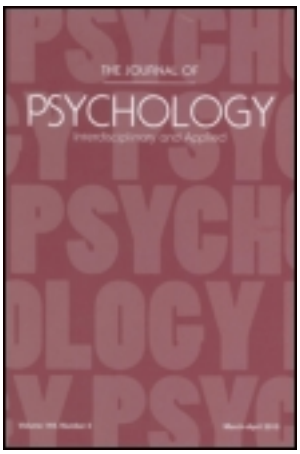

The Journal of Psychology: Interdisciplinary and Applied
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/vjrl20

Test-Retest Reliability of the Emotional Stroop Task: Examining the Paradox of Measurement Change

P. Eide [a], A. Kemp [a], R. B. Silberstein [a], P. J. Nathan [a] & C. Stough [b]

[a] Brain Sciences Institute Swinburne University of Technology

[b] Center for Neuropsychopharmacology Swinburne University of Technology

Version of record first published: 02 Apr 2010.

To cite this article: P. Eide , A. Kemp , R. B. Silberstein , P. J. Nathan & C. Stough (2002): Test-Retest Reliability of the Emotional Stroop Task: Examining the Paradox of Measurement Change, The Journal of Psychology: Interdisciplinary and Applied, 136:5, 514-520

To link to this article: http://dx.doi.org/10.1080/00223980209605547

The Journal of Psychology, 2002, *136*(5), 514–520

Test–Retest Reliability of the Emotional Stroop Task: Examining the Paradox of Measurement Change

P. EIDE
A. KEMP
R. B. SILBERSTEIN
P. J. NATHAN
Brain Sciences Institute
Swinburne University of Technology

C. STOUGH
Center for Neuropsychopharmacology
Swinburne University of Technology

ABSTRACT. The Emotional Stroop (ES) task (I. H. Gotlib & C. D. McCann, 1984) has been proposed as an experimental measure to assess the processing of emotion or the bias in attention of emotion-laden information. However, study results have not been consistent. To further examine its reliability for empirical research, the authors of this study administered the ES task to 33 participants on 2 separate occasions separated by 1 week. Results indicated that retest reliabilities for reaction times (RTs) derived from the 3 separate emotion conditions (manic, neutral, and depressive) across the 1 week interval were very high. However, consistent with previous research, the reliabilities were very low for the interference indices (manic and depressive). These low reliabilities reflect the very high intercorrelation between the RTs derived from the 3 conditions. The authors concluded that a better indicator of the reliability for this task is the individual RTs from each emotion condition.

Key words: emotion, emotional Stroop, reliability, test–retest

THE EMOTIONAL STROOP TASK (ES; Gotlib & McCann, 1984) is based on the original Stroop task (Stroop, 1935), which has previously been used to examine attentional processes and the well-known interference or *Stroop effect*. Results from studies employing the Stroop task have revealed that participants require a longer time to name the color of a stimulus when the word is incongruent than when it appears as a solid color square. That is, they have trouble saying "blue" when blue ink is used in printing the word "red."

Researchers are beginning to focus on the abnormal processing of emotion

or attention in emotion-related (e.g., affective) disorders (Austin et al., 1999; Franke, Maier, Hardt, & Frieboes, 1993; Hill & Knowles, 1991; Kinderman, 1994; Lemelin & Baruch, 1998; Tarbuck & Paykel, 1995). To assess the processing of emotional information in psychiatric disorders, such as depression, researchers may adapt traditional cognitive tests; the Stroop task has been adapted for research examining the processing of emotion-laden words (Gotlib & McCann, 1984; Williams & Nulty, 1986)

The ES task differs from the Stroop task in that emotional- and neutral-content words are presented instead of color-incongruent words and the interference effect is a result of emotional content rather than incongruence of color. Williams, Mathews, and MacLeod (1996) hypothesized that the ES task measures attentional bias because depressed individuals perform poorly at color naming when the words have a depressed content than when the words have a neutral or manic content. Response latencies from depressed individuals are longer when the stimulus material is negatively valenced than when the material is neutral or positively valenced (Kindt, Bierman & Brosschot, 1996; Segal, Gemar, Truchon, Guirguis, & Horowitz, 1995; Siegrist, 1997; Williams & Nulty, 1986).

It has been assumed that the emotional modification of the cognitive paradigm allows the measurement of inhibition of emotional information in much the same manner as the original cognitive paradigm measures inhibition of cognitive processes (Kindt et al., 1996). However, this has not been empirically evaluated. LeDoux (1990) suggested that processing emotional information is qualitatively different and that such information follows a different neural pathway than emotionally neutral (cognitive) information.

Results from previous studies using the ES task have been inconsistent (Williams et al., 1996) and may reflect the fact thatit is not a reliable measure. Reliability of the ES task has been investigated in only two previous studies (Kindt et al., 1996; Siegrist, 1997), with negative results. Kindt et al. found the reliability of emotional content words to be low (r = .19 and .25). This finding was replicated by Siegrist using self-relevant words (r = –.04).

A possible threat to the validity of these two studies is the confounding problem in the analysis of reliability of difference scores. This is the paradox for the measurement of change (Murphy & Davidshofer, 1994; Overall & Woodward, 1975). The paradox occurs when two variables are highly correlated. The reliabilities of the difference scores for these variables are then always low. The reason for this paradox is that all observed scores are assumed to be made up of true scores and error scores. If two scores are highly correlated, then the true scores must overlap considerably. Therefore, there will be hardly any difference between the true scores, and the difference seen will be almost entirely due to measurement error.

Address correspondence to P. J. Nathan, Brain Sciences Institute, 400 Burwood Road, Hawthorn, Victoria 3122, Australia; pnathan@bsi.swin.edu.au (e-mail).

Independent of this possible paradox effect, there were a number of minor problems in these previous studies that could be corrected in future research. First, although supposedly healthy controls were used in both studies, the researchers did not take into account the emotional condition of the participants, as there was no assessment of their emotional state. Second, the Kindt et al. (1996) study retested participants 3 months after the initial testing, although Carmines and Zeller (1979) recommended that retesting should be conducted no later than 1 month after the initial testing. Another potential problem is in the selection of the emotion-laden and non-laden words. Hill and Knowles (1991) attributed the lack of consistent findings to the selection of emotional words. They suggested that the emotional adjectives presented by Gotlib and McCann (1984) and Gotlib and Cane (1987) were the best types of words to produce an interference effect. Neither of the two previously published studies used the recommended emotional adjectives from the Gotlib studies.

Our aim in the present study was to examine the test–retest reliability of the ES task using the Gotlib adjectives. Of particular relevance to this objective was an assessment of the reliability of the interference effects. We examined the paradox within this context. We formed the following hypotheses:

1. If the variables are not highly correlated, the low reliability will indicate that the test should not be used in future research examining emotional processing in either nonclinical situations or for disorders of emotion.
2. If there is low test–retest reliability but the main variables are highly correlated, this will indicate that the test–retest statistic is not appropriate for this task.
3. A high test–retest reliability for the interference effect may indicate that the methodological differences between the present study and the two previously reported studies account for these differences and that the present methodology should be used in future studies assessing emotional processing.

Method

Participants

Participants were 33 members of the staff or student body at Swinburne University or acquaintances of the investigators (19 women and 14 men, with a mean age of 27 years [$SD = 5.54$]). We screened all the participants for any medical or psychological illnesses by having them complete a standard mental health questionnaire, and all provided informed written consent to participate. The research was approved by the Swinburne University Human Research Ethics Committee. All participants were of White European origin.

Materials

The task we administered was based on the ES task developed by Gotlib and McCann (1984) and consisted of color words presented on a computer screen using Arial Black font, size 72. In total, 120 words were selected from the list supplied by Gotlib and McCann. Of these words, 40 were depressive, 40 were neutral and 40 were manic. We presented the words in one of four colors; red, green, white, and blue. The object of the task was for participants to name the color of the words as fast as possible. We administered the Beck Depression Inventory II (BDI-II; Beck, Steer, & Brown, 1996) at both Time 1 and Time 2 to exclude any potential changes in mood over the testing interval.

Procedure

We tested the participants on two occasions, 1 week apart. We seated them approximately 1 m in front of a computer screen with headphones and a microphone to record response time (RT) for the ES task. We instructed the participants to say the color of the presented word as quickly as they could because we were recording their RTs. The words were presented for 1.5 s. Between each word presentation there was a blank screen for 250 ms and a fixation cross for 1 s. To minimize order effects, we gave half the participants the depressive words first, followed by neutral words, and finally the manic words. The remaining participants received the words in the opposite order (i.e., manic, followed by neutral and depressive words).

Results

BDI–II scores ranged from 0 to 13 and 0 to 12 at Sessions 1 and 2, respectively. The mean BDI–II score for Session 1 was 2.88 ($SD = 3.27$) and for Session 2, 2.19 ($SD = 2.74$); these means reflected very low scores on this scale. A two-tailed t test indicated that the change in the mean score was not significant, $t(31) = 1.71$, $p > .05$.

The mean RTs (from Session 1 to Session 2) were stable over time (Table 1). Correlations were also computed from RTs in the neutral, depressive, and manic conditions at Time 1 and Time 2. For neutral words, $r = .80$, $p < .01$; for depressive words, $r = .80$, $p < .01$; and for manic words, $r = .77$, $p < .01$. These correlations indicate that the RTs derived from each of the emotional conditions were consistent over time. For each category of words, the correlation was positive.

As stated previously, the ES task is an interference task. Hence, test–retest reliability is based on the correlations of the interference rather than the reaction times (RTs). To calculate the reliability of the interference effect, we subtracted the RTs for the emotional words from the neutral words (e.g., depressive interference $RT_{depressed} - RT_{neutral}$, and manic interference $RT_{manic} - RT_{neutral}$) for both

TABLE 1
Means and Standard Deviations for Reaction Time (ms) for the Different Stimulus Categories for the Emotional Stroop Task

Word type	Session 1		Session 2	
	M	*SD*	*M*	*SD*
Neutral	633.24	103.86	614.76	92.07
Depressive	657.64	105.36	625.55	102.89
Manic	627.42	120.85	627.74	100.20

Note. *N* varied between 30 and 32 due to missing values.

TABLE 2
Correlations of Emotional Words and Neutral Words for Reaction Time at Sessions 1 and 2

Word type	Time 1	Time 2
Depressed + Neutral	.90*	.90*
Manic + Neutral	.93*	.90*

Note. *N* varied between 28 and 32 due to missing values.
*$p < .001$.

Time 1 and Time 2. Test–retest reliability correlations for the interference effect of the ES words were $r = .24$, $p > .05$, for depressive words and $r = -.11$, $p > .05$, for manic words. These correlations were nonsignificant and unacceptably low for the purposes of test–retest reliability and indicate the lack of reliability for depressive and manic interference effect. Thus, the interference effects from Session 1 are not comparable to the interference effects seen at Session 2.

We conducted a further analysis on the RTs to examine the paradox for the reliability of difference scores. The correlations of RTs of emotional words with neutral words are contained in Table 2; all the correlations were positive, significant, and very strong, and *r* values were .90 or above. This result indicates that the RTs of the emotional words were strongly correlated with the RTs of the neutral words.

Discussion

Our major aim in the present study was to examine the test–retest reliabilities of the three emotion conditions as well as the interference indices. The inter-

ference indices had been previously hypothesized to reflect the bias in processing emotion-laden words. Our results indicated high test–retest reliabilities for the RTs derived from each emotion condition separately but low test–retest reliabilities for the interference indices.

Because the reliability of the interference indices was very low, additional analysis were conducted to confirm the paradox. The low test–retest reliability for the interference indices may have been because of the paradox of using change scores when two measures are highly correlated or, in fact, because of the very low reliability of the interference indices. RTs from all three conditions (depression, neutral, and manic) were highly correlated in the present study, indicating that test–retest reliability may not be a valid statistic to use in establishing the reliability of these indices. Perhaps it would be more suitable to use the test–retest reliabilities of the RTs derived from the words in the individual emotion conditions. These findings are consistent with past studies examining the reliability of the ES task (Kindt et al., 1996; Siegrist, 1997).

The modification of the present study using the Gotlib list of words produced test–retest correlations for RTs derived from the different emotion conditions similar to those reported in the two previous studies examining the reliability of the ES. The correlations previously reported were by Kindt et al. (r = .65 to .84) and by Siegrist (r = .84 to .91).

For both these previous studies (Kindt et al., 1996; Siegrist, 1997) as well as the current study, participants were recruited from a university population. The results of the present study indicate that the poor test–retest reliability of the interference effect of the ES is most likely to be due to the "paradox of measurement change." It is now important for researchers to determine how these interference reliability scores compare with scores in clinical populations, especially given that the ES task is considered an important tool in determining psychopathology in depressed and anxious individuals (Williams et al., 1996).

It is only possible to conclude from the present study that the ES task is reliable when measuring RTs derived from the different emotional conditions. Because of the difficulties in the analysis of difference scores, the reliability of the interference effects remains unknown. Nevertheless, because the RTs derived from the different conditions were highly correlated, it would be sensible to suggest that the interference produced by the ES is also likely to be highly reliable.

REFERENCES

Austin, M. P., Mitchell, P., Wilhelm, K., Parker, G., Hickie, I., Brodaty, H., et al. (1999). Cognitive function in depression: A distinct pattern of frontal impairment in melancholia? *Psychological Medicine, 29,* 73–85.

Beck, A. T., Steer, R. A., & Brown, A. K. (1996). *BDI-II manual.* San Antonio, TX: Harcourt Brace.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment.* Beverly Hills, CA: Sage.

Franke, P., Maier, W., Hardt, J., & Frieboes, R. (1993). Assessment of frontal lobe functioning of schizophrenia and unipolar depression. *Psychopathology, 26*(2), 76–84.

Gotlib, I. H., & Cane, D. B. (1987). Construct accessibility and clinical depression: A longitudinal investigation. *Journal of Abnormal Psychology, 96*(3), 199–204.

Gotlib, I. H., & McCann, C. D. (1984). Construct accessibility and depression: An examination of cognitive and affective factors. *Journal of Personality and Social Psychology, 47*(2), 427–439.

Hill, A. B., & Knowles, T. H. (1991). Depression and the "emotional" Stroop effect. *Personality and Individual Differences, 12*(5), 481–485.

Kinderman, P. (1994). Attentional bias, presecutory delusions and the self-concept. *British Journal of Medical Psychology, 67*(1), 53–66.

Kindt, M., Bierman, D., & Brosschot, J. F. (1996). Stroop versus Stroop: Comparison of a card format and a single-trial format of the standard color-word Stroop task and the Emotional Stroop task. *Personality and Individual Differences, 21*(5), 653–661.

LeDoux, J. E. (1990). Information flow from sensation to emotion: Plasticity in the neural computation of stimulus value. In M. Gabriel & J. Moore (Eds.), *Learning and computational neuroscience: Foundation of adaptive networks.* (pp. 3–52). Cambridge, MA: The MIT Press.

Lemelin, S., & Baruch, P. (1998). Clinical psychomotor retardation and attention in depression. *Journal of Psychiatry Research, 32*, 81–88.

Murphy, K. R., & Davidshofer, C. O. (1994). *Psychological testing: Principles and applications* (3rd ed.). Englewood, NJ: Prentice-Hall.

Overall, J., & Woodward, J. (1975). Unreliability of differences scores: A paradox for measurement of change. *Psychological Bulletin, 82*, 85–86.

Segal, Z. V., Gemar, M., Truchon, C., Guirguis, M., & Horowitz, L. M. (1995). A priming methodology for studying self-representation in major depressive disorder. *Journal of Abnormal Psychology, 104*(1), 205–213.

Siegrist, M. (1997). Test-retest reliability of different versions of the Stroop test. *The Journal of Psychology, 131*, 299–306.

Stroop, J. N. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 255*, 643–662.

Tarbuck, A. F., & Paykel, E. S. (1995). Effects of major depression on the cognitive function of younger and older subjects. *Psychological Medicine, 25*(2), 285–295.

Williams, J. M. G., Mathews, A., & MacLeod, C. (1996). The Emotional Stroop task and psychopathology. *Psychological Bulletin, 120*(1), 3–24.

Williams, J. M. G., & Nulty, D. D. (1986). Construct accessibility, depression and the Emotional Stroop task: Transient mood or stable structure? *Personality and Individual Differences, 7*(4), 485–491.

Original manuscript received January 8, 2001
Final revision accepted June 12, 2001