graduate school?
Probability of success in graduate school	Do you believe the author will be a successful graduate student? Are the proper qualifications listed? Does the author come across as mature, capable, and able to handle criticism, or does he or she write poorly or seem overly confident or arrogant?
Probability of success as a scientist	Clinical psychologists must be capable research scientists. Do you believe the author will be a successful researcher? Are the proper qualifications listed? Does the author come across as rational and level-headed, or does he or she appear to be interested only in clinical work?
Probability of success as a clinican	Clinical psychologists must be capable practitioners. Do you believe the author will be a successful clinician? Are the proper qualifications listed? Does the author express a mature rationale for pursuing a clinical profession, or does he or she simply describe a vague desire to help others?

(HLM). This approach was necessary because ratings were nested within participants. For each of the six rating dimensions, two-level models were constructed (Bryk & Raudenbush, 1992). In these analyses, the 75 repeated observations for each rater were modeled at Level 1, with individual variation among raters modeled at Level 2. Group comparisons were conducted by including a dummy coded variable (experimental vs. control) as a Level 2 fixed effect. We were interested in addressing two questions with the HLM models. The first concerned potential group differences in mean ratings. To examine this, Level 1 intercepts were centered within individuals, thus providing an estimate of each participant's mean rating across the 75 statements. These estimates were then compared across groups in tests of Level 2 fixed effects. Second, we were interested in examining whether group differences in rating trends were observed across the 75 statements. This question was addressed by testing for group differences in slopes. All models were constructed using full maximum likelihood. The general model for each of the six ratings was as follows:

$$\begin{aligned} \text{Level 1: } Y &= \beta_{0j} + \beta_{ij}(\text{repeated observation}) + r_{ij}; \\ \text{Level 2: } \beta_{0j} &= \gamma_{00} + \gamma_{01}(\text{group}) + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}(\text{group}) + u_{1j}. \end{aligned}$$

Taxometric analyses. Because MAMBAC has been outlined in detail elsewhere (e.g., Meehl, 1995a; Meehl & Yonce, 1994; Waller et al., 1996), it is described only briefly here. For each MAMBAC run, one indicator is specified as the input variable, and the other is specified as the output variable. The input variable is sorted, which also sorts the output variable with more or less efficiency depending on the validity of the indicators as markers of the measured trait. A sliding cut is then moved across the range of the input variable, and the mean of the output variable is calculated, both above and below the cut. The mean above the cut is subtracted from the mean below the cut, and obtained values are plotted across the entire range of the input variable. If discrete taxon groups exist, a marked peak is observed in this plot. In contrast, when a trait is distributed continuously, a U-shaped function is observed (see Figure 1). In the taxonic case, the location of the peak of the MAMBAC function corresponds to the input variable score that best differentiates between the groups. When the taxon base rate is .50, the peak appears near the median of the input variable. As the base rate becomes progressively lower, the peak shifts toward the right.

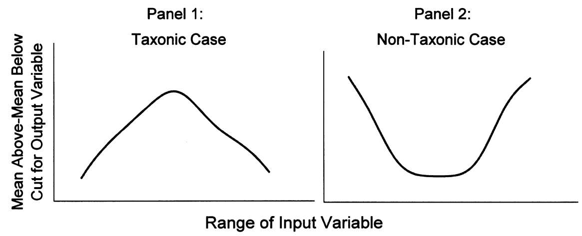

Figure 1. Smoothed mean above minus below a cut plots from Monte Carlo simulated data. The left plot (Panel 1) was derived from two univariate normal distributions of $n = 500$ and equal variance that were mixed. The effect size separating the distribution means was 1.5. The right plot (Panel 2) was derived from a single univariate normal distribution of $n = 1{,}000$.

MAMBAC analyses were performed separately for the experimental and control groups using a Statistica (1998) Basic program written by T. Beauchaine.[3] Within each group, there were 1,125 observations for each of the six rating dimensions (15 raters × 75 ratings). All pairwise sets of these rating dimensions were subjected to MAMBAC analyses.

Next, MAXCOV analyses were performed on each triadic combination of variables. Because MAXCOV has also been described in detail elsewhere (e.g., Beauchaine & Beauchaine, 2002; Meehl, 1995a; Meehl & Yonce, 1996), it is outlined only briefly here. When MAXCOV analyses are performed, one variable (designated the cut variable) is first sorted. As with MAMBAC, this also sorts the other two variables with more or less efficiency depending on the validity of each as markers of the measured trait. The cut variable is then divided into several intervals based on standard deviation (*SD*) units (e.g., one third or one fourth). Next, the covariance of the two remaining indicators is calculated within each interval, and the resulting function is plotted. Similar to MAMBAC, taxonic data yield a marked peak in the MAXCOV plot. The location of this peak occurs within the hitmax interval, or the value of the cut variable that best differentiates the taxon and nontaxon groups. In contrast, nontaxonic data produce relatively flat MAXCOV plots.

MAXCOV yields point estimates of the taxon base rate, the taxon and nontaxon group means on the sort-

[3] This program is available for download at http://tbeauchaine.psych.washington.edu/tb/cbs/taxometrics.htm.

ing variable, the false-positive and false-negative rates of group assignment, and Bayesian-derived probability estimates of taxon group membership for each individual (see Beauchaine & Beauchaine, 2002; Meehl & Yonce, 1996; Waller & Meehl, 1998). All MAXCOV analyses were conducted using a C++ program that we have described elsewhere (Beauchaine & Beauchaine, 2002). A one-third standard deviation interval size was used across a range of −3.0–3.0 *SD* on the sorting variable.[4] When too few data points ($n < 10$) were available within an interval to calculate a reliable covariance term, an interpolated value from adjacent one-third *SD* windows was substituted (see Meehl & Yonce, 1996). Estimates of the taxon base rate, the hitmax value, the true-positive rate, and the false-positive rate were entered into Bayes's theorem, from which a probability of taxon group membership was derived. In the simplest three-variable case, these probabilities are given by the formula,

$$\Pr(t|x^+y^-z^+) = \frac{Pp_{tx}q_{ty}p_{tz}}{Pp_{tx}q_{ty}p_{tz} + Qp_{cx}q_{cy}p_{cz}}, \qquad (1)$$

where $\Pr(t|x^+y^-z^+)$ represents the probability of taxon group membership given scores above the hitmax value on x and z and below the hitmax value on y, P represents the taxon base rate, Q represents the complement-class base rate, p_{tx} represents the true positive rate for variable x as derived from the MAXCOV analysis (see below), p_{cx} represents the false-positive rate for variable x as derived from the MAXCOV analysis, $q_{tx} = 1 - p_{tx}$, and $q_{cx} = 1 - p_{cx}$.

Originally, we planned to subject only those variables that appeared taxonic in the MAMBAC runs to subsequent MAXCOV analyses (see Meehl, 1995a, 2001). However, nearly all MAMBAC plots from the experimental group suggested latent taxonic structure, compared with few if any from the control group (see below). Thus, all variables were included in the MAXCOV analyses of experimental group data. For comparison purposes, all variables were also included in the control group MAXCOV analyses.

Results

Descriptive Statistics

Descriptive statistics and results of the HLM analyses are reported for each rating dimension in Table 2. Skewness and kurtosis values were acceptable for all variables. As indicated by the tests of Level 2 variance components, significant variability in mean ratings was observed among participants for all rating dimensions. However, tests of Level 2 fixed effects indicated no significant differences in mean ratings across groups. Although not reported in Table 2, inspection of the slope parameters also suggested significant variability among participants, all χ^2s(28) > 43.7, all *p*s < .03. Once again, however, significant group differences as assessed from tests of Level 2 fixed effects were not found, all *t*s(28) < 0.29, all *p*s > .77. Moreover, the Level 1 slope parameters were nonsignificant for all rating dimensions, suggesting no linear trends across rating trials for the sample as a whole, all *t*s(2246) < 1.42, all *p*s > .15.

Interrater reliability was assessed by computing intraclass correlation coefficients (ρ_I) for each rating dimension using Eisenhart's Type II analysis of variance model (see Snijders & Bosker, 1999). Separate intraclass correlation coefficients were calculated for the experimental and control groups. Results indicated substantial agreement among raters, with ρ_Is ranging from .68 to .80. No significant group differences were observed, all *F*s(1148) < 1.61, all *p*s > .20. None of these findings suggest an enhanced or attenuated likelihood of producing taxonic outcomes.

Taxometric Analyses

MAMBAC plots are presented in Figure 2. For the experimental group (top panel), nearly all of the 30 plots exhibited a marked peak, suggesting latent taxonic structure. In addition, the location of the peaks indicated a taxon base rate somewhat below .50. In contrast, few if any of the MAMBAC functions from the control group (bottom panel) suggested taxonic structure, as all plots were U-shaped.

MAXCOV analyses of the experimental group data also suggested latent taxonic structure. Because there were 120 trivariate combinations of indicators (60 for each group), separate MAXCOV plots are not presented. However, representative plots for both the experimental and control groups are depicted in Figure 3. Taxonicity was assessed by examining the Jöreskog and Sörbom (2001) goodness-of-fit (GFI) index comparing the observed and predicted covariance matrices

---

[4] The range of −3–3 was chosen after examining *z*-scores for all variables. Of the 2,250 data points, only 13 (0.5%) fell outside of the ±3 interval. Thus, extending the range beyond this would have created MAXCOV windows with virtually no data points.

Table 2
Descriptive Statistics and Group Comparisons for Each Rating Dimension

Dimension	Experimental group (n = 15)				Control group (n = 15)				HLM Level 2 effects			
									Variance component[a]		Fixed effect[b]	
	M	SD	Skew	Kurt	M	SD	Skew	Kurt	$\chi^2_{(28)}$	p	$t_{(28)}$	p
Believability	7.07	1.86	−.01	−.27	7.13	1.93	−.81	.41	974.62	< .001	0.73	.47
Clarity/succinctness	6.39	2.10	−.02	−.49	6.75	2.01	−.48	−.14	1,081.87	< .001	1.32	.20
Persuasiveness	6.17	2.28	.07	−.80	6.42	2.23	−.63	−.27	568.61	< .001	0.13	.90
Probability of grad school success	6.51	2.34	−.25	−.61	6.88	2.02	−.36	.07	776.50	< .001	0.71	.49
Probability of success as scientist	6.34	2.49	−.24	−.59	6.54	2.25	−.68	.06	753.95	< .001	0.17	.87
Probability of success as clinician	6.37	2.37	−.27	−.46	6.42	2.17	−.66	.08	727.63	< .001	0.31	.76

Note. HLM = hierarchical linear modeling; Skew = skewness; Kurt = Kurtosis.
[a] Test of Level 2 variability among individuals in mean ratings. [b] Test of Level 2 variability between groups in mean ratings.

(Waller & Meehl, 1998). GFI values above .90 are typically not obtained from dimensional data. For the experimental group ratings, the GFI was .92.[5] Moreover, as illustrated in the left panel of Figure 4, 1,074 of the 1,125 ratings (95.5%) were assigned a probability of taxon group membership in the upper or lower deciles of the (0,1) probability interval by Bayes's theorem. This pattern is also suggestive of latent taxonic structure (Waller & Meehl, 1998). The base rate of the taxon was .34.[6]

MAXCOV analyses of the control group data yielded a GFI of .80 and were therefore not consistent with a taxonic model. As depicted in the right panel of Figure 3, Bayesian probability estimates of taxon group membership placed only 458 of 1,125 ratings (40.7%) in the upper or lower deciles of the (0,1) interval.

Discussion

Our objective in conducting this experiment was to determine whether manipulating the cognitive sets of raters could produce taxonic structure in rating-scale data. By telling some raters that essays were written by graduate students who either excelled or struggled, we were able to produce categorically distributed ratings of both essay quality and the likelihood of authors' future success. In contrast, participants who were not instructed in this manner did not produce categorically distributed ratings.

---

[5] Several articles in the structural equation modeling literature have noted limitations with the GFI, including sensitivity to sample size and the use of a nonrealistic baseline with no variance and no covariance (Hu & Bentler, 1998; Marsh, Balla, & McDonald, 1988; Tanaka, 1993). Because a large sample was used in the present study, the former concern may be less problematic. However, the nonrealistic baseline issue remains. Despite this caveat, the GFI has discriminated between taxonic and nontaxonic data in previous research, and it is the only fit index that has been applied to the MAXCOV procedure. We were therefore reluctant to use an alternative fit index without first exploring its operating characteristics in a simulation study.

[6] As pointed out by an anonymous reviewer, the question of whether participants' ratings drifted toward categories over time is also of interest. Unfortunately, ratings were not time stamped, so this question could not be addressed with precision. However, we did split the experimental group ratings in half on the basis of the order in which they were posted (first half vs. last half). Both halves yielded taxonic outcomes based on MAMBAC analyses.

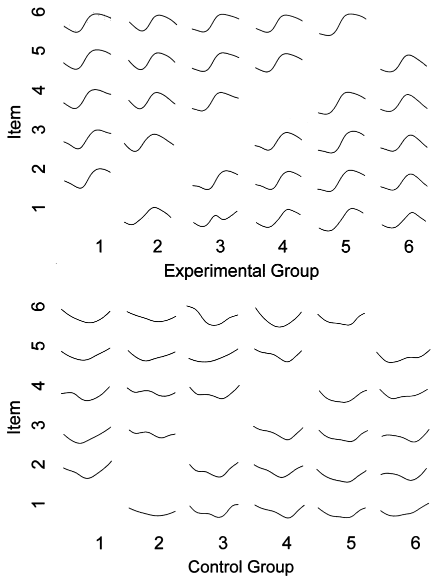

Figure 2. Smoothed mean above minus below a cut plots of all pairwise combinations of indicators. Top: Results for the experimental group. Bottom: Results for the control group. All plots were generated with y-axis scaling between −1.8 and 2.8.

It is worth noting that a brief and straightforward manipulation affected ratings conducted over the next several weeks. These results are especially persuasive regarding the potential impact of cognitive biases on taxometric analyses because participants' expectations were manipulated implicitly. Thus, the instructions given to participants induced an inference about taxonicity, and no ongoing manipulation was required to maintain the resulting cognitive bias. Moreover, this bias was almost certainly much weaker than the consolidated, theoretically based expectations that researchers and expert observers are likely to bring to a rating task.

These results have potential implications for future taxometrics research and suggest that a sole reliance on rating-scale data may be ill-advised when testing taxonic hypotheses, particularly if there is reason to believe that raters endorse a categorical model of the construct being assessed. To return to the example of attachment, there is no way to determine whether findings that are consistent with discrete attachment classifications are the result of raters' beliefs, which are likely to be categorical, if the indicators used for assessment are all rating-scale scores. As noted above, many experts in the field endorse categorical models of both infant and adult attachment, and many also believe that valid inferences cannot be offered about an individual's attachment pattern without extensive training from specific people in the field. The present study suggests that this practice carries the real danger of producing false latent taxonic structure in analyses of expert ratings of attachment behaviors.

Although there may be contexts in which expert ratings of attachment status are of value, researchers who wish to test taxonic hypotheses might be advised to use more objective sources of data. For example, raters who are naive to attachment theory could code objective frequencies of attachment-related behaviors, which could then be aggregated to produce attachment scores. Objectively coded attributes, such as behavioral frequency counts, carry far less bias than ratings requiring significant judgment and inference on the part of raters (Hoyt & Kerns, 1999). Moreover, it is commonplace for raters to be kept blind to experimental hypotheses and the group status of partici-

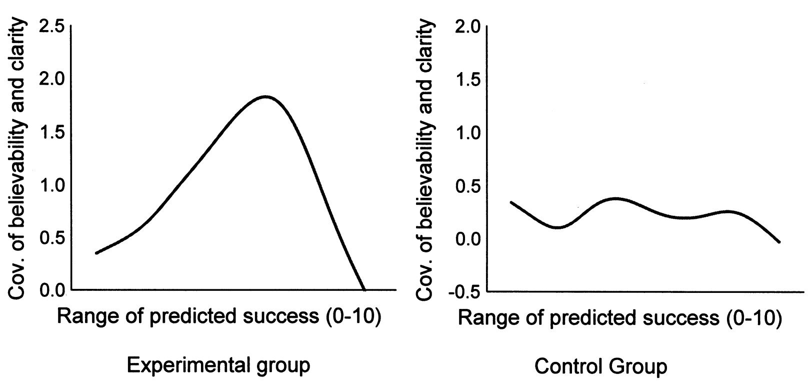

Figure 3. Smoothed maximum covariance plots of the covariance of essay believability and clarity across the entire range of predicted success in graduate school for the experimental group (left panel) and for the control group (right panel). Cov. = covariance.

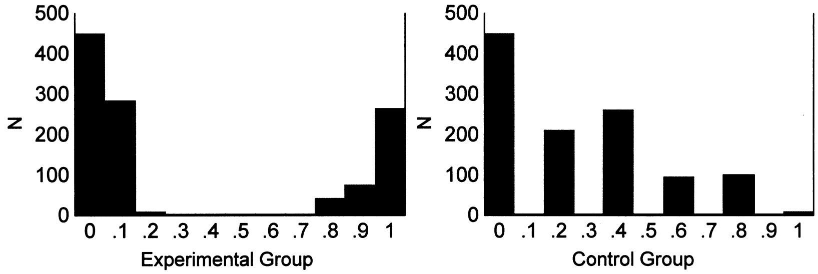

Figure 4. Bayesian-estimated probabilities of taxon membership for ratings generated by experimental group participants (left panel) and by control group participants (right panel).

pants in other observational coding contexts (e.g., Beauchaine, Strassberg, Kees, & Drabick, 2001; Snyder, Edwards, McGraw, Kilgore, & Holton, 1994).

Similar to the example of attachment, evidence for latent typologies of eating disorders has also been derived from taxometric analyses of expert ratings of symptom severity (Williamson et al., 2002). Consequently, these results may be subject to the influence of cognitive biases as well and should therefore be replicated using more objective measures before conclusions are reached about the ontological status of eating disorders as dimensional versus discrete.

Although some might argue that expert observers would be more resistant to rating biases than naive raters, empirical evidence suggests otherwise. As mentioned previously, human beings think categorically by nature (see Malt, 1993; Rosch & Lloyd, 1978; Smith, 1995) and become more confident in making categorical decisions with increasing experience (Dawes et al., 1989; Simon et al., 2001). Taken together, these findings suggest that taxometric analyses based on scale scores rendered by experts may be ripe for detecting spurious taxa. Although the implications for self- and informant-report measures are less clear, such data might also be subject to false-positive findings in taxometric analyses if respondents adhere to implicit categorical models of the traits they are rating. Given this possibility, the conclusion that certain psychological constructs are indeed taxonic may need to be reexamined. Examples include pathological dissociation (Waller et al., 1996; Waller & Ross, 1997) and psychopathy (Harris et al., 1994), each of which has been analyzed using exclusively self-report measures. In contrast, the taxometric status of schizotypy has been confirmed using not only self-report data (e.g., Korfine & Lenzenweger, 1995; Lenzenweger, 1999) but also neuromotor performance and objective behavioral markers (Erlenmeyer-Kimling et al., 1989; Tyrka et al., 1995). In addition, Woodward et al. (2000) used objective behavioral frequencies in their taxometric analysis of infant reactivity. Thus, stronger cases can probably be advanced for schizotypy and infant reactivity as discrete behavioral syndromes. Because the effects of potential cognitive biases in self- and informant-report indices can be significant, obtaining at least some objective indicators might be preferred in all taxometrics research (see also Beauchaine, in press).

Our findings may also have implications for diagnostic assessment. Clinicians, for example, are often faced with categorical decisions about the diagnostic status of clients or patients. Although there are many reasons to believe that most personality characteristics and psychiatric disorders are distributed continuously (e.g., Beauchaine, in press; Klein & Riso, 1993), prevailing models of psychopathology are explicitly categorical (*Diagnostic and Statistical Manual of Mental Disorders,* 4th ed., text rev.; American Psychiatric Association, 2000). Thus, biases in the thinking of diagnosticians about psychiatric conditions may be prevalent, resulting in misclassification of individuals who fall in the midrange along continua of symptomatology. It may therefore be useful to educate clinicians regarding potential biases in their thinking that are likely to be induced by the prevailing diagnostic system. However, we should not assume that educative efforts will eliminate rater biases. Rather, current evidence suggests rater training can reduce but not abolish such effects (Hoyt & Kerns, 1999).

One finding that was not anticipated was the base rate of the latent taxon (.34). Our instructions advised participants in the experimental group that "about half of the essays were written by quite successful graduate students, while the other half were written by students who struggled." Thus, we expected that if a latent taxon was identified, the base rate would be around .50. Although speculative, this discrepancy may represent a general bias among participants against rendering too many higher-than-average ratings. We await the results of future studies to determine if this shift in base rates replicates. If it does, taxometric analyses could provide a means of examining the bias further, as it may be of independent interest.

Alternatively, it is possible that the base rate shift reflected a somewhat valid inference among raters that only a minority of the essays were written by future graduate students. In fact, less than 10% of the statements were authored by those who were eventually admitted to the Clinical Psychology program at Stony Brook. Given that most students who attend competitive graduate programs apply to several universities, it is possible that the observed 34% base rate approximated the true distribution of eventual graduate students. This would suggest that the experimental instructions induced a dichotomous rating set of greater predictive validity than would have been observed had raters adhered to a base rate of 50%. Because we do not have information linking the essays to graduate school attendance, this possibility cannot be ruled out. However, it could also be explored in future research.

A question of potential interest that was not addressed in this study is to what extent individual raters were influenced by the manipulation. In other words, did most participants who received categorical instructions rate categorically or did some rate continuously? Although we suspect that most necessarily rated categorically to produce the observed pattern of results, this could not be verified with only 75 ratings per person. Meehl (1995a) has suggested that at least 300 data points be used when conducting taxometric investigations, and we have found 200 to be an absolute minimum (Beauchaine & Beauchaine, 2002). Although we considered this issue when planning the experiment, we concluded on the basis of established literature (e.g., Dawes et al., 1989; Simon et al., 2001) that the task needed to be somewhat involving to maximize the likelihood of inducing categorical beliefs. Given this, we could not reasonably expect participants to commit to rating 200–300 essays.

We also did not address the question of whether inducing a continuous cognitive set among raters can obscure latent taxonic structure in a construct that is known to be typological. In our view, this is less likely given the literature outlined above, suggesting that humans think categorically by nature and drift toward categorical ratings on continuous dimensions over time. Thus, it may be more difficult to induce continuous biases than typological biases, yet this remains an empirical question that can be addressed in future research.

With the development of CCK procedures such as MAMBAC and MAXCOV, taxometric methods are increasingly recognized as important tools in construct validation research. In performing our own taxometric investigations, we became concerned about analyses that were based solely on rating-scale measures given the tendency of humans to categorize information. It appears that this concern was well founded. However, the present analyses do not undermine the logic or value of taxometric procedures. To the contrary, they continue Meehl's 30-year commitment to consistency tests and to understanding the operating characteristics of these methods through empirical analysis. No statistical procedure is self-interpreting (Meehl, 1995a) or entirely self-correcting. Recognizing that observer expectations can produce pseudotaxonicity in research based exclusively on rating-scale data is an important step toward making the best use of CCK and other taxonomic search methods.

References

Ainsworth, M. D. S., Blehar, M. C., Waters, E., & Wall, S. (1978). *Patterns of attachment: A psychological study of the Strange Situation.* Hillsdale, NJ: Erlbaum.

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC: Author.

Beauchaine, T. P. (in press). Taxometrics and developmental psychopathology. *Development and Psychopathology.*

Beauchaine, T. P., & Beauchaine, R. J., III (2002). A comparison of maximum covariance and *k*-means cluster analysis in classifying cases into known taxon groups. *Psychological Methods, 7,* 245–261.

Beauchaine, T. P., Strassberg, Z., Kees, M. R., & Drabick, D. A. G. (2001). Cognitive response repertoires to child noncompliance by mothers of aggressive boys. *Journal of Abnormal Child Psychology, 30,* 89–101.

Blanchard, J. J., Gangestad, S. W., Brown, S. A., & Horan, W. P. (2000). Hedonic capacity and schizotypy revisited: A taxometric analysis of social anhedonia. *Journal of Abnormal Psychology, 109,* 87–95.

Blashfield, R. K., & Aldenderfer, M. S. (1988). The methods and problems of cluster analysis. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experi-*

mental psychology (2nd ed., pp. 447–473). New York: Plenum Press.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models.* Newbury Park, CA: Sage.

Cantor, N., & Genero, N. (1986). Psychiatric diagnosis and natural categorization: A close analogy. In T. Milton & G. Klerman (Eds.), *Contemporary directions in psychopathology: Toward the DSM–IV* (pp. 233–256). New York: Guilford Press.

Cantor, N., & Mischel, W. (1979). Prototypes in person perception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 12, pp. 3–52). Orlando, FL: Academic Press.

Cleland, C., & Haslam, N. (1996). Robustness of taxometric analysis with skewed indicators: I. A Monte Carlo study of the MAMBAC procedure. *Psychological Reports, 79,* 243–248.

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin, 90,* 218–244.

Crowell, J. A., & Treboux, D. (1995). A review of adult attachment measures: Implications for theory and research. *Social Development, 4,* 294–326.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243,* 1668–1674.

Erlenmeyer-Kimling, L., Golden, R. R., & Cornblatt, B. A. (1989). A taxometric analysis of cognitive and neuromotor variables in children at risk for schizophrenia. *Journal of Abnormal Psychology, 98,* 203–208.

Flanagan, E. H., & Blashfield, R. K. (2002). Psychiatric classification through the lens of ethnobiology. In L. E. Beutler & M. L. Malik (Eds.), *Rethinking the DSM: A psychological perspective. Decade of Behavior* (pp. 121–145). Washington, DC: American Psychological Association.

Fraley, R. C., & Spieker, S. J. (in press). Are infant attachment patterns continuously or categorically distributed? A taxometric analysis of Strange Situation behavior. *Developmental Psychology.*

Fraley, R. C., & Waller, N. G. (1998). Adult attachment patterns: A test of the typological model. In J. A. Simpson & W. S. Rholes (Eds.), *Attachment theory and close relationships* (pp. 74–114). New York: Guilford Press.

Gangestad, S. W., Bailey, J. M., & Martin, N. G. (2000). Taxometric analyses of sexual orientation and gender identity. *Journal of Personality and Social Psychology, 78,* 1109–1121.

Gangestad, S. W., & Snyder, M. (1985). “To carve nature at its joints”: On the existence of discrete classes in personality. *Psychological Review, 92,* 317–349.

George, C., Kaplan, N., & Main, M. (1985). *The adult attachment interview.* Unpublished manuscript, University of California at Berkeley.

Golden, R. R., & Meehl, P. E. (1979). Detection of the schizoid taxon with MMPI indicators. *Journal of Abnormal Psychology, 88,* 212–233.

Grayson, D. A. (1987). Can categorical and dimensional views of psychiatric illness be distinguished? *British Journal of Psychiatry, 151,* 355–361.

Grove, W. M., & Meehl, P. E. (1993). Simple regression-based procedures for taxometric investigations. *Psychological Reports, 73,* 707–737.

Haertel, E. H. (1990). Continuous and discrete latent structure models for item response data. *Psychometrika, 55,* 477–494.

Harris, G. T., Rice, M. E., & Quinsey, V. L. (1994). Psychopathy as a taxon: Evidence that psychopaths are a discrete class. *Journal of Consulting and Clinical Psychology, 62,* 387–397.

Haslam, N., & Beck, A. T. (1994). Subtyping major depression: A taxometric analysis. *Journal of Abnormal Psychology, 103,* 686–692.

Hoyt, W. T., & Kerns, M. D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4,* 403–424.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3,* 424–453.

Jöreskog, K. G., & Sörbom, D. (2001). *LISREL 8: User’s reference guide* [Computer software manual]. Lincolnwood, IL: Scientific Software International.

Kendell, R. E. (1989). Clinical validity. *Psychological Medicine, 19,* 45–55.

Klein, D. N., & Riso, L. P. (1993). Psychiatric disorders: Problems of boundaries and comorbidity. In C. G. Costello (Ed.), *Basic issues in psychopathology* (pp. 19–66). New York: Guilford Press.

Korfine, L., & Lenzenweger, M. F. (1995). The taxonicity of schizotypy: A replication. *Journal of Abnormal Psychology, 104,* 26–31.

Lenzenweger, M. F. (1999). Deeper into the schizotypy taxon: On the robust nature of maximum covariance analysis. *Journal of Abnormal Psychology, 108,* 182–187.

Lenzenweger, M. F., & Korfine, L. (1992). Confirming the latent structure and base rate of schizotypy: A taxometric analysis. *Journal of Abnormal Psychology, 101,* 567–571.

Macmillan, N. A., & Creelman, C. D. (1990). Response bias: Characteristics of detection theory, threshold theory and "nonparametric" indices. *Psychological Bulletin, 107,* 401–413.

Malt, B. (1993). Concept structure and category boundaries. In G. V. Nakamura, D. L. Medin, & R. Taraban (Eds.), *Categorization by humans and machines. The psychology of learning and motivation: Advances in research and theory* (Vol. 29, pp. 363–390). San Diego, CA: Academic Press.

Marsh, H. W., Balla, J. W., & McDonald, R. P. (1988). Goodness-of-fit indices in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391–411.

Masters, G. N. (1985). A comparison of latent trait and latent class analyses of Likert-type data. *Psychometrika, 50,* 69–82.

Meehl, P. E. (1973). MAXCOV-HITMAX: A taxonomic search method for loose genetic syndromes. In P. E. Meehl (Ed.), *Psychodiagnosis: Selected papers* (pp. 200–224), Minneapolis: University of Minnesota Press.

Meehl, P. E. (1995a). Bootstraps taxometrics: Solving the classification problem in psychopathology. *American Psychologist, 50,* 266–275.

Meehl, P. E. (1995b). Extension of the MAXCOV-HITMAX taxometric procedure to situations of sizable nuisance covariance. In D. Lubinski & R. V. Dawis (Eds.), *Assessing individual differences in human behavior: New concepts, methods, and findings* (pp. 81–92). Palo Alto, CA: Davies-Black.

Meehl, P. E. (1996). MAXCOV pseudotaxonicity. *American Psychologist, 51,* 1184–1186.

Meehl, P. E. (2001). Comorbidity and taxometrics. *Clinical Psychology: Science and Practice, 8,* 507–519.

Meehl, P. E., & Yonce, L. J. (1994). Taxometric analysis: I. Detecting taxonicity using two quantitative indicators using means above and below a sliding cut (MAMBAC procedure). *Psychological Reports, 74,* 1059–1274.

Meehl, P. E., & Yonce, L. J. (1996). Taxometric analysis: II. Detecting taxonicity using covariance of two quantitative indicators in successive intervals of a third indicator (MAXCOV procedure). *Psychological Reports, 78,* 1091–1227.

Meyer, T. D., & Keller, F. (2001). Exploring the latent structure of the perceptual aberration, magical ideation, and physical anhedonia scales in a German sample. *Journal of Personality Disorders, 15,* 521–535.

Murphy, E. A. (1964). One cause? Many causes? The argument from the bimodal distribution. *Journal of Chronic Disease, 17,* 301–324.

Nagin, D. S. (1999). Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychological Methods, 4,* 139–157.

Nagin, D. S., & Tremblay, R. E. (2001). Analyzing developmental trajectories of distinct but related behaviors: A group-based method. *Psychological Methods, 6,* 18–34.

Rajendar, G. K. (1996). The influence of positive and negative wording and issue involvement on responses to Likert scales in marketing research. *Journal of the Market Research Society, 38,* 235–246.

Rosch, E., & Lloyd, B. (1978). *Cognition and categorization.* New York: Wiley.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88,* 413–428.

Semin, G. R., & Rosch, E. (1981). Activation of bipolar prototypes in attribute inferences. *Journal of Experimental Social Psychology, 17,* 472–484.

Simon, D., Pham, L. B., Le, Q. A., & Holyoak, K. J. (2001). The emergence of coherence over the course of decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1250–1260.

Skilling, T. A., Quinsey, V. L., & Craig, W. M. (2001). Evidence of a taxon underlying serious antisocial behavior in boys. *Criminal Justice and Behavior, 28,* 450–470.

Smith, E. E. (1995). Concepts and categorization. In E. Smith & D. Osherson (Eds.), *Thinking: An invitation to cognitive science.* (Vol. 3, 2nd ed., pp. 3–33). Cambridge, MA: MIT Press.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis.* Thousand Oaks, CA: Sage.

Snyder, J., Edwards, P., McGraw, K., Kilgore, K., & Holton, A. (1994). Escalation and reinforcement in mother-child conflict: Social processes associated with the development of physical aggression. *Development and Psychopathology, 6,* 305–321.

Sroufe, L. A., & Waters, E. (1977). Attachment as an organizational construct. *Child Development, 48,* 1184–1199.

Statistica [Computer software]. (1998). Tulsa, OK: Statsoft.

Strube, M. J. (1989). Evidence for the Type in Type A behavior: A taxometric analysis. *Journal of Personality and Social Psychology, 56,* 972–987.

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models.* Newbury Park, CA: Sage.

Tyrka, A. R., Cannon, T. D., Haslam, N., Mednick, S. A., Schulsinger, F., Schulsinger, H., & Parnas, J. (1995). The latent structure of schizotypy: I. Premorbid indicators of a taxon of individuals at risk for schizophrenia-spectrum disorders. *Journal of Abnormal Psychology, 104,* 173–183.

Waldman, I. D., & Lilienfeld, S. O. (2001). Applications of taxometric methods to problems of comorbidity: Perspectives and challenges. *Clinical Psychology Research and Practice, 8,* 520–527.

Waller, N. G., & Meehl, P. E. (1998). *Multivariate taxometric procedures: Distinguishing types from continua.* Newbury Park, CA: Sage.

Waller, N. G., Putnam, F. W., & Carlson, E. B. (1996). Types of dissociation and dissociative types: A taxometric analysis of dissociative experiences. *Psychological Methods, 1,* 300–321.

Waller, N. G., & Ross, C. A. (1997). The prevalence and biometric structure of pathological dissociation in the general population: Taxometric and behavior genetic findings. *Journal of Abnormal Psychology, 106,* 499–510.

Waters, E., & Beauchaine, T. P. (2003). Are there really patterns of attachment? Comment on Fraley and Spieker (2003). *Developmental Psychology, 39,* 417–422.

Williamson, D. A., Womble, L. G., Smeets, M. A. M., Netemeyer, R. G., Thaw, J. M., Kutlesic, V., & Gleaves, D. H. (2002). Latent structure of eating disorders symptoms: A factor analytic and taxometric investigation. *American Journal of Psychiatry, 159,* 412–418.

Woodward, S. A., Lenzenweger, M. F., Kagan, J., Snidman, N., & Arcus, D. (2000). Taxonic structure of infant reactivity: Evidence from a taxometric perspective. *Psychological Science, 11,* 296–301.

Received June 13, 2002
Revision received October 28, 2002
Accepted November 1, 2002 ■