Global Ratings of Essays About Trauma: Development of the GREAT Code, and Correlations with Physical and Mental Health Outcomes

Bridget Klest
Jennifer J. Freyd

ABSTRACT. Past research has demonstrated in a variety of contexts that writing about emotional topics can benefit physical health and general well being. Most of this prior research has used the Linguistic Inquiry and Word Count program (LIWC, Pennebaker & Francis, 1996), but not global essay ratings, to assess what aspects of written essays might be associated with such benefits. Yet scoring rubrics are commonly used in the field of education to score global aspects of student writing. The current study used a sub-sample of essays from a larger research project on trauma, writing and health to develop a global rating rubric for essays about trauma based on rubrics used in education. The resulting rubric was reliably applied to participants' essays about trauma. Global ratings of essay organization were correlated with improvements in physical and mental health measures at a

Bridget Klest, MA, and Jennifer J. Freyd, PhD, are affiliated with Department of Psychology, University of Oregon.

Address correspondence to: Bridget Klest, MA, Department of Psychology, 1227 University of Oregon, Eugene, OR 97403-1227 (Email bklest@uoregon.edu).

This research was supported by the Northwest Health Foundation, Grant Number 2001-255 *Child Abuse and Health: An Intervention* (Freyd, PI). The manuscript preparation was also supported in part by the Trauma and Oppression Research Fund at the University of Oregon Foundation. The authors are grateful to the numerous contributions to this research made by Ann Yee, James Pennebaker, and members of the Freyd Dynamics Lab at the University of Oregon.

Journal of Psychological Trauma, Vol. 6(1) 2007
Available online at http://jpsyt.haworthpress.com

doi:10.1300/J513v06n01_01

six-month follow-up. Properties of the rubric and correlations with outcome measures are discussed. doi:10.1300/J513v06n01_01

KEYWORDS. Writing, trauma, health, global ratings, statistical reliability

A growing body of research suggests that writing about emotional experiences may provide benefits to physical health and general well-being (Pennebaker, 1997). In a seminal study on this topic, Pennebaker, Kiecolt-Glaser, and Glaser (1988) found that healthy undergraduate students who wrote about distressing experiences in a controlled study showed a drop in visits to the student health center and an increase in cellular immune response following writing, compared with students who wrote about a trivial topic.

A decade later, researchers found that written emotional disclosure could increase lymphocyte levels, marking improved immune function (Petrie, Booth, & Pennebaker, 1998). Another study revealed that such disclosure resulted in better immune response to a hepatitis B vaccination compared with a control group (Petrie, Booth, Pennebaker, Davis, & Thomas, 1995). Writing about emotional experiences appears to bolster the immune system and benefit health.

Physical health is not the only aspect of well-being that might benefit from writing. A study of individuals who had recently lost their jobs found that participants who wrote about the emotions associated with job loss were more likely to find new jobs within the months following writing than were control participants (Spera, Buhrfeind, & Pennebaker, 1994). Writing about stressful events has also been shown to improve academic performance of college students (Lumley & Provenzano, 2003; Pennebaker & Francis, 1996).

Most studies on the health benefits of writing have focused on health improvements in already healthy research participants. Writing about stressful events has led to disease-specific improvements in symptoms in some patient populations, such as research participants with mild to moderately severe asthma or rheumatoid arthritis (Smyth, Stone, Hurewitz, & Kaell, 1999). However, some researchers has found limited or no beneficial effects of written disclosure, particularly in patient populations, or in populations defined by exposure to psychological stress (Harris, 2006). Thus although writing appears to be beneficial across settings for healthy participants, it is yet unclear what

makes writing beneficial in some patient samples and not in others. It seems that writing has the potential to be a major public-health intervention for reducing symptoms and healthcare use among healthy participants. This may be true among physically and psychologically distressed populations as well if research can determine what makes writing effective in these populations.

Trauma is known to be related to a variety of physical and mental health symptoms (Felitti et al., 1998). It would seem that writing about traumatic experiences might be a useful intervention for reducing such symptoms, but empirical results have been mixed. While some researchers have found health benefits of writing about trauma among frequent healthcare users (Gidron et al., 2002), other researchers (Batten, Follette, Hall, & Palm, 2002; Freyd, Klest, & Allard, 2005) have found no overall benefit of writing about traumatic experiences. Studies that have focused specifically on the impact of writing about trauma in people who report experiencing significant trauma such as sexual assault, or other betrayal or life-threat traumas (e.g., Batten et al., 2002, Freyd et al., 2005, respectively) have found no benefit of writing about trauma. This raises questions about the potential efficacy of written disclosure as a beneficial intervention for trauma.

The answer may lie in the observation that not all writing is equal in predicting health benefits. Pennebaker and Francis (1996) developed a computer program called the Linguistic Inquiry and Word Count (LIWC) to analyze the content of participants' essays and assess what aspects of writing might be related to improvements in health. The LIWC counts the number of words used in participants' essays that fall into specific categories, as defined by groups of related words in the program's dictionary (Pennebaker & Francis, 1996). Studies using this program, to analyze essay content, have revealed that heavier use of insight-related and causal words predict improvement in health, as do heavier use of positive emotion words, and using a moderate number (contrasted with very high or very low numbers) of negative emotion words (Pennebaker, 1997).

The results of the LIWC's computer text analysis are notable, but may not provide a complete picture of what makes essay-writing beneficial to research participants. For example, the use of causal and insight words has been assumed to reflect a tendency toward constructing a coherent narrative (Pennebaker & Francis, 1996), but coherence is one aspect of written essays that the LIWC cannot directly assess, and this assumption has as yet not been supported with research. In fact, in the one study attempting to assess whether causal and insight words were related to construction of a coherent narrative, no reliable correlation between the two was found (Graybeal, Sexton, & Pennebaker, 2002).

THE ROLE OF COHERENCE

Exposure therapy is considered to be among the most effective treatments for posttraumatic stress disorder (Riggs, Cahill, & Foa, 2006). A key element of exposure therapy is the development, through repeated imaginal exposure or retelling, of a coherent and integrated narrative about the targeted traumatic event (Riggs et al., 2006, Shipherd, Street, & Resick, 2006). It is suggested that creating a coherent story, through imaginal exposure, from previously fragmented emotions and memories might alleviate symptoms of avoidance, hyperarousal, and dissociation that are characteristic of PTSD (Riggs et al., 2006, Shipherd et al., 2006). Freyd (1996) has similarly hypothesized that transforming memories that are sensory in nature into a more sharable form (e.g., the language of a coherent and sharable narrative) might alleviate symptoms such as flashbacks, and "re-living" the event, while making the memories more consciously accessible. Although there is debate in the literature regarding the nature of traumatic memories (Sivers, Schooler, & Freyd, 2002), it is relatively well-documented that exposure therapy, with a focus on developing a coherent trauma narrative, is effective in treating trauma symptoms (Shipherd et al., 2006). However, the global coherence of written narratives has not often been assessed in writing studies, and as far as we know, has never been assessed in essays about trauma.

To date, we know of only one study that has assessed global ratings of essay characteristics in a writing intervention study such as those described above. In that study, Graybeal and colleagues (2002) had raters judge undergraduate participants' essays using 7-point scales, answering questions such as "to what degree does this essay tell a story?," and "to what degree does this essay have a moral or a message?" While this study was pioneering in its attempt to assign global ratings to essays, the particular rating system used had several problems associated with it. First, the authors fail to mention how the rating system was developed. The questions used in the rating system may have face validity, however it is impossible to determine whether they were measuring "story-making," as was asserted by the authors. The inter-rater reliabilities for essay ratings were relatively low, with alphas reported as "above .60" which is conventionally the absolute minimum to be considered reliable. These reliability statistics were calculated for four raters, and time and resource considerations generally dictate that fewer raters is better; with fewer raters reliability tends to drop. It is unclear whether raters were given any training on this rating system prior to applying it to the studied essays, which may account for some inconsistency. The assertion made by the authors that this study challenges the idea that a well-constructed story is beneficial to physical and mental health is weak at best due to these and other limitations of the study, some of which are mentioned by the authors

(Graybeal et al., 2002). Thus a better rating system is needed in order to assess to what degree characteristics of an essay as a whole might contribute to the health improvements noted in previous writing studies.

A handful of researchers have used rating systems to code the coherence and cohesion of children's' fictional narratives (e.g., Cain, 2003; Shapiro & Hudson, 1991). Such rating schemes have generally coded coherence by noting the presence or absence of traditional story components (e.g., beginning, middle, and ending) and cohesion by noting the use of appropriate connectives. Additionally, overall coherence has been rated by matching structural elements of children's stories to criteria mapped out in scoring rubrics (Cain, 2003; Shapiro & Hudson, 1991). Although the particular criteria used in these studies were geared toward rating children's fictional stories, they provide a useful template for use with trauma essays written by adults.

Scoring rubrics, defined generally, are guides for assigning scores representing the overall quality of something, placing it into an ordinal category. Rubrics differ from other kinds of assessment tools in that they provide descriptions of the characteristics of each scoring level individually. For example, each score has a set of descriptive qualities. An evaluator might want to assess whether the ideas in a paper are connected in such a way that there is a good flow from one idea to the next, and the ideas come together to form a coherent story. This is precisely the type of subjective (and potentially important) quality that a computer program, such as the LIWC, cannot currently evaluate. And, unarmed with objective scoring criteria, two raters might have very different subjective impressions of the same essay. However, by using a rubric which lays out criteria for evaluating this subjective factor, it is more likely that raters will agree on the meaning of the factor and in their ratings.

The use of rubrics is not a novel concept, and in fact, rubrics are not unlike other assessment tools in psychology. For example, psychological disorders are often diagnosed by matching symptoms against the criteria laid out in diagnostic categories. However, the use of rubrics to assess writing in psychological research has not, as far as we know, been reported before.

Applying rubrics to research participants' essays about trauma may prove extremely useful. In particular, if developing a coherent and sharable narrative is key in the treatment of PTSD, as is suggested by research on exposure therapy and by shareability theory, scoring narrative coherence (the degree to which the essay has an overall plan or structure) and cohesion (how logically and easily the essay makes transitions between sentences, topics, and ideas) might provide clues about written disclosure as a trauma intervention.

STUDY AIMS AND HYPOTHESES

The first aim of the current study was to develop a reliable, valid, and easily implemented global rating system for essays about trauma, based on scoring rubrics. Additionally, the rating system was compared with the LIWC. The LIWC is a very fast and easy system for scoring essays. And if LIWC scores are highly correlated with global ratings, then it is possible that using a global rating system will not add to our prediction of health outcomes. The second aim was to determine whether global qualities of participants' essays were associated with improvements in health.

It was hypothesized that a valid global rating system could be developed, and reliably and economically applied to essays about trauma. It was also hypothesized that ratings for organization would be correlated with health outcomes such that better essay scores would predict long-term improvements in physical and mental health symptoms.

METHOD

The data for this study were collected as part of a larger study investigating the relationship between betrayal trauma and physical and mental health for adults with chronic pain and chronic health problems. The study was designed as an intervention, investigating whether writing about personal experiences of betrayal trauma (interpersonal trauma perpetrated by a close other) and/or completing a survey about such trauma might have health benefits for this population, similar to the investigation by Gidron and colleagues (2002) of frequent health-care clinic users.

Participants completed a battery of questionnaires, including the Brief Betrayal Trauma Survey (BBTS, Goldberg & Freyd, 2006) during a pre-test session, prior to completing three weekly 20-minute writing sessions. The BBTS assesses a variety of traumatic experiences including life-threat traumas such as accidents and natural disasters, as well as interpersonal traumas such as sexual assault by a stranger, and betrayal traumas such as sexual or emotional abuse by a close other. All participants in this sample endorsed having experienced at least one traumatic event on the BBTS. Rates of lifetime exposure to a DSM PTSD criterion A trauma have been estimated at around 70% in the general population (e.g., Resnick, Kilpatrick, Dansky, Saunders, & Best, 1993). The BBTS captures serious interpersonal traumas that might not be readily included as criterion A stressors (such as emotional abuse), and given this and the strong relationship between chronic health problems and exposure to

trauma (e.g., Felitti et al., 1998), it is not particularly surprising that all of our participants endorsed experiencing at least one traumatic event.

Participants wrote about the most disturbing or distressing event they had experienced that involved at least one other person, instructions which probed specifically for betrayal traumas. Six months after the third writing session, participants were asked to complete the same battery of questionnaires they had been given at the beginning of the study. A complete description of data collection procedures for this study can be found in Freyd and colleagues (2005).

PARTICIPANTS

Participants for this study were 40 community adults, 25 women and 15 men, ranging in age from 19 to 63 years ($M = 40.23$, $SD = 12.25$). Participants were selected from a larger sample (Freyd et al., 2005), and were recruited based on having experienced chronic pain and/or other chronic health problems. Although the sample was not directly recruited based on having experienced trauma, we expected many participants would have trauma histories (based on the strong correlation between chronic health problems and trauma) and this was indeed the case in our sample. The sample was demographically diverse, with a large range in educational attainment (8 years [8th grade] to 20 years [PhD], $Mdn = 13$, $M = 13.94$, $SD = 2.63$) and annual income ($0 to $32,000, $M = $10,830, $SD = $8,783), and ethnic/racial diversity roughly representative of the community from which participants were drawn, with ethnic minorities being slightly overrepresented (27 White, 6 Native American, 3 Hispanic, 1 Black, 3 no response).

MATERIALS

The symptom measures used in this study were time-bound such that participants were instructed to report how frequently they had experienced each symptom during the past month. In contrast, the original measures asked participants to report on symptom frequency experienced over longer periods of time. This was done for comparison purposes between scores obtained prior to the intervention and those obtained following the intervention, approximately 6 months later. Reliability and validity statistics presented in the descriptions below are for administrations of the measures without time-bound instruction; no statistics are available for time-bound administrations.

Pennebaker Inventory of Limbic Languidness, time bound (PILL-t; Pennebaker, 1982). The PILL-t assesses the degree to which participants have experienced each of 54 physical health symptoms (e.g., headaches, chest pains, abdominal pain) during the last month, on a 5-point likert scale ranging from 0 (never) to 4 (almost every day). The PILL-t also asks participants how many days they have been sick, how many days activity has been restricted due to illness, and how many visits to a doctor they have made in the last month. The recommended way to score the symptom part of the PILL-t is to sum up the total number of items on which individuals score 3 or higher (indicating about once a week or more frequent), resulting in a score ranging from 0 to 54. Using this scoring method, a mean score of 17.9 (SD = 4.5) was obtained on a sample of 939 college students in the original study assessing the PILL's psychometric value, which was found to be high in terms of reliability and validity (Pennebaker, 1982).

Trauma Symptom Checklist 40, time bound (TSC40-t; Briere & Runtz, 1989). The TSC-40-t is a 40-item checklist, assessing symptoms commonly associated with the experience of traumatic events. Respondents are asked to indicate how frequently they experienced each symptom on a scale of 0 (never) to 3 (very often). The TSC-40 is composed of 6 symptom subscales: anxiety, depression, dissociation, sexual abuse trauma index, sexual problems, and sleep disturbances. Sample items include "anxiety attacks" and "trouble getting along with others." The TSC-40-t is scored by summing responses, for a resulting score falling between 0 and 120, with higher scores indicating greater trauma symptomatology. The average TSC-40 score in a study of 438 female students was 66.8, and for those who had experienced child and/or adult abuse, the mean ranged from 70.4 to 77.4 (Gold, Milan, Mayall, & Johnson, 1994). The measure has been shown to have good reliability and validity (Briere & Runtz, 1989; Elliott & Briere, 1992).

Dissociative Experiences Scale, time bound (DES-t; Carlson & Putnam, 1993). This 28 item questionnaire assesses the frequency with which participants have had particular dissociative experiences during the past month. Respondents select a percentage ranging from 0 to 100, increasing in 10% intervals, to indicate how frequently each item is experienced. Items range from normal dissociative experiences such as "spacing out" during a conversation with someone to more unusual experiences such as not recognizing oneself in the mirror. The overall DES score is obtained by averaging the 28 item scores, yielding a score ranging from 0 to 100. Scores above 20 suggest the presence of highly dissociative experiences and that further clinical assessment is warranted, whereas scores below 10 fall within the normal range of dissociative experiences (Carlson & Rosser-Hogan, 1993). The DES has been to shown to have very good validity and reliability, and good overall

psychometric properties in a number of studies (see Briere, 1997 for a review). A relationship between the development of dissociative symptoms and traumatic experiences has been documented (e.g., Bernstein & Putnam, 1986, Chu & Dill, 1990).

PROCEDURE

Participants completed four sessions over the course of six months. The first session involved completing a questionnaire packet and a 20-minute writing session. The second and third sessions happened one and two weeks following the first session, and involved writing only. The fourth session occurred six months later, and involved completing the same questionnaire packet used in the first session.

Code Development. We developed a code to assess global aspects of participants' essays, in an attempt to determine whether particular characteristics of these essays influence the effectiveness of the writing intervention. The Global Ratings of Essays About Trauma (GREAT) code was modeled after rubrics used to assess the writing skills of students in second through twelfth grades. Information on these rubrics and the rubrics themselves were obtained from school websites and state education department websites in Oregon, California, Alaska, and Illinois (IGAP, 1993; Language arts, 1997; Official scoring guide, 2002; Scoring guides, 2000).

Each of these rubrics contains a number of scoring dimensions, from language conventions and paragraphing to ideas and content. For this study, an analytic rubric with scoring guides for organization was created using some criteria from these educational rubrics and some criteria developed by the authors. Each dimension was scored on a 5-point scale, where a score of 1 indicated that the essay was generally uncodable, and a score of 5 indicated excellent demonstration of the trait being scored. Scores of 2, 3, and 4 were assigned to essays falling between these two extremes. Each score was associated with a set of descriptive scoring criteria to assist in making objective ratings. For example, in coding the coherence of an essay the raters are given criteria related to the structure of the essay for each possible score. A score of 3 requires that the writer frequently includes off-topic digressions, a 4 indicates few digressions, and a 5 is given only when there are no off-topic digressions. All scales were ordinal, with higher scores indicating better essays. The rating criteria are attached in the Appendix, and complete coding instructions are available from the authors upon request.

The criteria used to score organization were drawn from several educational rubrics, and edited and combined to be relevant to a variety of narrative

essays written by adults. Most educational rubrics are based on the assumption that the rubric will be used to score students at the same educational level, all writing about the same assigned topic. This was not the case in the current study; therefore, several educational rubrics were pieced together, taking only the most general parts of each one so as to apply equally to writers of varying abilities and varying topics. The rubrics used as models were originally used to score the writing of students in second through twelfth grades. The goal in using such a broad educational range was to create a code that would not correlate with level of educational attainment, a possible confound with essay quality.

Organization was coded using sub-rubrics for coherence and cohesion. The coherence score assessed the degree to which an essay had an overall plan or structure, including a related beginning, middle, and conclusion. The cohesion score assessed the degree to which sentences, paragraphs, and ideas transitioned easily and progressively. These two sub-rubrics were combined to create an overall organization score.

Prior to coding the essays for analysis, a subset of essays was used as practice to refine the scoring rubrics and establish interrater reliability. A total of 120 essays were collected in this study (from 40 participants who each wrote on three occasions). Most of these were used for training purposes and all were coded for final analysis. No essay was coded by the same coder in both the training phase and the final coding phase. In the first phase of practice coding, essays written by 20 participants (a total of 60 individual essays) were rated by two coders. The first 33 essays coded revealed ambiguous wording in the rubrics, which were changed to leave less room for interpretation. The remaining 27 essays in this subset were rated by two coders as practice essays and to check interrater reliability. A second subset of 42 essays was coded by two different raters, 30 for training, and 12 to check reliability. Our goal was to have reliability coefficient alphas at or above .75 (using intraclass correlation) prior to coding essays for final data analysis. Overall, reliability statistics for practice essays after training on the code were between .73 and .89.

Coding. Essays were coded in two phases, using two coding pairs. Each coding pair rated the essays that had been used to train the other coding pair, and did not rate any essay they had previously coded while training on the code. In the first phase, two coders rated the essays of 20 participants (60 essays total). Both coders rated each essay. Coding of these final essays took place over the course of three weeks. Coders rated six to seven sets of essays per week, three to four sets in one sitting. These two coders were the first author and a research assistant involved in developing and revising the code. In order to determine the ease with which untrained research assistants could use the code, two new coders who were blind to all hypotheses were recruited. In

the second phase, this second team of coders rated a different set of 60 essays after a period of training.

LIWC Analysis. Essays were analyzed using the Linguistic Inquiry and Word Count program (LIWC, Pennebaker & Francis, 1996). Dr. James Pennebaker volunteered his services for this portion of the analysis. A number of dimensions of each essay were scored using this program. For the purposes of this study, only causal words and insight words will be discussed. Causal and insight words have been predicted to relate to coherent narrative development (Graybeal et al., 2002).

DATA ANALYSIS AND RESULTS

Interrater reliabilities were checked using intraclass correlation. Intraclass correlation was used primarily because previous research in this area by Graybeal and colleagues (2002) reported coefficient alphas for interrater reliability, and using the same reliability procedure facilitated comparisons between these two studies. Also, intraclass correlation is acceptable for use with both continuous and ordinal data, which makes it particularly useful in cases of continuous data with a somewhat restricted range (Streiner, 1995). Interrater reliabilities were first computed for each coding category using all essays individually-three essays from each participant. However, since essays by the same participant are by definition non-independent observations, essays from each of the three writing sessions were evaluated for reliability separately. Next, for simplicity of later data analysis, a single score in each category was computed for each participant by calculating the average score for that participant for all three essays. Reliabilities for these composite scores were also calculated using intraclass correlation. See Table 1 for a summary of these results.

Symptom change scores were calculated by subtracting scores at the final session from scores at the first session on the PILL-t, DES-t, and TSC40-t so that difference scores greater than 0 indicate an improvement in symptoms, and negative difference scores indicate symptom increases. Descriptive statistics were calculated for all variables used in subsequent data analysis (see Table 2). Correlations between demographic variables and outcome variables were assessed to determine what, if any, variables should be included as covariates. Age, income, and ethnicity were not correlated with any predictors or outcome variables ($rs < .20$, $ps > .20$). Gender was correlated with change in DES scores ($r = .31$, $p < .05$) such that men tended to show a slight increase and women a slight decrease in dissociation over time, and educational attainment and gender were controlled for in all regression analyses.

TABLE 1. Interrater Reliability Alpha Coefficients for GREAT Scores from the Average of Two Raters, with Reliability Estimates for a Single Rater in Parentheses ($p < .001$)

	Session 1 Essays	Session 2 Essays	Session 3 Essays	Average Scores for 3 Essays
Organization	.85 (.74)	.89 (.80)	.91 (.83)	.90 (.81)

TABLE 2. Description Statistics

		Minimum	Maximum	Mean	Std. Deviation
Organization		2.50	4.17	3.4917	.41164
TSC40-t					
	change	-65.00	31.00	-1.4359	17.33994
	pretest	23	134	76.4872	18.79232
	posttest	46.00	147.00	77.7500	19.92132
PILL-t					
	change	-14.00	12.00	-.7000	6.15276
	pretest	5.00	40.00	20.4250	9.62605
	posttest	3.00	43.00	21.1250	10.83723
DES-t					
	change	-40.00	31.07	-.4821	9.99985
	pretest	.71	39.29	9.6161	8.66757
	posttest	.71	46.07	10.0982	10.76180
LIWC Causal Words		.39	2.06	1.1239	.37215
LIWC Insight Words		1.21	4.02	2.6353	.74685

Previous research has found that increased use of causal and insight words across writing sessions is predictive of positive outcome (Pennebaker & Francis, 1996). Change scores were calculated for LIWC causal and insight words, as well as for organization, by subtracting first session scores from final session scores. Positive change scores indicate increases, and negative change scores indicate decreases.

To determine whether computer-calculated essay fluency was a potential confounding factor, each essay was scored using the Flesch ease-of-reading

scale (available in most word-processing programs). Flesch scores are calculated using sentence length, word length, and paragraph length, and reflect readability and "grade-level" of a piece of writing. Although there has been debate about the validity of Flesch scores and other readability scores (Oakland & Lane, 2004), such scores are widely used and we wanted to rule out the possiblity that surface-level factors were influencing organization ratings. In our data set, Flesch readability scores were not correlated with rubric-scored organization ($r = .20$, $p > .20$), or with outcome measures ($rs < .20$, $ps > .20$). Thus fluency was assumed not to be a confounding factor and was left out of the rest of our analyses.

Simultaneous entry regression analyses were run to determine what proportion of variance in outcome scores was attributable to coded dimensions of essays, compared with LIWC word counts and other possible sources of variance such as demographic variables. In these analyses, average coded organization ratings, LIWC causal words, and LIWC insight words were entered as predictors, along with gender and educational attainment as covariates. Three analyses were run with this set of predictors, one for each of the three outcome varibles which were change in PILL-t scores, change in TSC40-t scores, and change in DES-t scores.

A second set of regressions was perfomed that was nearly identical to the first, but used change in organization scores and LIWC scores over time (instead of average scores across time) as predictors. Again, gender and educational attainment were entered as covariates, and separate regression analyses were run with each of the three symptom change scores at outcome variables.

These regression analyses also allowed us to determine the degree to which variance in symptom reduction attributable to coded organization overlapped with variance attributable to LIWC causal and insight words. By comparing the squared semi-partial correlation coefficients between each of these predictors and the outcome variables, we were able to determine whether each predictor contributed to outcome variance over and above the contribution of other predictors.

In the first set of analyses, using average scores on essay ratings as predictors, in each case only organization was a significant predictor of outcomes such than higher rated organization predicted greater symptom improvement (see Table 3). In the second set of regressions using change scores on essay ratings as predictors, there were no significant or marginally significant effects (see Table 3).

TABLE 3. Relationships Between Symptom Improvement and GREAT Coding and LIWC ($N = 40$, $df = 6, 32$)

	Average across time					Change over time				
			Squared Semi-partial rs					*Squared Semi-partial rs*		
Dependent Measure	*Adjusted* R^2	*F*	Organization	LIWC Causal	LIWC Insight	*Adjusted* R^2	*F*	Organization	LIWC Causal	LIWC Insight
PILL-t	.16	2.77*	**.17***	.01	.00	.09	.37	.00	.01	.02
TSC40-t	.22	2.45*	**.25****	.01	.05	.03	1.12	.06	.04	.01
DES-t	.14	1.90†	**.08†**	.06	.00	.03	1.25	.03	.00	.02

†$p < .10$, *$p < .05$, *$p < .01$

DISCUSSION

Several of this study's hypotheses were supported. First, we found that the GREAT code could be reliably applied to essays about trauma. In fact, reliability levels were uniformly high, with alphas between .84 and .93 for the coding used in data analysis. Coding by two raters was sufficient to achieve these high levels of reliability, and coding by a single rater would yield slightly lower, but still acceptable, reliability alphas between .73 and .87. This suggests that the GREAT code can be reliably and economically applied to research participants' essays about trauma.

The results of this study also establish preliminary predictive validity of the GREAT code. Organization scores were significantly predictive of decreases in physical and mental health symptoms, and marginally associated with decreases in dissociation. Additionally, causal and insight words as measured by the LIWC were not predictive of outcomes, and did not overlap significantly with organization scores. Taken together, these results suggest that the GREAT code, particularly organization, measures a quality of essays that has predictive power and that is not captured by other coding systems currently used by researchers in this area. The GREAT code is a potentially useful research tool for deepening understanding about the mechanisms underlying the health benefits of expressive writing.

The finding that better essay organization is related to symptom reduction at a six-month follow-up suggests that narrative coherence may indeed play an important role in the relationship between health and writing. The health benefits of expressive writing may, to some degree, depend on narrative coherence. This has been hypothesized but never directly tested with a reliable, valid coding scheme prior to the GREAT code and the current study.

These results have implications for interventions and treatments for people with physical and mental health symptoms. If writing a coherent narrative is beneficial to health, perhaps providing instruction on how to do this could help people with previously less coherent narratives gain similar benefits. Con-

versely, it could also be true that people with better health trajectories are able to write more coherent narratives. Perhaps a writing course focused on developing a coherent trauma narrative could be an effective intervention for trauma survivors, and research on such an intervention could help determine whether the relationship observed in this study is a causal one.

One question raised by the current study is whether the health benefits of expressive writing seen in other studies are actually due to writing about emotional topics, or whether they are perhaps solely due to formation of coherent essays. It is possible that asking participants to write about the most traumatic event they have ever experienced prompts more organized, coherent narratives than instructions to write about time management. If this is the case, the health benefits seen in previous studies might be accounted for simply by differences in essay organization. Intuitively, and because of the large body of research finding benefits to writing about emotional topics, it seems that emotional expression is integral to receiving health benefits from writing. However, in light of the results of the current study, it seems important for future research to address this question, pitting emotional expression against organization, and assessing the possibility that the combination of these two factors is necessary for writing to benefit health.

Also puzzling is the lack of any relationship between LIWC causal and insight words and outcomes. Pennebaker and colleagues have found such a relationship in many previous studies, although not in all (Graybeal et al., 2002). It is possible that in some samples, these dimensions do not predict outcome. In addition, our written disclosure instructions were trauma focused as opposed to focused on less severe emotional experiences. Other researchers (Batten et al., 2002) have suggested that writing about significant trauma may differ from writing about other types of experiences.

There are several limitations to the current study which should be addressed in future research. First, essays by only 40 participants were included in the data analysis for this study. With such a small sample size, the power of the analyses was relatively low. Although it is remarkable that significant relationships were uncovered even with this low power, important relationships may have been missed. It is possible that moderate correlations between variables were non-significant only because of the small sample size in this study, and that other relationships between variables in this study do in fact exist but went undetected. In addition, although no essay was rated by the same person for both training and analysis, the same set of essays was used for both code development and final coding. It is possible that high reliability coefficients resulted at least partially from this overlap. New and larger sets of writing samples are needed to assess the reliability of the findings in this study.

A second limitation is that the participants in this sample were not representative of the general population of trauma survivors. Our sample was generally quite low-income, including several homeless and many unemployed participants. Additionally all participants in our sample had chronic health problems. Thus it is possible that the findings in this paper might not generalize to other populations, or that particular expected relationships (such as the relationship between proportion of causal and insight words and outcomes) were not observed. It will be important to test our findings with more representative samples in the future.

Future research using the GREAT code should assess its reliability and validity in not only more representative samples, but also more diverse essay types. All the participants in this sample wrote essays related to trauma, and although the GREAT code was primarily developed to assess the coherence of trauma essays, it may be useful for other types of essays as well. Scoring control essays in which participants write emotionally neutral but still potentially coherent stories might help parse out the relative contributions of emotional expression and coherence in the health benefits of narrative writing.

Although more research is needed, the GREAT code appears to be a reliable and valid research tool, and narrative coherence appears to be related to symptom improvements. This research is potentially important to the area of trauma and health, and more specifically, to exploration of the health benefits of narrative. Writing is a particularly exciting intervention because it is cost effective and potentially accessible to vast numbers of people. The GREAT code may complement current research tools for assessing narrative writing and has the potential to help determine what it is that makes writing so beneficial for some and not others. With an answer to that question, writing might be adapted to become a healing intervention for the masses.

REFERENCES

Batten, S. V., Follette, V. M., Hall, M. L. R., & Palm, K. M. (2002). Physical and psychological effects of written disclosure among sexual abuse survivors. *Behavior Therapy,* 33, 107-122.

Bernstein, E., & Putnam, F. (1986). Development, reliability, and validity of a dissociation scale. *Journal of Nervous and Mental Disease, 174*, 727-735.

Briere, J. (1997). Psychological assessment of Adult Posttraumatic States. Washington, DC: American Psychological Association.

Briere, J., & Runtz, M. (1989). The trauma symptom checklist (TSC-33): Early data on a new scale. *Journal of Interpersonal Violence, 4,* 151-163.

Cain, K. (2003). Text comprehension and its relation to coherence and cohesion in children's fictional narratives. *British Journal of Developmental Psychology, 21,* 335-351.

Carlson, E. B., & Putnam, F. W. (1993). An update on the dissociative experiences scale. *Dissociation, 6,* 16-27.

Carlson, E. B., & Rosser-Hogan, R. (1993). Mental health status of Cambodian refugees ten years after leaving their homes. *American Journal of Orthopsychiatric, 63,* 223-231.

Chu, J. A., & Dill, D. L. (1990). Dissociative symptoms in relation to childhood physical and sexual abuse. *American Journal of Psychiatry, 147,* 887-892.

Elliott, D. M., & Briere, J. (1992). Sexual abuse trauma among professional women: Validating the trauma symptom checklist-40 (TSC-40). *Child Abuse and Neglect, 16,* 391-398.

Felitti, V. J., Anda, R. F., Nordenberg, D., Williamson, D. F., Spitz, A. M., Edwards, V., Koss, M. P., & Marks, J. S. (1998). Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults. *American Journal of Preventive Medicine, 14(4),* 245-258.

Freyd, J. J. (1996). *Betrayal Trauma: The Logic of Forgetting Childhood Abuse.* Cambridge, MA: Harvard University Press.

Freyd, J. J., Klest, B., & Allard, C. B. (2005). Betrayal trauma: Relationship to physical health, psychological distress, and a written disclosure intervention. *Journal of Trauma & Dissociation, 6(3),* 83-104.

Gidron, Y., Duncan, E., Lazar, A., Bidernan, A., Tandeter, H., & Shvartzman, P. (2002). Effects of guided written disclosure of stressful experiences on clinic visits and symptoms in frequent clinic attenders. *Family Practice, 19,* 161-166.

Gold, S. R., Milan, L. D., Mayall, A., & Johnson, A. E. (1994). A cross-validation study of the trauma symptom checklist: The role of mediating variables. *Journal of Interpersonal Violence, 9,* 12-26.

Goldberg, L. R., & Freyd, J. J. (2006). Self-reports of potentially traumatic experiences in an adult community sample: Gender differences and test-retest stabilities of the items in a brief betrayal-trauma survey. *Journal of Trauma & Dissociation, 7(3),* 39-63.

Graybeal, A., Sexton, J. D., & Pennebaker, J. W. (2002). The role of story-making in disclosure writing: The psychometrics of narrative. *Psychology and Health, 17,* 571-581.

Harris, A. H. S. (2006). Does expressive writing reduce health care utilization? A meta-analysis of randomized trials. *Journal of Consulting and Clinical Psychology, 74,* 243-252.

IGAP narrative scoring guide. (1993). Retrieved November 17, 2003, from http://www.gower.k12.il.us/Staff/WRITEON/32 narr.htm - Nar%20Guide

Language arts six-trait analytic scoring guide. (1997). Fairbanks North Star Borough School District. Retrieved November 17, 2003, from http://www.northstar.k12.ak.us/curriculum/currguide/langarts/sixtrait.html

Lumley, M. A., & Prvenzano, K. M. (2003). Stress management through written emotional disclosure improves academic performance among college students with physical symptoms. *Journal of Educational Psychology, 95,* 641-649.

Oakland, T., & Lane, H. B. (2004). Language, reading, and readability formulas: Implications for developing and adapting tests. *International Journal of Testing, 4,* 239-252.

Official scoring guide: Writing. (2002). Oregon Department of Education. Retrieved November, 17, 2003, from http://www.ode.state.or.us/asmt/scoring/guides/2003-04/writingscoringguide0304.pdf

Pennebaker, J. W. (1982). *The Psychology of Physical Symptoms.* New York: Springer-Verlag.

Pennebaker, J. W. (1997). Writing about emotional experiences as a therapeutic process. *Psychological Science, 8(3),* 162-166.

Pennebaker, J. W., & Francis, M. E. (1996). Cognitive, emotional and language processes in disclosure. *Cognition and Emotion, 10,* 601-626.

Pennebaker, J. Wl, Kiecolt-Glaser, J. K., & Glaser, R. (1988). Disclosure of traumas and immune function: Health implications for psychotherapy. *Journal of Consulting and Clinical Psychology,* 56, 239-245.

Petrie, K., Booth, R. J., & Pennebaker, J. W. (1988). The immunological effects of thought suppression. *Journal of Personality and Social Psychology, 75,* 1264-1272.

Petrie, K., Booth, R. J., & Pennebaker, J. W., Davison, K. P., & Thomas, M. G. (1995). Disclosure of trauma and immune response to a hepatitis B vaccination program. *Journal of Consulting and Clinical Psychology, 63,* 787-792.

Resnick, H. S., Kilpatrick, D. G., Dansky, B. S., Saunders, B. E., & Best, C. L. (1993). Prevalence of civilian trauma and posttraumatic stress disorder in a representative samle of women. *Journal of Consulting and Clinical Psychology, 61,* 984-991.

Riggs, D. S., Cahill, S. P., & Foa, E. B. (2006). Prolongued exposure treatment of posttraumatic stress disorder. In Follete, Victoria M. & Ruzek, Josef I (Eds). *Cognitive-behavioral terhapies for trauma (2nd ed. Pp. 96-116).* New York, NY: Guilford Press.

Scoring guides for writing. (2000). Cajon Valley Union Schools. Retrieved November 17, 2003, from http://www.cajon.k12.ca.us/standards/scoring guides-pdf/index.htm.

Shapiro, L. R., & Hudson, J. A. (1991). Tell me a make-belief story: Coherence and cohesion in young children's picture-elicited narratives. *Development Psychology, 27,* 960-979.

Shipherd, J. C., Street, A. E., & Resick, P. A. (2006). Cognitive therapy for posttraumatic stress disorder. In Follette, Victoria & Ruzek, Josef I (Eds). *Cognitive-behavioral therapies of trauma (2nd ed.,* pp. 96-116). New York, NY: Guildford Press.

Sivers, H., Schooler, J., & Freyd, J. J. (2002). Recovered memories. In V. S. Ramachandran (Ed.). *Encyclopedia of Human Brain, Volume 4,* (pp. 169-184). San Diego, California and London: Academic Press.

Smyth, J. M., Stone, A. A., Hurewitz, A., & Kaell, A. (1999). Effects of writing about stressful experiences on symptom reduction in patient with asthma or rheumatoid arthritis. *JAMA, 281,* 1304-1309.

Spera, S. P., Buhrfeind, E. D., & Pennebaker, J. W. (1994). Expressive writing and coping with job loss. *Academy of Management Journal, 37,* 722-733.

Streiner, D.L. (1995). Learning how to differ: Agreement and reliability statistics in psychiatry. *Canadian Journal of Psychiatry, (40)*, 60-66.

doi:10.1300/J513v06n01_01

RECEIVED: 09/28/06
REVISED: 01/24/07
ACCEPTED: 01/25/07

APPENDIX
GREAT Coding Rubric

Organization

Coherence: How good is the overall plan or structure of the essay? Does the story progress logically, with a beginning, middle, and conclusion? If the reader is able to determine a beginning, middle, and end to the story that is the main focus of the essay, the essay is coded a 3 or higher. If not, it is a 2 or lower.

1	2	3	4	5
Not enough was written to code this essay, or the essay is not understandable to the reader.	Possible evidence of attempted structure, but structure is hard to infer. The story focuses on more than one event, none of which have enough detail to give the story a clear focus, or there is not much detail provided about the focus event. Organization is rough, though it may not be completely absent.	Has basics of structure, including a roughly defined beginning, middle, and end. Has one main focus but also includes less important events/details that do not add to the reader's understanding, or, fails to provide important details that would add to the reader's understanding Frequently gets off topic.	Has good structure, including a beginning, middle and end in logical order. Tells about one specific event in detail with only minor digressions. Once or twice includes less important details that do not add to the reader's understanding.	Has good structure, including a beginning, middle and end in logical order. Tells about one specific event in detail. Does not make digressions.

Cohesion: How well does the essay transition sentence-to-sentence and topic-to-topic? Is the essay choppy or does it flow easily?

1	2	3	4	5
Not enough was written to code this essay, or the essay is not understandable to the reader.	Many sentences do not flow easily one to the next. Transitions are	Some sentences flow easily one to the next. At times transitions are	Many sentences flow easily one to the next. Most transitions are easy to	Sentences flow easily one to the next, with only one or two exceptions. .

APPENDIX (continued)

	usually hard to follow. The reader can only understand the progression of ideas by making inferences. Writing is generally choppy.	easy to follow, at times they are not. Ideas sometimes follow one another logically, and sometimes do not. Writing is not particularly choppy, but not particularly easy to read.	follow. The reader may, rarely, have to make inferences to understand why one idea follows another. Generally easy to read.	Transitions are easy to follow. The reader does not have to make inferences to understand the progression of ideas. Can be read quickly and effortlessly.