Accuracy of retrospective memory and covariation estimation in patients with obsessive–compulsive disorder

Andrew T. Glostera,*, David C.S. Richardb, Joseph Himlec,d, Ellen Koche, Heather Ansone, Laura Lokersd, James Thorntonf

a Institute for Clinical Psychology and Psychotherapy, Technical University of Dresden, Germany
b Department of Psychology, Rollins College, USA
c School of Social Work, University of Michigan, USA
d Department of Psychiatry, University of Michigan, USA
e Department of Psychology, Eastern Michigan University, USA
f Department of Economics, Eastern Michigan University, USA

Received 31 December 2007; received in revised form 27 February 2008; accepted 27 February 2008

Abstract

Assessment methods relying on biased or inaccurate retrospective recall may distort knowledge about the nature of disorders and lead to faulty clinical inferences. Despite concerns about the accuracy of retrospective recall in general and in particular with obsessive–compulsive disorder (OCD) patients, the accuracy of retrospective recall for one’s own symptoms assessed in vivo is unknown in this population. This study used a prospective ecological momentary assessment (EMA) methodology to create a criterion against which to assess recall accuracy in OCD patients. Although results indicated that patients’ retrospective recall of OCD symptoms was fairly accurate, they consistently overestimated the magnitude of OCD symptom covariation with non-OCD facets (e.g., sleep duration, contemporaneous stress level, etc.). Findings suggest that even when recall of OCD symptoms is accurate, patients may be inaccurate in estimating symptom covariation. The findings have implications for the research, case conceptualization, and assessment of OCD, and may extend to other disorders.

Keywords: Obsessive–compulsive disorder (OCD); Retrospective recall; Covariation estimation; Memory bias; Ecological momentary assessment (EMA)

Accurate clinical case conceptualization frequently relies on a client’s ability to retrospectively recall behavioral dimensions of problems. Clinical interviews and questionnaires presuppose that a client’s recall is reasonably accurate, although research has questioned this assumption. For example, data collected in vivo and patients’ retrospective recall have shown discrepancies for estimates of pain intensity (Stone, Broderick, Shiffman, & Schwartz, 2004), frequency of eating behaviors (Stein & Corte, 2003), use of coping behaviors (Stone et al., 1998), frequency of panic attacks (de Beurs, Lange, & Van Dyck, 1992), timing of smoking lapses...
(Shiffman et al., 1997), and presence of anxious cognitions (Marks & Hemsley, 1999). Inaccurate patient recall has the potential to negatively influence treatment planning and compromise treatment efficacy (Haynes, Leisen, & Blaine, 1997).

No studies to date have assessed recall accuracy of ideographic and ecologically valid symptoms in patients with obsessive–compulsive disorder (OCD). This is problematic for several reasons. First, some laboratory studies have documented memory biases in patients with OCD (e.g., Deckersbach et al., 2002; Fontenelle, Mendelowicz, Mattos, & Versiani, 2006; Savage et al., 1996; Zitterl et al., 2001). Such deficits have been hypothesized to contribute to the maintenance of this disorder. However, not all studies have found evidence of memory deficits (Jelinek, Moritz, Heeren, & Naber, 2006; Moritz, Jacobsen, Willenborg, Jelinek, & Fricke, 2006). Based on extant studies, it appears as if recall accuracy varies as a function of experimental methodology and type of memory assessed (for a review see Muller & Roberts, 2005) and may be moderated by depressive symptoms (Wilhelm, McNally, Baer, & Florin, 1997). Based on extant studies, it appears as if recall accuracy varies as a function of experimental methodology and type of memory assessed (for a review see Muller & Roberts, 2005) and may be moderated by depressive symptoms (Wilhelm, McNally, Baer, & Florin, 1997).

A caveat to interpreting the extant literature is that previous laboratory studies have employed standardized, rather than personally relevant stimuli. One exception is the study of Tolin et al. (2001) in which participants recalled objects after six 10-s presentations, which they had rated as safe, neutral, or unsafe. In this study, participants with OCD recall accuracy of the presented objects did not differ from anxious and non-anxious controls, but did differ with respect to lower confidence in their memory. The ecological validity of this study was limited, however, because the participants remained in the laboratory. Ecologically valid studies in which participants interact with their natural environment are needed to test the generalizability of laboratory results, especially with regard to accuracy in recall of clinically relevant autobiographical behavior. To date, the accuracy of ecologically valid retrospective recall has been investigated primarily with regard to overt behaviors (e.g., smoking, drinking, purging, etc.). The relative accuracy of retrospectively obtained reports of latent processes (e.g., thoughts, emotions, moods, etc.) is less precisely understood. This is troubling given that many theories of anxiety and mood disorders posit a central role for cognition as a causal or maintaining factor (e.g., Beck & Emery, 1985).

If the accuracy of recalled behavioral intensities, frequencies, and durations are compromised, it is also possible that patients inaccurately estimate covariation of different behaviors assumed to maintain the disorder. For instance, the common clinician inquiry as to what factors may exacerbate a target problem requires patients to judge covariance of symptoms, mood, and environmental stimuli. Possible examples range from the general “What makes your symptoms better?” to specific items on assessment instruments such as “How much distress do your obsessive thoughts cause you?” (Yale-Brown Obsessive–Compulsive Scale (Y-BOCS); Goodman et al., 1989). To our knowledge, no studies have directly examined clinical participant’s accuracy in estimating covariations of one’s own symptoms. Other lines of research suggest, however, that patients will have difficulty accurately reporting covariation of their own behavior. For example, using a non-clinical population O’Brien (1995) presented advanced clinical psychology graduate students hypothetical self-monitoring data of a patient’s target behaviors (e.g., headache frequencies, durations, and intensities) and hypothesized controlling factors (e.g., stress level, number of arguments, sleep duration, and number of pain killers taken) and asked them to estimate variable intercorrelations. Students overinflated estimations of weak relationships while underestimating the strength of strong relationships. O’Brien’s findings were consistent with other research on illusory correlation (de Jong, Merckelbach, Bögels, & Kindt, 1998; Pauli, Montoya, & Martz, 2001; Tomarken, Mineka, & Cook, 1989) and basic studies on judgment (Kahneman, Slovic, & Tversky, 1982; Nisbett & Ross, 1980), which has shown that humans are largely inaccurate when estimating...
covariations in laboratory tasks. However, the accuracy of patients’ estimations of covariation magnitudes for their own clinically relevant behaviors remains unknown.

The first purpose of the current study was to extend previous laboratory research by assessing the degree of discrepancy between criterion data obtained in vivo and retrospective recall in OCD patients over a 1-week period. This time frame was selected to reflect the typical interval assessed in psychotherapy. The second purpose was to investigate accuracy of OCD patients in estimating symptom covariation. Three a priori hypotheses were tested. First, to guard against erroneous interpretations, we hypothesized that in vivo data collection would show reactivity. Second, we hypothesized OCD patients would inaccurately recall behavioral dimensions of their OCD symptoms. Third, we hypothesized OCD patients would inaccurately estimate covariation of OCD symptoms with other contextual variables and non-OCD behaviors. In addition to testing the generalizability of laboratory findings of memory bias in OCD, this study aimed to examine patients’ ability to recall types of information commonly asked in psychiatric interviews and standard assessment instruments.

Method

Sample

Participants

Fifty-nine outpatients with diagnoses of OCD from an anxiety disorders clinic at a major university medical school in Midwestern USA were assessed for eligibility. In addition to the structured clinical interview used with the patients enrolled in the study, all patients were previously diagnosed with OCD at the anxiety clinic as established by a clinician in conjunction with multidisciplinary team review. Following screening, 43 patients were enrolled in the study after 16 were excluded for not meeting a priori screening requirements. Seven participants subsequently dropped out and one was removed after completing the study due to handheld computer failure. Of the seven dropouts, two participants cited time constraints. We were unable to contact the other five.

The final sample was 35 treatment-seeking OCD patients, 48.57% (n = 17) of whom were female. The sample was mostly Caucasian (n = 29, 82.86%), with Asian (n = 4, 11.43%), African-American (n = 1, 2.86%), and Hispanic (n = 1, 2.86%) individuals represented. The mean age was 36.31 years (SD = 12.21 years, range = 20–62) with 15.88 years of education (SD = 2.46). Participants received a $25 gift card to a local retailer for their participation.

Inclusion and exclusion criteria

Inclusion criteria were: (a) OCD diagnosis using the Anxiety Disorders Interview Schedule (ADIS-IV; Brown, DiNardo, & Barlow, 1994); (b) a score of 16 or more on the Y-BOCS (Goodman et al., 1989) or greater than 18 on the Obsessive–Compulsive Inventory–Revised (OCI-R; (Foa et al., 2002) and (c) age 18–65. Exclusion criteria were: (a) co-morbidity with disorders associated with memory impairment (i.e., psychotic disorder, substance dependence, dementia, and bi-polar disorder); (b) disabilities that prohibited participation (e.g., poor vision, inability to hear a handheld computer beep); (c) English illiteracy; and (d) suicidality.

Measures

In addition to a demographic inventory, the following measures were administered:

**Obsessive–Compulsive Inventory–Revised (OCI-R; Foa et al., 2002)**

The OCI-R is the short-form of the Obsessive-Compulsive Inventory-Revised (Foa et al., 2002; Foa, Kozak, Salkovskis, Coles, & Amir, 1998). The OCI-R contains six subscales, each measuring a facet of OCD (i.e., washing, obsessing, hoarding, ordering, checking, and neutralizing). Each subscale consists of three items, for a total of 18 items. Internal consistency for the OCI-R subscales ranged from .82 to .90 and 2-week test–retest reliability for the subscales ranged from .74 to .91 (Foa et al., 2002). Using a 5-point scale ranging from 1 to 5, each question asks “how much the experience has distressed or bothered you during the past month.” To
properly contextualize questions for this study, the reference interval for questions was changed from the past month to the past week.

**Yale-Brown Obsessive–Compulsive scale (Y-BOCS; Goodman et al., 1989)**

The Y-BOCS is a clinician-administered interview designed to measure severity of obsessions and compulsions in the past week. The Y-BOCS consists of 10 items, five each for obsessions and compulsions. One week test–retest reliability for the Y-BOCS was .90 (Kim, Dysken, & Kuskowski, 1990) and inter-rater reliability was $r = .98$ (Goodman et al., 1989). Alpha coefficients have ranged from .88 (Goodman et al., 1989) to .69 (Woody, Steketee, & Chambless, 1995).

**Beck Depression Inventory-II (BDI-II; Beck, Steer, & Brown, 1996)**

The BDI-II is a 21-item self-report questionnaire that measures the presence and severity of depressive symptoms and suicidal ideation (Beck et al., 1996). One week test–retest reliability for the BDI-II is .93 and internal consistency was .92 among outpatients (Beck et al., 1996).

**Anxiety Disorders Interview Schedule for DSM-IV (Adult version) (ADIS-IV; Brown et al., 1994)**

The ADIS-IV is a semi-structured assessment interview for differential diagnosis of DSM-IV Axis I and anxiety disorders. Kappa’s for the ADIS-IV have been between .60 and .86 across disorders, except for Dysthymic Disorder ($\kappa = .22$ (Brown, DiNardo, Lehman, & Campbell, 2001). The ADIS-IV uses a 0–8 scale, with 4 being the cutoff for diagnostic presence. The ADIS-IV was used to confirm participants were accurately diagnosed with OCD and to assess for comorbid conditions. Subsections not relevant to the study were excluded.

**Assessors**

Assessors were advanced doctoral clinical psychology students and a clinical social worker (MSW). All viewed training video vignettes of the clinician-administered interviews (Y-BOCS and ADIS-IV). Assessors’ ratings were required to concur with at least 80% of those of an expert on training vignettes of the Y-BOCS and ADIS-IV. Any discrepancies from the criterion were discussed and the rationale for the correct rating was provided by the principal investigator. In addition, each assessor observed one Y-BOCS and ADIS-IV interview before administering at least two supervised interviews. During these observations, ratings between the principal investigator and the trainees were compared and disagreements discussed. If rating discrepancies exceeded 1 point on either the Y-BOCS total score or ADIS-IV diagnostic severity scale, assessors were required to review the training videos.

**Procedure**

The study was conducted over 2 weeks with observations in weekly intervals. Time 1 observations occurred on the first day of participation. Time 2 observations occurred on the 8th day of the study. Time 3 observations occurred on the 15th day of the study. Each point of contact with participants (i.e., Times 1–3) is described in detail below.

**Time 1 (screening assessment)**

Prospective participants were screened by telephone and informed of their research rights with verbal informed consent solicited prior to data collection. During the telephone screening, participants answered demographic questions and verbally completed the OCI-R, Y-BOCS, and BDI-II. Questions were modified to query for symptoms over the previous week. Participants meeting inclusion criteria were scheduled for ecological momentary assessment (EMA) training at Time 2. Finally, participants were instructed to informally monitor symptoms during the ensuing week with the statement: “Please pay attention to your symptoms over the next week”.
Time 2 (EMA training)

One week later, participants met with a researcher and signed the informed consent. Participants verbally completed the OCI-R, Y-BOCS, and BDI-II a second time. As before, OCI-R questions were modified to query for the previous week. OCD diagnosis was confirmed by having participants complete the ADIS-IV.

Participants then received a Palm Zire 21 handheld computer (personal data assistant, PDA) and were instructed in its use. The handheld computer had a monochrome, 5 cm × 5 cm screen. Data were recorded using stylus input. Training included a demonstration of how to operate the handheld computer, initiate data recording, record data, and recognize the signaling tone. Participants practiced recording data, asked questions, and were informed the PDA time stamped all responses. Participants were also informed that each time they completed an assessment within 15 min of being signaled, their name would be entered in a lottery giveaway for a PDA. A notepad in the PDA case allowed participants to record any PDA difficulties or malfunctions. Participants were asked to return 7 days later for their last appointment (Time 3). Participants were contacted during the 1st, 3rd, and 6th days of EMA self-monitoring to address any questions regarding the use of the PDA.

In the week interval between Times 2 and 3 (days 8 to 14 of the study) participants completed questionnaires on the handheld computer contingent on an audible signal every 4 h (i.e., 10 a.m, 2 p.m, 6 p.m., and 10 p.m.). Item stems read, “In the last four hours…” followed by the 18 OCI-R items. Participants monitored washing, obsessing, hoarding, ordering, checking, and neutralizing behaviors because these had the strongest OCI-factor loadings and represented common OCD behaviors (Foa et al., 2002).

The computer algorithm for each item first presented a screening question to determine whether the participant had engaged in the target OCD behavior in the previous 4 h. If not, the algorithm skipped to the next OCD item. Otherwise, the program branched to two questions assessing frequency and duration of the symptom. Following the 18 OCD items, the computer-administered supplemental items that assessed previous night’s sleep duration, physical location, whether the participant was alone, mood, stress level, anxiety, loneliness, and whether he or she experienced distress following subjectively defined interpersonal conflict during the previous 4 h. These supplemental items represented variables commonly utilized in EMA studies, were measured using visual analog scales, and were used for the covariation analyses.

Time 3 (retrospective recall and covariation estimation)

On the last day of the study, participants returned the handheld computer and verbally completed the OCI-R, Y-BOCS, and BDI-II. Participants estimated the total frequency and average duration of each OCD symptom queried by the PDA over the preceding week. For example, participants were asked, (a) “In the last week (when you used the handheld computer), how many times did you…” and (b) “On average, how long did you spend doing … each day?” Retrospective estimates of daily behavioral frequencies and durations were then compared with EMA aggregated data of mean daily behavioral frequencies and durations. The difference between a participant’s retrospective estimate and the EMA collected data was the index of recall accuracy.

After estimating frequency and duration of all EMA-OCD items, participants completed a graphically supported, interactive, 27-slide PowerPoint didactic on the meaning of a correlation. The purpose of the didactic was to teach participants the meaning of covariation so they could reliably estimate covariation relationships between OCD symptoms and supplemental variables (e.g., amount of sleep, mood, etc.). Participants learned two major concepts: direction of relationship (i.e., positive, negative, and zero relation) and magnitude of relationship (i.e., strong, mild, weak). To simplify the learning process, participants learned correlations using a −100 (perfect negative correlation) to +100 (perfect positive correlation) scale instead of using decimals. Further, a visual analog scale that used a color-coded, double-pointed arrow helped to facilitate learning. Labels reading “positive correlation,” “negative correlation,” and “no correlation” anchored the arrow. Other descriptive labels further clarified the meaning of a given correlation’s direction and magnitude. Embedded in the tutorial were three concept quizzes to assess comprehension. Participants were required to successfully answer these questions before continuing. After completing the tutorial, participants estimated correlations between select OCD symptoms and supplemental EMA items encountered during the past week. Participants estimated correlations using the same color-coded visual analog scale used in training. For each item, the stem sentences learned during the tutorial were adapted for the specific content
of that question and appeared next to the arrow (e.g., “As sleep per night went up, counting went up” and “As sleep per night went down, counting went down”).

Because selection of behaviors for the covariation estimation task required taking into account which behaviors each participant experienced, behaviors with the highest frequency on each OCI-R subscale were selected for participants. Other high-frequency behaviors were also included, up to a total of six behaviors. Thus, depending upon the participant, correlations between supplemental EMA items and three–six OCD behaviors were estimated.

Participants were then asked to rate their confidence in their memory, perceived accuracy in symptom recall, and perceived accuracy in covariation judgments. Finally, they were asked about their awareness of the study’s hypotheses.

**Results**

**Participants’ symptom severity**

During the telephone screening interview at Time 1, participants completed the OCI-R, Y-BOCS, and BDI-II. Mean scores for participants at Time 1 were as follows: OCI-R total score = 25.80 (SD = 9.38), Y-BOCS severity score = 22.10 (SD = 3.93), and BDI-II total score = 22.00 (SD = 13.61). At Time 2, participants completed the ADIS-IV to verify Axis I diagnoses. As measured by the ADIS-IV, the mean OCD severity level was 5.44 (SD = 0.81). The mean number of comorbid diagnoses in the sample was 1.31 (SD = 1.13). Social phobia occurred most frequently (n = 10), followed by major depression (n = 9), panic/agoraphobia (n = 8), generalized anxiety disorder (n = 4), specific phobia (n = 2), posttraumatic stress disorder (n = 1), dysthymia (n = 1), body dysmorphic disorder (n = 1), and trichotillomania (n = 1).

**Data screening**

Prior to analysis, data were screened for outliers. For EMA data, an outlier was defined as a data point that was at least 10 times greater than the mean value for the same item for any given participant over the course of the EMA week. First, impossible values (e.g., recording 700 min as the duration of a behavior in a 4 h period) were identified. Three participants recorded impossible values. Given the nature of recording on a handheld computer, inadvertent recording errors were also possible and inconsistencies within participant across the week were identified. When previously recorded values clearly pointed to recording errors (e.g., 120 min recorded consistently across the week followed by 1200 min), the suspect value was changed to be consistent with the rest of that participant’s data. This occurred 4 times out of 21,800 participant entries. The remaining outliers were changed to zero to prevent the artificial inflation of EMA values and guard against changes that favored experimental hypotheses.

**Response rate**

Overall, participants responded to 872 of a possible 969 EMA observations for an overall response rate of 90.0%. Of these, 702 (80.5%) occurred within ± 15 min of the audible signal, 752 (86.2%) within ± 30 min, and 798 (91.5%) within ± 60 min. The mean latency between signal and response was 21.14 min (SD = 54.89). The response rate for this study is comparable to response rates in previous studies (i.e., 94%, Stone et al., 2004).

**Reactivity (Hypothesis 1)**

In order to assess the presence of reactivity, one-way within-subjects ANOVAs were conducted across the three time points independently for the total scores from the OCI-R, BDI-II, and Y-BOCS. The mean and standard deviations for these measures are presented in Table 1. With respect to the OCI-R total score, results indicated a significant time effect, Wilk's Λ = .65, F(2,33) = 8.78, p < .001, multivariate η² = .35. Follow-up polynomial contrasts indicated a significant linear effect with the mean OCI-R total score decreasing.
Table 1
Means and standard deviations of measures at Times 1–3

<table>
<thead>
<tr>
<th>Measure</th>
<th>Time 1</th>
<th>Time 2</th>
<th>Time 3</th>
<th>Overall F</th>
</tr>
</thead>
<tbody>
<tr>
<td>OCI-R total</td>
<td>25.80 (9.38)</td>
<td>21.04 (9.81)</td>
<td>19.97 (11.10)</td>
<td>8.78***</td>
</tr>
<tr>
<td>Washing</td>
<td>3.37 (3.57)</td>
<td>2.69 (2.89)</td>
<td>2.29 (2.59)</td>
<td></td>
</tr>
<tr>
<td>Obsessing</td>
<td>6.71 (3.56)</td>
<td>5.57 (3.63)</td>
<td>5.86 (3.85)</td>
<td></td>
</tr>
<tr>
<td>Hoarding</td>
<td>4.26 (3.94)</td>
<td>2.83 (3.62)</td>
<td>2.83 (3.16)</td>
<td></td>
</tr>
<tr>
<td>Ordering</td>
<td>4.46 (3.39)</td>
<td>4.27 (3.54)</td>
<td>3.74 (3.78)</td>
<td></td>
</tr>
<tr>
<td>Checking</td>
<td>3.89 (3.18)</td>
<td>2.91 (3.32)</td>
<td>2.83 (3.11)</td>
<td></td>
</tr>
<tr>
<td>Neutralizing</td>
<td>3.11 (3.45)</td>
<td>2.77 (3.82)</td>
<td>2.43 (3.93)</td>
<td></td>
</tr>
<tr>
<td>Y-BOCS total</td>
<td>22.10 (3.93)</td>
<td>22.20 (5.01)</td>
<td>20.91 (4.57)</td>
<td>2.93</td>
</tr>
<tr>
<td>Obsession</td>
<td>11.23 (2.22)</td>
<td>11.01 (3.02)</td>
<td>10.13 (2.62)</td>
<td></td>
</tr>
<tr>
<td>Compulsion</td>
<td>10.87 (2.38)</td>
<td>11.01 (2.47)</td>
<td>10.79 (2.60)</td>
<td></td>
</tr>
<tr>
<td>BDI-II total</td>
<td>22.00 (13.61)</td>
<td>18.23 (13.92)</td>
<td>17.69 (13.86)</td>
<td>10.50***</td>
</tr>
</tbody>
</table>

Note: * p < .05; ** p < .01; *** p < .001; n = 35 for all comparisons.

Patients were identified consistent with HIPPA regulations and as approved by the institutional review board.

Paired-samples t-tests were also conducted for this hypothesis and failed to yield any significant differences across any comparisons.

1Subjects were identified consistent with HIPPA regulations and as approved by the institutional review board.

2Paired-samples t-tests were also conducted for this hypothesis and failed to yield any significant differences across any comparisons.

Retrospective recall of OCD symptom frequency and duration (Hypothesis 2)

At Time 3, participants were asked to recall the total frequency and duration of behaviors measured by the EMA-OCD items over the previous week. Random intercept regression was used to analyze participants’ retrospective recall accuracy relative to the EMA criterion data. Random-intercepts regression capitalizes on the repeated measures nature of the data, in which error is partialled into separate terms. The failure to do so, as in a t-test, leads to an increased risk of Type II error (Hedeker, 2004; Rowland & Thornton, 2003). The random intercept regression models used a feasible generalized least square (FGLS) estimator. FGLS accounts for correlation among the errors, thus yielding unbiased and efficient estimates of recall accuracy. Dummy variables were coded to allow for statistical testing of the effect of the variable in question (e.g., retrospective recall vs. EMA mean). The error term $c_{it}$ was broken down into $u_i$ (unexplained variance that varies across participants but not over time) and $e_{it}$ (unexplained variance that varies across participants and time). This procedure capitalized on the richness of EMA data, accounted for unobserved variables that varied across individuals, and thereby decreased the risk of Type II error. Finally, this procedure is one of the most efficient and accurate means of accounting for missing data (King, King, & Bachrach, 2001). For these analyses, participants’ EMA-OCD data were summed per day.
I thought about good and bad numbers. I arranged things in a particular order. I avoided throwing something away because I was afraid I might need it later. I washed my hands longer than necessary. I found it difficult to control my own thoughts. I found myself repeatedly counting objects. I found myself not wanting to touch an object touched by someone else. I had troubling thoughts that were difficult to get rid of. I saved or collected something I don’t need. I checked stove burners, water taps, and/or light switches after turning them off. I arranged things in a particular order. I thought about good and bad numbers. I washed or cleaned myself because I felt contaminated. I repeated certain numbers resulted in a significant coefficient of recall. With the exception of these two items, differences between retrospectively recalled average daily duration and EMA values did not exceed 15 min. This suggests that after accounting for the predictor variables, participants retrospectively recalled a mean of 139.3 more instances per day of “repeating certain numbers” than what was recorded on the handheld computer.

**Behavioral duration**

The second set of random-intercepts regression analyses examined participants’ retrospective recall accuracy with respect to average daily behavioral duration of the OCD behaviors after accounting for the same variables identified above and noted in Table 3. Coefficients for the dummy variable recall (EMA vs. retrospective recall of behavioral duration) are presented in Table 3 along with inferential statistics for each EMA-OCD item’s regression model. Only the items measuring repeatedly counting objects and repeating certain numbers resulted in a significant coefficient of recall. With the exception of these two items, differences between retrospectively recalled average daily duration and EMA values did not exceed 15 min. This suggests that after accounting for the predictor variables, participants were relatively accurate when retrospectively recalling average daily behavioral durations. The two EMA-OCD items with significant coefficients of recall indicated that participants’ retrospective recall of average daily durations was overestimated by 36 min for...
repeatedly counting objects (item 4), \( z(1.25) = 2.88, p < .004 \), and 28 min for repeating certain numbers (item 10), \( z(10) = 2.44, p < .05 \).

The implication of the results is that, on average, our sample of OCD patients appeared to recall reasonably accurately both the frequency and duration of nearly all OCD symptoms.

**Retrospective estimations of symptom covariation (Hypothesis 3)**

To test the hypothesis that participants’ estimates of symptom covariation were different from those found in the EMA data, participants estimated the relationship of selected OCD items with the six supplementary variables. For example, a participant might estimate the relationship between obsessive thinking and sleep duration, mood, stress level, anxiety, loneliness, and distress following an interpersonal fight. After these estimates were completed, the process was reiterated for a second OCD symptom with the same supplemental variables. Each participant provided a minimum of 18 covariation estimates (at least 3 OCD behaviors \( \times 6 \) supplemental variables).

A series of paired-samples \( t \)-tests were conducted to evaluate whether retrospective covariation estimates (hereafter referred to as estimated correlations) were statistically different from correlations computed from the EMA data (hereafter referred to as EMA correlations). Estimated correlations were compared with EMA correlations separately for behavioral frequencies and behavioral duration. Using a Bonferroni family-wise correction, \( z \) was set at .008 for any individual comparison. Table 4 presents means, standard deviations, and test statistics of estimated and EMA correlations collapsed across individuals, behavior, and rank-order of behavioral frequency and duration.

Results showed that participants’ estimated correlations frequently departed from EMA correlations to a degree that could not be accounted for by chance or measurement error. In general, participants consistently

---

**Table 3**

Summary of random intercept regression analysis for average daily behavioral duration

<table>
<thead>
<tr>
<th>Item</th>
<th>df</th>
<th>Recall coeff.</th>
<th>( z )</th>
<th>( p )</th>
<th>95% CI</th>
<th>Overall ( R^2 )</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>I saved or collected something I don’t need</td>
<td>168</td>
<td>-1.65</td>
<td>-0.29</td>
<td>0.77</td>
<td>-12.87–9.56</td>
<td>0.06</td>
<td>0.00</td>
</tr>
<tr>
<td>I checked something I didn’t need to</td>
<td>186</td>
<td>-2.29</td>
<td>-0.35</td>
<td>0.72</td>
<td>-14.97–10.40</td>
<td>0.19</td>
<td>0.66</td>
</tr>
<tr>
<td>I got upset if things were not arranged properly</td>
<td>169</td>
<td>-2.12</td>
<td>-0.34</td>
<td>0.74</td>
<td>-14.41–10.17</td>
<td>0.21</td>
<td>0.12</td>
</tr>
<tr>
<td>I found myself repeatedly counting objects</td>
<td>125</td>
<td>36.08</td>
<td>2.88</td>
<td>0.004**</td>
<td>11.51–60.65</td>
<td>0.17</td>
<td>0.46</td>
</tr>
<tr>
<td>I found myself not wanting to touch an object touched by someone else</td>
<td>143</td>
<td>-5.94</td>
<td>-0.56</td>
<td>0.5</td>
<td>-26.70–14.82</td>
<td>0.19</td>
<td>0.19</td>
</tr>
<tr>
<td>I found it difficult to control my own thoughts</td>
<td>146</td>
<td>-9.95</td>
<td>-0.53</td>
<td>0.60</td>
<td>-47.00–27.10</td>
<td>0.20</td>
<td>0.69</td>
</tr>
<tr>
<td>I collected things I don’t need</td>
<td>134</td>
<td>4.42</td>
<td>1.22</td>
<td>0.22</td>
<td>-2.69–11.54</td>
<td>0.09</td>
<td>0.00</td>
</tr>
<tr>
<td>I repeatedly checked doors, windows, drawers, etc.</td>
<td>134</td>
<td>4.01</td>
<td>0.98</td>
<td>0.33</td>
<td>-4.04–12.07</td>
<td>0.11</td>
<td>0.49</td>
</tr>
<tr>
<td>I got upset when someone changed the way I arranged things</td>
<td>151</td>
<td>-1.80</td>
<td>-0.43</td>
<td>0.67</td>
<td>-10.07–6.46</td>
<td>0.21</td>
<td>0.43</td>
</tr>
<tr>
<td>I repeated certain numbers</td>
<td>100</td>
<td>28.596</td>
<td>2.44</td>
<td>0.015*</td>
<td>5.66–51.53</td>
<td>0.21</td>
<td>0.61</td>
</tr>
<tr>
<td>I washed or cleaned myself because I felt contaminated</td>
<td>169</td>
<td>-5.50</td>
<td>-0.18</td>
<td>0.86</td>
<td>-6.02–5.02</td>
<td>0.16</td>
<td>0.23</td>
</tr>
<tr>
<td>Unpleasant thoughts came into my mind against my will</td>
<td>222</td>
<td>-2.30</td>
<td>-0.09</td>
<td>0.93</td>
<td>-53.51–48.91</td>
<td>0.21</td>
<td>0.59</td>
</tr>
<tr>
<td>I avoided throwing something away because I was afraid I might need it later</td>
<td>151</td>
<td>-4.77</td>
<td>-0.19</td>
<td>0.85</td>
<td>-53.72–44.18</td>
<td>0.08</td>
<td>0.00</td>
</tr>
<tr>
<td>I checked stove burners, water taps, and/or light switches after turning them off</td>
<td>117</td>
<td>5.52</td>
<td>1.61</td>
<td>0.11</td>
<td>-1.19–12.24</td>
<td>0.12</td>
<td>0.11</td>
</tr>
<tr>
<td>I arranged things in a particular order</td>
<td>151</td>
<td>-4.91</td>
<td>-0.62</td>
<td>0.54</td>
<td>-20.45–10.62</td>
<td>0.38</td>
<td>0.13</td>
</tr>
<tr>
<td>I thought about good and bad numbers.</td>
<td>84</td>
<td>12.66</td>
<td>1.08</td>
<td>0.28</td>
<td>-10.33–35.64</td>
<td>0.37</td>
<td>0.19</td>
</tr>
<tr>
<td>I washed my hands longer than necessary</td>
<td>101</td>
<td>-8.00</td>
<td>-0.12</td>
<td>0.91</td>
<td>-14.38–12.77</td>
<td>0.10</td>
<td>0.00</td>
</tr>
<tr>
<td>I had troubling thoughts that were difficult to get rid of</td>
<td>246</td>
<td>-11.82</td>
<td>-0.34</td>
<td>0.73</td>
<td>-79.29–55.64</td>
<td>0.11</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Note: *\( p < .05 \); **\( p < .008 \). Models accounted for (a) the number of years participants reported OCD symptoms, (b) OCD severity as measured by the ADIS-IV at Time 2, (c) a proxy test for one-week memory, (d) self-reported confidence in the accuracy of their retrospective recall following the recall task, (e) prior knowledge of the hypothesis, (f) depression severity as measured by the BDI-II, (g) sex, and (h) age.
overestimated the degree of relationship between OCD symptoms and variables measured by the supplemental items. Significant overestimates were found for correlations between behavioral frequency and: stress \(t(117) = 7.58, p < .001, d = .70\); anxiety \(t(114) = 9.15, p < .001, d = .85\); and loneliness \(t(101) = 3.57, p < .001, d = .35\). Significant overestimates were found for correlations between behavioral duration and: stress \(t(117) = 7.02, p < .001, d = .64\); anxiety \(t(113) = 7.83, p < .001, d = .73\); loneliness \(t(101) = 4.04, p < .001, 40\); and distress following subjectively defined interpersonal fights \(t(30) = 2.88, p < .007, d = .52\). It is noteworthy that the absolute magnitude of estimated symptom covariation was always greater than the covariation present in the EMA data. Specifically, results indicate that the highest EMA correlation between OCD symptom fluctuation and any supplemental item was only .21 (for the relationship between OCD behavioral duration and self-reported levels of stress and anxiety) and accounted for just 4% of variation in OCD symptoms. Nevertheless, participants estimated a mean correlation of .51 between OCD symptoms and levels of stress and anxiety, thereby estimating that 26% of the variation in OCD symptoms could be explained by contemporaneous levels of stress and anxiety. Participants overestimated, but never underestimated, covariation magnitudes. This result suggests that participants understood the concept of covariation while simultaneously showing an overestimation bias, otherwise one should expect a random pattern of both under- and overestimations of relationship magnitudes.

**Discussion**

OCD patients in our sample were accurate reporters of their symptoms but consistently overestimated the magnitude of symptom covariation with non-symptomatic variables. Although inaccuracies with respect to retrospective recall of EMA-assessed criterion values were largely unsubstantiated, evidence of recall bias was observed in participants’ ability to estimate covariance of their symptoms. Results of this study also support the contention that EMA does not result in significant reactivity.

Results suggested that reactivity did not occur in response to the EMA recording method. Instead, reactivity in this study was observed only between Time 1 and 2. These results add to the growing body of literature that has failed to detect reactivity to EMA (e.g., Cruise, Broderick, Porter, Kaell, & Stone, 1996; Hufford, Shields, Shiffman, Paty, & Balabanis, 2002; Stein & Corte, 2003). The failure to detect reactivity during an EMA recall
paradigm is important because it eliminates one threat to internal validity, namely that differences noted during recall tasks are confounded by reactivity to EMA self-monitoring.

The current study also reiterated the importance of including a baseline period prior to the collection of EMA criterion values. Multiple factors likely caused the nearly 20% decline in symptom scores observed from Time 1 to 2. Hypothesized factors include increased awareness of actual symptoms, increased familiarity with questions asked during the assessment, instructions to “pay attention to your symptoms,” and improvement in symptoms by simply enrolling in the study. The sum of these effects resulted in statistically observable reactivity from Time 1 to 2, whereas the addition of EMA self-monitoring from Time 2 to 3 did not. Importantly, decreases in symptom severity observed from Time 1 to 2 occurred despite longstanding length of symptoms and treatment for most participants.

Results failed to detect consistent, statistically significant differences between retrospectively recalled OCD symptoms and EMA criterion data. These largely non-significant results were consistent for both behavioral frequency and duration. When participants were inaccurate, they tended to underestimate the frequency of their behaviors relative to the criterion value while overestimating duration. However, the coefficient of recall in the random-effects analyses suggested that across the 18 EMA-OCD behaviors, participants’ inaccuracies were of little clinical importance. Differences between retrospectively recalled daily frequencies and EMA calculated daily frequencies exceeded 15 occurrences per day only once (i.e., “I repeated certain numbers”). Similarly, average daily inaccuracy in estimated duration exceeded 15 min per day on only two variables (i.e., “I checked something I didn’t need to” and “I repeated certain numbers”). Most likely, the magnitude of differences between participants’ retrospective recall and EMA criterion data were not large enough to jeopardize the validity of a clinical case conceptualization. These results are consistent with recent findings that failed to detect significant memory dysfunction in OCD patients (Moritz et al., 2006).

These results stand in partial contrast to previous studies that reported inaccuracies in retrospective recall accuracy using EMA designs (Stein & Corte, 2003; Stone et al., 2004). Several factors may account for these differences. First, participants in this study recorded and recalled more behaviors (e.g., obsessions vs. bulimic episodes). Second, target behaviors in this study were recorded using a 0 to infinity scale for behavioral frequency and a 0–240 min scale for behavioral duration. In contrast, Stone et al. utilized a 100-point visual analog scale to record participants’ pain intensity. Finally, OCD patients may differ from eating disordered patients and pain patients in a way that predisposes them to be more accurate than other populations. Future research is needed to clarify which of these hypotheses is more plausible.

This study was the first to examine patients’ ability to estimate symptom covariation with supplemental variables. Participants consistently overestimated the magnitude of correlations between OCD behaviors and supplemental items. These effects were most robust for the relationship between OCD behaviors and the supplemental items stress and anxiety, though nearly all estimated correlations were overestimated.

Like graduate students, OCD patients tend to report illusory correlations. These results are consistent with O'Brien (1995), who found that advanced graduate students’ symptom covariation estimates based on hypothetical self-monitoring data were inaccurate. Graduate students overestimated weak correlations and underestimated strong correlations. This study extended O'Brien’s findings of the overestimation of weak correlations. Whether OCD patients also underestimate strong correlations, like the graduate students in O’Brien’s study, could not be tested because strong correlations calculated between OCD symptoms and supplemental items were not observed. Although the covariation estimation in the present study used different stimuli and required the patients to rely on retrospective memory, the results are largely consistent with the research base on illusory correlations (Kahneman et al., 1982; Nisbett & Ross, 1980).

Overestimates of symptom covariation could have a negligible impact on treatment conceptualization or could lead clinicians to concentrate on attempting to modify relationships that are illusory. Future studies are needed to understand the pervasiveness of the overestimation effect, its consequences, and the degree to which it can be observed in other patient and non-patient populations.

Overall, this study suggests that, in contrast to other populations (see Stein & Corte, 2003; Stone et al., 2004) OCD patients accurately recalled the frequency and duration of their OCD symptoms as measured in vivo. However, OCD patients overestimated symptom covariation with non-OCD variables. The importance of this finding lies in the following question: how accurate are patients’ reports of symptom covariation when provided in a clinical context? Our results suggest that researchers and clinicians should regard the accuracy of
covariation estimates obtained regularly in the course of clinical interviews and implicitly in many assessment instruments used in research with skepticism. The implications for clinical assessment and case formulation, especially assessment that relies on assessing functional relations between behaviors and variables that maintain their frequency and duration, could be significant. Specifically, treatment plans that are constructed from a foundation of largely inaccurate covariation estimation may be expected to be less efficacious than expected.

Although OCD patients accurately recalled dimensions of personally meaningful symptoms, they were largely inaccurate in the more complicated task of covariation estimation. This suggests that research on memory in OCD should carefully consider the complexity of the paradigm and the setting in which the tasks are given.

Replication of a covariation overestimation effect will require researchers to explain why it occurs. It is possible that the cognitive load required to consciously attend to factors that covary with symptoms exceeds the capacity of humans (Kahneman et al., 1982). Proponents of statistical decision making would likely endorse such a position (e.g., Garb, 1998).

Although the sample size in this study was large for an EMA study, additional participants would have increased statistical power. Because a number of OCD symptoms were not endorsed by all participants, the actual sample sizes for individual items were lower than the total \( N \) of 35 participants. The idiosyncratic nature of OCD meant that most participants endorsed only a portion of the EMA-OCD questions and that non-endorsed questions would nearly always lead to perfect, though theoretically less interesting, retrospective recall.

Although we acknowledge that the control condition (Week 1 of the study) could have been augmented by a non-OCD control group, we were concerned that OCD items would have had little relevance for non-affected controls who would not have endorsed many, if any, OCD items. As a result, the recall task for control subjects would have been less about recall and more about rote recall of “zeros” for both frequency and duration dimensions. This would render the recall task fundamentally different than for those affected by the disorder, and would result in unanalyzable data in the control group due to high levels of multicollinearity.

Finally, we acknowledge that the criterion data collected via the EMA is also subject to error. However, we agree with the observation that a gold standard need not be perfect, but rather the most accurate measure of behavior available (Kraemer, 1992).

Results of this study suggest several directions for continued inquiry. Future studies are needed to clarify why data from this study stand in contrast to results from previous investigations of retrospective recall accuracy using similar designs (Stein & Corte, 2003; Stone et al., 2004). Of primary interest is whether this study achieved divergent results as a function of the methodology, the frequency of EMA queries, the nature of the questions, or if OCD patients are predisposed to retrospectively recall their symptoms more accurately than other patients. Future research should also clarify the external validity of laboratory studies detecting inaccuracies in OCD patients on various measures of memory.

Because this is the first study to examine patients’ accuracy in estimating symptom covariation, further studies are needed to understand the implications of symptom overestimation and the factors that mediate covariation estimation errors, both with OCD and other disorders. Furthermore, the supplemental items utilized in the covariation estimation tasks represent only a fraction of those variables currently hypothesized to be related to OCD symptoms. Future studies should include other populations and variables to assess the generalizability of the present findings.

These results suggest that clinicians treating patients with OCD will receive relatively accurate retrospective reports from patients regarding frequency and duration of symptoms in the previous 7 days. However, the same cannot be said for covariation estimates. The latter result implies that case conceptualizations for OCD and possibly other disorders that include estimates of functional covariation may be of suspect validity if the main source of covariation estimation was the patient. Further, treatments based on inaccurate covariation estimates may be less potent than anticipated since a smaller proportion of variance in the OCD symptom will be shared with the covariate. Thus, modification of dimensions of the covariate should not be expected to have a collateral effect on the OCD symptom. An issue to consider is whether current treatment packages over- or underestimate covariation with variables that constitute significant components of the treatment plan.
Acknowledgments

This research was supported in part by the Blue Cross Blue Shield of Michigan Foundation (Grant no. 1045.SAP) and the Eastern Michigan University Graduate School Research Fund (Grant no. 116350), both awarded to the first author. Further, we wish to acknowledge the support of Lisa O'Donnel, Elizabeth Yaelingh-Scoffins, Elizabeth Nelson, and Susan Miller for coordinating and implementing data collection. Finally, we wish to thank Sylvia Helbig, Jens Klotsche, and Katrin Hummel for helpful comments on previous drafts.

References


