Pergamon

S0191-8869(96)00133-X

Person. individ. Diff. Vol. 21, No. 5, pp. 653–661, 1996

0191-8869/96 $15.00 + 0.00

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

Merel Kindt,[1]* Dick Bierman[2] and Jos F. Brosschot[1]

[1]Department of Clinical Psychology, University of Amsterdam, Roetersstraat 15, 1018 WB, Amsterdam, The Netherlands and [2]Department of Psychonomics, University of Amsterdam, The Netherlands

(Received 12 March 1996)

Summary—The card format and the single-trial format of the Stroop task are used alternately for the same purposes in general cognitive studies and in emotion studies. However, no convergent validity or test–retest reliability has ever been shown. In the present study, a card format and a single-trial format of a standard color-word Stroop and an emotional Stroop (i.e. spider words) were administered to a normal sample and re-administered after 3 months. Neither for the standard Stroop effect, nor the emotional Stroop effect, was there convergence between the two formats. The test–retest reliability of the standard Stroop effects of both formats was low but significant and the test–retest reliability of the emotional Stroop effects was very low. The results suggest that the two formats measure different underlying mechanisms and that both mechanisms are unstable. It is concluded that the application of cognitive paradigms within emotional research is of value if combined with the appropriate psychometric research. Copyright © 1996 Elsevier Science Ltd.

INTRODUCTION

The last decade has shown an emergence of the application of cognitive psychology to the investigation of emotional processes. Information-processing paradigms (e.g. Stroop task, Dichotic Listening task, Visual Deployment task), which have been developed to study cognitive phenomena, are now frequently applied to emotional phenomena (see for a review Logan & Goetsch, 1993). The usual approach is simply to add to a standard design an extra stimulus category of threat stimuli and an extra group of anxious *Ss*. The typical finding is that anxious individuals are distorted in processing threat related information, compared to neutral information. These distortions, demonstrated by differences in processing between emotional and neutral stimuli, are quantified by measuring reaction times. Differences between RTs to emotional and neutral stimuli are interpreted as emotional bias. Application of emotional bias research may be of interest when used for individual assessment—like an indicator for therapy effect or relapse—instead of using it only for group effects. Applying cognitive paradigms for this kind of research may add value to the usual assessment methods in that it is far less sensitive to demand characteristics and that it may reflect the background cognitive fear-network of emotionally disturbed individuals. Two important requirements for this kind of research are that processing bias is a stable phenomenon, and that it is open to manipulation at the same time. However, although recent years have shown a development of these kind of applications (Logan & Goetsch, 1993) no psychometric studies have been carried out to investigate the reliability of the modifications of these cognitive paradigms.

The implicit assumption of such studies is that the emotional modification of the cognitive paradigm measures qualitatively the same as the original cognitive paradigm. However, it has been demonstrated that emotional information is processed qualitatively differently and follows a different pathway than neutral information (LeDoux, 1990). LeDoux stated that the processing of emotional information involves an automatic and crude, "quick and dirty", transmission route, which signals to the amygdala that a significant stimulus is present. This suggests that different principles are valid for the original cognitive paradigms when compared with the emotional para-

* To whom all correspondence should be addressed.

digms. One of the most extensively used cognitive paradigms concerning emotional processing in anxious individuals is 'the emotional Stroop task', a modification of the standard color-word Stroop task (1935). Although there is still controversy about the mechanism which is measured by the standard color-word Stroop, research has shown that the easier a distractor stimulus activates the cognitive network, i.e. the more the distractor stimulus is automatized or the higher its accessibility is, the more interference is revealed (see for a review MacLeod, 1991). However, there is evidence that higher accessibility of emotional information does not increase the emotional Stroop interference. There are several reasons for doubting that this 'principle of automaticity' explains interference by emotional Stroop words.

When the distractor stimuli are words, the degree of automaticity may be operationalized as the effect of lexical frequency (frequency of use in common language) or familiarity of the words. According to Klein (1964) this would account for the greater interference produced by common words than by rare words, assuming common words are read (i.e. responded to) faster and more automatically than rare words. It is attractive to explain the finding that anxious individuals show more interference for threat words than for neutral words in terms of their high degree of familiarity with the threat words. This explanation, however, is in contrast with findings from studies in which these differences in interference disappeared after psychological treatment (Lavy & van den Hout, 1993; van den Hout & Arntz, 1993; Mathews, Mogg, Kentish & Eysenck, 1995; Mattia, Heimberg & Hope, 1993; Mogg, Bradley, Millar & White, 1995; Watts, McKenna, Sharrock & Trezise, 1986). It is not plausible that the degree of familiarity with the threat words is decreased after treatment. On the contrary, these words will even be much more familiar. Hence, emotional interference must be due to the emotional salience of the stimuli rather than to specific cognitive expertise (Watts, 1986).

Furthermore, in two recent studies (Riemann, Amir & Louro, submitted; Riemann & McNally, 1995) it was shown that the emotional Stroop interference was not attributable to lexical frequency. Despite large differences in lexical frequency between neutral words and emotional words, no interference for the emotional words was present in normal control *S*s, whereas this effect was present in anxious *S*s (Riemann *et al.*, submitted; Riemann & McNally, 1995). Hence, although automaticity explains the difference in the interference between non-words and words, this is probably not the case for the emotional Stroop effects. It is unlikely that the differences in readability that may still exist between reasonably well known emotional and neutral word sets, account for bias effects related to their meaning.

Some other research suggests that emotional interference is caused by a different process to the process underlying standard Stroop interference. MacLeod and Mathews (1991) showed that a difference between anxious and normal *S*s in the processing of threatening information only reveals itself in cognitive paradigms that require assignment of priorities to simultaneously available processing options. Threatening information is always prioritized compared to neutral information (see also LeDoux, 1990). For the emotional Stroop task, this means that the degree of distortion in color-naming threat words is determined by the degree to which processing resources are *not* allocated to the emotional information. This can also explain the finding, mentioned above, that anxious *S*s after treatment showed less interference for threat words than for neutral words (e.g. Mogg *et al.*, 1995). Thus, it appears that the emotional Stroop task is quite different from the standard Stroop task and that it may even measure a different phenomenon or mechanism. Considering this, it is surprising that little attention has been devoted to the psychometric properties of the method. In the present study the reliability of the emotional Stroop task and the original standard Stroop task, on which the emotional Stroop task is based, were investigated.

In studying the psychometric properties of the Stroop paradigm, it should be noted that there are important differences in the formats used to display the Stroop stimuli. In the original standard Stroop task, all stimuli of one word condition were presented on one single card. Since cognitive psychology requires a more analytic methodology whereby individual stimuli can be presented and timed, a single-trial method of the standard Stroop was developed. Dalrymple-Alford and Budayr (1966) found that a single-trial format yielded the same group effects as a card format, but they did not compare Stroop performance on an individual basis. Notwithstanding the fact that a computerized single-trial format has become predominant in the field of cognitive psychology, in clinical psychology the card format of the emotional Stroop task is still frequently used (e.g. Mathews &

Sebastian, 1993; Mattia *et al.*, 1993). In fact, in this field the computerized single-trial format and the card format of the emotional Stroop task are used as an alternative for similar research questions. However, so far no attempt has been made to show the degree of convergence between the two methods on the level of individual *S*, neither for the emotional Stroop task, nor for the standard Stroop task. In a recent study, we did not find any convergence between the two formats in children, i.e. for the standard Stroop interference nor for the emotional Stroop interference (Kindt, Bierman & Brosschot, submitted). One suggested explanation was that the two formats measure different mechanisms and/or one or both mechanisms are unstable.

The card format of the Stroop task obviously differs from the single-trial format in that stimuli are presented in a context of distractor words. Hence, in the card format a considerable part of the interference may be due to hindrance by context stimuli rather than hindrance of only the distractor word itself, which is integrated in the target stimulus. There is evidence that context stimuli can yield interference in a card Stroop task. This does not mean that the same mechanisms are involved in the interference generated by context when compared to interference generated by the integrated Stroop stimuli (MacLeod, 1991). Therefore, the single-trial format and the card format may measure different mechanisms.

Another possible reason for a lack of convergence in our previous study (Kindt *et al.*, submitted) is that the underlying mechanism assessed by one or both formats is unstable. Instability of the underlying mechanism may be due to limitation of the cognitive resources, a condition that may fluctuate in time. The finding of Dalrymple-Alford and Budayr (1966) that the two standard Stroop formats revealed the same group effects does not contradict this suggestion. For a given sample, on the average incongruent color words may be more difficult to ignore than non-words. However, the *S*s responsible for the group effect need not have been the same for the two formats. In addition, in a recent study by Dalgleish (1995) the two formats of the emotional Stroop task were administered to two samples of *S*s, which yielded different results. However, no study has compared the two formats within one and the same sample.

In sum, it is still unknown whether the card format and single-trial format of the emotional Stroop task and the standard Stroop task measure the same mechanism and whether this mechanism is stable. The present study was designed to investigate the test–retest reliability and convergent validity of the card format and the single-trial format of the standard and emotional Stroop task. The emotional Stroop task consisted of spider words and neutral words. Spider related information belongs to those unlearned emotional stimuli for which humans are believed to be evolutionary prepared to react to with priority (Seligman, 1971). For example, normal subjects show slower habituation of orienting response to pictures of spiders than to neutral stimuli (Öhman, Frederikson & Hugdahl, 1978) and a more rapid acquisition of conditioned fear response to these spider stimuli (Öhman, Erixon & Löfberg, 1975). Furthermore, spider words were shown to induce more interference than neutral words (Kindt *et al.*, submitted; Lavy & van den Hout, 1993; Lavy *et al.*, 1993; Watts *et al.*, 1986). If neither of the two formats shows adequate test–retest reliability, the use of emotional interference for the purpose of individual assessment is not justified. If there is no convergent validity for the two formats, whilst the test–retest reliability of both formats is moderate or high, different mechanisms must be involved that produce interference in the two Stroop formats. Exploratively, we investigated whether the degree of spider fear is related to: (a) convergence of the two formats; and (b) test–retest reliability of the two formats.

METHOD

Subjects

The sample consisted of undergraduate psychology students, who received course credit for their participation. Sixth-three *S*s (20 males and 43 females) participated in the first two sessions in which they performed a Card format and a Single-trial format of the Stroop task. The mean age was 21.4, ranging from 18 to 41 years (SD = 4.2). The Spider Phobia Questionnaire (SPQ) was used to assess subjective anxiety to spiders (Klorman, Weerts, Hastings, Melamed & Lang, 1974). The SPQ consists of 31 statements of self-reported fear that had to be answered with "true" or "false". *S*s completed this questionnaire in the second session after they had performed the Stroop task.

Thirteen *S*s dropped out before the two test–retest sessions of the Card format and the Single-trial format of the Stroop task. The mean age of the remaining 50 *S*s (14 males and 36 females) was 21.5, ranging from 18 to 37 years (SD = 3.7). Due to illness, two *S*s did not perform the retest of the Card format. hence, 48 *S*s performed the retest of the Card format and 50 *S*s performed the retest of the Single-trial format. None of the *S*s suffered from color blindness.

Materials

The Card format task as well as the Single-trial format consisted of both a Standard Stroop task and an Emotional Stroop task. The word sets in the Standard color-word were: (i) 'incongruent color words' (red, blue, yellow and green) and (ii) 'non-words' in the Dutch language (loav, tmelv, ernif, muga). These word sets were matched on word length and number of syllables. The word sets in the Emotional Stroop task were: (iii) Dutch 'spider words' (spider, web, hairy, legs, crawl) and (iv) Dutch 'control words' (sparrow, nest feather, flying, bird). Spider words and control words were also matched on word length and syllables.

Each word set consisted of 20 stimuli. The four non-words were presented in each of the four colors (red, blue, yellow and green), and once again in one of the four colors. The incongruent color words were never presented in ink of the same color name; every color word was presented twice in two of the four colors and once in one of the four colors. All words in the Emotional Stroop task were presented four times, each in one of the four colors. A total of 80 stimulus words were presented to each *S*. The *S* also received 20 practice stimuli consisting of non-experimental words.

The stimuli were presented in four blocked trials, each block consisting of one of the four word sets. There were 24 orders of presentation of the word sets. The order of presentation of the stimuli within each word set was one of two fixed random orders. Hence, there were 48 different orders of presentation of words. The only restrictions were that neither a word nor a color appeared more than twice in succession.

Apparatus

The Stroop words were presented to the *S* via an Apple Macintosh LC-II with a color monitor. In the Card format, stimuli were presented on the screen in four rows of five words. On the presentation of each word set, a fixation arrow appeared on the screen indicating where the *S*s should start in naming the color of each word. As soon as the word set appeared on the screen, *S*s started in naming the colors of the words. They read left to right and top to bottom. Timing began with the appearance of a fixation cross in the left top corner of the screen and stopped with the last color name. RTs for each word set were indicated by the experimenter on the keyboard.

In the Single-trial format, color naming responses were detected by a voice key, connected to the computer. Before the task, a voice test was applied to adjust the microphone to the individual average voice level. The RTs were recorded with millisecond accuracy. The presentation software recorded response latencies per word, operationally defined as the interval between stimulus word presentation and the detection of the vocal response. Errors were marked with the use of a standard key board, operated by the experimenter. The words appeared in the centre of the computer screen. On each color naming trial, a little fixation dot appeared at the centre of the screen 1500 msec before word onset. The word was displayed until the subject reacted, with a time-out of 3000 msec. If there was no record of a response within 3000 msec, the trial was considered as missing and registered as an error.

Procedure

Testing was conducted on four separate occasions. In the first session, *S*s were administered one of the two formats of the Stroop task: the Card format or the Single-trial format. On the second occasion, 1 week later, *S*s were given the remaining Stroop format test after which they filled in the Spider Phobia Questionnaire. To assess the test–retest reliability of the two formats of the Stroop task, *S*s were asked by mail to take part in two other sessions. There were 3 months between the second and third session. They performed the two retest Stroop tasks in the same order as before, with 1 week between the third and fourth session. The *S*s were tested individually. In each Stroop task, they were instructed to name aloud as fast as possible the color of the ink in which each word or non-word was written, while ignoring the meaning of the word. The Single-trial format started

with a voice-level test and a color blindness test, in which *Ss* were asked to name the color of blocks which were presented in the middle of the screen. Both Stroop formats started with 20 practice stimuli.

Design

The Stroop task consisted of a standard Stroop task and an emotional Stroop task. The within-*Ss* factor was Word Set (incongruent color words vs non-words in the analysis of the Standard Stroop effects and spider words vs control words in the analysis of the Emotional Stroop effects). In the first two sessions *Ss* were randomly allocated to one of the four conditions:

1. first the Single-trial format followed 1 week later by the Card format, both with the same order of word set ($N = 15$);
2. first the Single-trial format followed 1 week later by the Card format, with a different order of word set ($N = 16$);
3. first the Card format followed 1 week later by the Single-trial format, both with the same order of word set ($N = 16$);
4. first the Card format followed 1 week later by the Single-trial format, with a different order of word set ($N = 16$).

The between-*Ss* factor was Order of Presentation (Single-trial format–Card format vs Card format–Single-trial format). In the first session, *Ss* were randomly assigned to one of the 48 different orders of word sets. In the two test–retest sessions, the *Ss* were presented the same order of format and word set as in the first two sessions. Convergence and test–retest reliability were calculated by within *Ss* Pearson correlations.

RESULTS

Card format

Mean RTs of the Card format are presented for each stimulus type in Table 1. The mean RTs were subjected to two separate two-way ANOVAs for repeated measures with Word Set as a within-*Ss* factor and Order of Presentation as a between-*Ss* factor.

Analysis of the Standard Stroop data revealed a significant main effect of Word Set [$F(1,59) = 196.53$, $P < 0.0001$], with mean RTs 38% longer for the incongruent color words (17.5 sec per card) than for the non-words (12.7 sec per card). There was a main effect of Order of Presentation [$F(3,59) = 3.21$, $P < 0.05$]. However, we observed no interaction effect between Word Set and Order of Presentation [$F(3,59) = 2.66$, NS]. An analysis of the Emotional Stroop data showed that all *Ss* needed 5% more time to name the color of the spider words (12.9 sec per card) than to name the control words (12.3 sec per card). This difference was significant [$F(1,59) = 9.08, P < 0.01$]. There was also a significant main effect of Order of Presentation [$F(3,59) = 5.5, P < 0.01$], but there was no interaction effect [$F(3,59) = 1.09$, NS]. This indicates that there was neither a practice effect for the Standard Stroop effect nor for the Emotional Stroop effect.

Table 1. Mean color-naming latencies and standard errors for the Card format (in seconds) and the Single-trial format (in msec) in the first two sessions, and Percent of Errors (PE) for the Single-trial format

	Card format ($N = 63$)		Single-trial format ($N = 63$)		
Stimulus type	*M*	*SE*	*M*	*SE*	*PE*
Non-words	12.7	0.27	592	12	1.19
Color words	17.5	0.43	702	16	2.15
Control words	12.3	0.27	583	12	0.56
Spider words	12.9	0.27	597	14	0.64

Single-trial format

Mean color-naming latencies and the percentage of errors (PE) were calculated for each *S*. Mean RTs and PEs of the Single-trial format are presented for each stimulus type in Table 1. Errors and outlier latencies below 300 msec were eliminated from the analyses and outliers above 3000 msec were not recorded by the computer. The basic error rate was low, averaging 4.5%. Again analyses of variance for repeated measures were performed with one within-*S*s factor (Word Set) and one between-*S*s factor (Order of Presentation). Analysis of the RTs for the Standard Stroop yielded a significant difference of Word Set [$F(1,59) = 116.03$, $P < 0.0001$] with 19% longer RTs for incongruent color words (702 msec) than for non-words (592 msec). There was no effect of Order of Presentation [$F(3,59) = 1.45$, NS], and Word Set did not interact with Order of Presentation [$F(3,59) = 1.09$, NS]. Analysis of the PEs for the Standard Stroop revealed no effect of Word Set [$F(1,59) = 2.36$, NS], no effect of Order of Presentation [$F(3,59) = 0.87$, NS], and no interaction between Word Set and Order of Presentation [$F(3,59) = 0.15$, NS].

Analysis of the RTs for the Emotional Stroop showed no main effect of Word Set [$F(1,59) = 2.66$, NS] with mean RTs 2.5% longer for spider words (597 msec) than for control words (583 msec). The main effect of Order of Presentation [$F(3,59) = 0.17$, NS] and the interaction between Word Set and Order of Presentation [$F(3,59) = 0.08$, NS] did not reach statistical significance. Analysis of the PEs for the Emotional Stroop showed no effect of Word Set [$F(1,59) = 0.06$, NS], no effect of Order of Presentation [$F(3,59) = 1.26$, NS], and no interaction between Word Set and Order of Presentation [$F(3,59) = 1.65$, NS]. In general, both the Standard Stroop effect and the Emotional Stroop effect are half as large on the Single-trial format than on the Card format.

Convergent validity and test–retest reliability

Pearson's product–moment correlation coefficients were calculated between the color-naming RTs for each of the word categories on the Card format and the Single-trial format and between the tests and retests of these formats. The correlations are presented in Table 2. All the correlations were highly significant, indicating convergence and test–retest reliability with respect to the assessment of reaction–speed of both formats.

The most interesting correlations are those between the interference indices. Standard interference indices ($RT_{color} - RT_{non\text{-}words}$) and spider interference indices ($RT_{spider} - RT_{control}$) were computed for the Card format and the Single-trial format. Correlations between the interference indices are presented in Table 3. As can be seen from this table, there are practically no significant correlations between the standard interference indices on the two formats. These results indicate a lack of convergence between the two formats with respect to their measurement of Standard Stroop effect.

Table 2. Correlations between the colour-naming latencies

	Convergence between the Card format and the Single-trial format ($N = 63$)	Test–retest reliability of the Card format ($N = 48$)	Test–retest reliability of the Single-trial format ($N = 50$)
Non-words	0.52***	0.63***	0.73***
Color words	0.35**	0.64***	0.66***
Control words	0.38**	0.75***	0.73***
Spider words	0.57***	0.65***	0.84***

†trend; *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

Table 3. Correlations between the interference indices

	Convergence between the Card format and the Single-trial format ($N = 63$)	Test–retest reliability of the Card format ($N = 48$)	Test–retest reliability of the Single-trial format ($N = 50$)
Standard Stroop Interference	0.00	0.49***	0.29*
Spider Stroop Interference	0.10	0.19	0.25**

†trend; *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

This is remarkable when we take into account that both formats showed strong Standard Stroop effects. There was also no convergence between the two formats on the spider interference indices. For both types of interference indices there was sufficient variance for correlations to appear if they existed.

The test–retest reliability on the standard interference indices were moderate for the two formats. The correlation between the spider interference indices was low and non-significant for the Card format. For the Single-trial format there is only a trend to correlate for the spider interference indices.

Additional data analyses

Additional analyses were performed to investigate in more detail the Emotional Stroop data. Firstly, we wished to exclude the possibility that in the single-trial format the spider interference indices were just error measurements instead of systematically interrelated scores. Cronbach's alpha was calculated between the RTs of the different spider words when compared to the mean RT of the neutral words. Alpha was 0.84, which confirmed that these spider interference measures were not a random error, but were due to their common denominator, i.e. the spider-related content.

Secondly, as was suggested in the introduction the low reliability may be caused by resource fluctuations (e.g. loss of concentration). Highly anxious *S*s may suffer less from such resource fluctuations, because they are processing emotional stimuli that have a special significance for them. This may be reflected in higher convergence and/or test–retest reliability on the emotional Stroop task. To investigate this, *S*s were divided in three groups, on basis of the distribution of the SPQ-scores: high fear group (SPQ > 12), medium fear group (6 < SPQ < 13) and low fear group (SPQ < 7). Correlations between the spider interference indices of the two formats and between the tests and retests were calculated for the three different groups. The only notable divergence from the correlations calculated for the whole sample (see Table 3) was that the test–retest correlation of the Single-trial format was remarkably higher in the high fear group ($r = 0.54$, NS, $N = 8$) than in the medium fear group ($r = -0.10$, NS, $N = 13$) and in the low fear group ($r = 0.17$, NS, $N = 27$).

DISCUSSION

The present findings show that for the standard Stroop effect there is no convergent validity for the card format and the single-trial format. This is also the case for the emotional Stroop effect. Since there were no interactions between order of presentation and word set, the lack of convergence can not be explained by a practice effect. The test–retest reliability of the standard Stroop interference on the card format was moderate but significant on the single-trial format. The test–retest reliability of the spider interference on the card format was very low. For the single-trial format, the test–retest reliability of the spider interference was low but approached significance.

The lack of test–retest reliability on the emotional Stroop task indicates that processing bias for threatening information, as measured by the emotional Stroop task, might be an unstable phenomenon, caused by fluctuations on the allocation of resources (e.g. loss of concentration). To a certain extent, the same holds for the standard Stroop effect. One reason for resource fluctuations may be the double instruction which is given to the *S*s, that is to be fast *and* to be accurate. The combination of a speed instruction and accuracy instruction is in fact ambiguous and may therefore cause oscillations between these two strategies. The highest test–retest correlation for the standard Stroop effect was found on the card format, which was also the most difficult task in the experiment. This can be inferred from the fact that this task yielded the largest Stroop effect. If more resources are required to perform a task, then less fluctuation may be expected in resource allocation.

Reasoning along the same line, it is possible that since anxious individuals find it more difficult to ignore threatening words, their resources may show less fluctuations. This may give more stability to the performance of anxious individuals on the emotional Stroop task. Indeed, when only high fear subjects were included, the additional analysis revealed a marked increased in the test–retest reliability of the emotional Stroop interference on the single-trial format ($r = 0.54$). However, caution is needed when interpreting these findings because of the small number of high fear *S*s, and the absence of this finding for the card format. It should be mentioned that, in general, the delay in

response times on threat words is always less than on color-incongruent words. Therefore, the emotional Stroop task seems to drain fewer resources than the standard Stroop task. As a consequence, emotional Stroop performance will be more sensitive to resource fluctuations than standard Stroop performance. Therefore, although the test–retest reliability in anxious *S*s was the highest, it is unlikely that it will ever equal or exceed the already moderate test–retest reliability of the standard Stroop task. On the other hand, if the emotional Stroop task measures another process than the standard Stroop task, another factor may influence the resource fluctuations, i.e. the emotional significance of the stimuli. In that case, the test–retest reliability of the emotional Stroop task may be as high as was found in our high fear *S*s.

The combined finding of the moderately stable performance on the standard Stroop tasks and the fact that there was no convergence between the two formats, suggests that the two formats measure different mechanisms in the standard Stroop task. This may be explained as follows. Interference, caused by distractor stimuli presented in the context of the target, may involve another selective processing mechanism. This mechanism produces a type of interference that probably differs from the interference caused by a distractor integrated in the target stimulus. Although context stimuli produce interference effects, these effects are relatively small when compared to integrated stimuli (see for a review MacLeod, 1991). Different mechanisms may account for this divergence in results depending on the point in the cognitive process at which interference occurs. In an integrated task, the distractor stimulus is presented in the focus of attention, whilst contextual distractor stimuli are presented outside the focus of attention. As a consequence, it may be more difficult to inhibit the processing of an integrated distractor stimulus than the processing of context stimuli. The interference, caused by context stimuli in the card format, may occur at a more superficial level of processing than the interference caused by the integrated distractor stimulus present in both formats. Thus, the part of the interference in the card format caused by context stimuli, may be due to a different mechanism to that which causes the interference revealed by the single-trial format. However, we can not be sure about it, because of the apparent instability of the standard and emotional Stroop effects.

In summary, the Stroop task, when used in emotion research, appears to be only suitable to determine group differences between anxious and non-anxious *S*s. It is probably not suitable for assessing individual interference as a stable characteristic. Group effects may be robust in spite of individual fluctuations in the allocation of resources, because these fluctuations are embedded in the variance of the group. If other work confirms our finding that interference on the emotional Stroop task is more stable for *S*s with a high degree of fear, individual assessment of emotional Stroop interference should be restricted to these *S*s. On the other hand, if the test–retest reliability disappears with decrease of fear, individually assessed emotional interference can not be used for predicting individual relapse in previously anxious individuals. This makes the Stroop task a poor instrument for evaluating individual therapy effects on information processing. The present findings suggest that the two formats of the Stroop task measure different mechanisms and that these mechanisms are unstable, especially for the emotional Stroop effect. Both formats are frequently used for the same purpose of study. Hence, further research is needed to unravel the exact differences between the two formats, for the standard Stroop effect as well as the emotional Stroop effect. We conclude that the present study shows that cognitive paradigms are not applicable to emotional research without psychometric data.

Acknowledgements—We gratefully acknowledge Oliver Gebhardt for his grammatical advice, and Eldrid Robberstadt, who assisted in data collection.

REFERENCES

Dalgleish, T. (1995). Performance on the emotional Stroop task in groups of anxious, expert, and control subjects: A comparison of computer and card presentation formats. *Cognition and Emotion, 9*, 341–362.

Dalrymple-Alford, E. C. & Budayr, B. (1966). Examination of some aspects of the Stroop color-word test. *Perceptual and Motor Skills, 23*, 1211–1214.

Kindt, M., Bierman, D. & Brosschot, J. F. (submitted). Cognitive bias in spider fear and control children: assessment of emotional interference by a card format and a single-trial format of the Stroop task.

Klein, G. S. (1964). Semantic Power measured through the interference of words with color-naming. *American Journal of Psychology, 77*, 576–588.

Klorman, R., Weerts, T. C., Hastings, J. E., Melamed, G. B. G. & Lang, P. J. (1974). Psychometric description of some specific fear questionnaires. *Behavior Therapy, 5*, 401–409.
Lavy, E. & van den Hout, M. (1993). Selective attention evidenced by pictorial and linguistic Stroop tasks. *Behavior Therapy, 24*, 645–657.
Lavy, E., van den Hout, M. & Arntz, A. (1993). Attentional bias and spider phobia: conceptual and clinical issues. *Behaviour Research and Therapy, 31*, 17–24.
LeDoux, J. E. (1990). Information flow from sensation to emotion: Plasticity in the neural computation of stimulus value. In Gabriel, M. & Moore, J. (Eds) *Learning and computational neuroscience: Foundations of adaptive networks*. Cambridge, MA: MIT Press.
Logan, A. C. & Goetsch, V. L. (1993). Attention to external threat cues in anxiety states. *Clinical Psychology Review, 13*, 541–559.
MacLeod, C. M. (1991). Half a century of research on the Stroop Effect: An integrative review. *Psychological Bulletin, 109*, 163–203.
MacLeod, C. & Mathews, A. (1991). Biased cognitive operations in anxiety: Accessibility of information or assignment of processing priorities. *Behaviour Research and Therapy, 29*, 599–610.
Mathews, A., Mogg, K., Kentish, J. & Eysenck, M. (1995). Effects of psychological treatment on cognitive bias in generalised anxiety disorder. *Behaviour Research and Therapy, 33*, 293–303.
Mathews, A. & Sebastian, S. (1993). Suppression of Emotional Stroop effects by fear-arousal. *Cognition and Emotion, 7*, 517–530.
Mattia, J. I., Heimberg, R. G. & Hope, D. A. (1993). The revised Stroop color-naming task in social phobics. *Behaviour Research and Therapy, 31*, 305–313.
Mogg, K., Bradley, B. P., Millar, N. & White, J. (1995). A follow-up study of cognitive bias in generalised anxiety disorder. *Behaviour Research and Therapy, 33*, 927–935.
Öhman, A., Erixon, G. & Löfberg, I. (1975). Phobias and preparedness: Phobic versus neutral pictures as conditioned stimuli for human autonomic responses. *Journal of Abnormal Psychology, 84*, 41–45.
Öhman, A., Frederikson, M. & Hugdahl, K. (1978). Orienting and defensive responding in the electrodermal system: Palmar–dorsal differences and recovery rate during conditioning to potentially phobic stimuli. *Psychophysiology, 15*, 93–101.
Riemann, B. C., Amir, N. & Louro, C. E. (submitted). Cognitive processing of personally-relevant information in panic disorder.
Riemann, B. C. & McNally, R. J. (1995). Cognitive processing of personally relevant information. *Cognitive and Emotion, 9*, 325–340.
Seligman, M. R. E. P. (1971). Phobias and preparedness. *Behavior Therapy, 2*, 307–320.
Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 225*, 643–662.
Watts, F. N. (1986). Cognitive processing in phobias. *Behavioural Psychotherapy, 14*, 295–301.
Watts, F. N., McKenna, F. P., Sharrock, R. & Trezise, L. (1986). Color naming of phobia-related words. *British Journal of Psychology, 77*, 97–108.