*British Journal of Psychology* (1994), **85**, 413–432 *Printed in Great Britain*





# The sentence–picture verification task: Methodological and theoretical difficulties

**Maxwell J. Roberts***
*Department of Psychology, University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ*

**David J. Wood** and **David J. Gilmore**
*Department of Psychology, University of Nottingham, UK*

The sentence–picture verification task has received a great deal of attention over the years, and it has been a source of many interesting findings concerning the relationships between individual differences in people's abilities, task performance and strategy selection. However, there are a number of problems with interpreting data from this task as it appears to be particularly difficult to reliably classify subjects into distinct strategy groups. The purpose of this work was to review the various classification systems used by previous researchers and assess their validity. No system was found to be without problems, and it is suggested that previous conclusions derived from this task may not be justified.

The sentence–picture verification task is an important paradigm that has been used to investigate the relationship between individual differences in cognitive abilities and strategy selection. Interesting findings have come to light in the past, for example, those of MacLeod, Hunt & Mathews (1978) suggested that subjects have distinct preferences for representing and manipulating information (either as verbal propositions, or as spatial representations) and that these preferences are predictable from psychometric test scores, suggesting that this task is perhaps acting a possible indicator of styles of thought. However, recent work by Marquer & Pereira (1990) (see also Evans, 1990 in reply) suggests that the task is still not fully understood.

The task in its usual form consists of a number of trials, each consisting of two phases. In the first (the comprehension phase), subjects encode a simple sentence describing a pair of objects, one object being *above* or *below* the other. In addition half of the sentences are negated so that one object is *not above* or *not below* the other (these are known as *negative trials* as opposed to the *affirmative trials*). For the decision phase of a trial, the sentence is then removed and replaced by a picture that shows one of the objects above the other. The subject has to decide whether or not the sentence was a true or false description of the picture. For example, if the sentence is:

* Requests for reprints.

THE STAR IS ABOVE THE PLUS

and this is replaced with the picture

+
*

then the correct answer is false. If the same picture replaces this sentence:

THE PLUS IS NOT BELOW THE STAR

then the correct answer is true.

There are thus four types of trial; true affirmatives (TA), false affirmatives (FA), false negatives (FN) and true negatives (TA). The first example above is an FA trial, the second is a TN trial.

The task was first investigated by Wason (e.g. Wason & Jones, 1963; Wason 1972, 1980) from the point of view of how people process negatives. However, in addition to having to process negatives, subjects are faced with the problem of how to make comparisons between linguistic information in the sentence and spatial information in the picture. Some form of conversion must take place in order to compare the two.

## Strategies for solving the problems

### *The verbal strategy*

With this strategy, the sentence is encoded as a verbal proposition, and then compared with a propositional encoding of the picture. Carpenter & Just (1975) developed the most complete model of this strategy, which they named the Constituent Comparison Model. The two pairs of propositions are compared in a series of steps, component by component, until all components of both have either been matched or (if there is a mismatch) have been resolved and then matched. Resolving a mismatch causes a truth index to be changed from true to false (or back again), and on completion of all steps the answer may be read off from the truth index. Each of the four types of trial requires a unique number of steps for evaluation. For decision times, if it is assumed that encoding the picture and making the final response requires $k$ units of time, and that all comparison steps require equal (arbitrary) periods of time, then this means that TA trials (the easiest) require $k$ units of time, FA trials, $k+1$ units, FN trials, $k+4$ units and TN trials (the hardest) $k+5$ units. This model thus clearly predicts the rank ordering and the expected sizes of the differences between the four decision stage latencies. When plotted on a graph, with mean decision time against the number of comparisons, the means should lie on a straight line. This strategy may thus also be described as a *linear strategy* (see Fig. 1 below).

As to the exact nature of the processes taking place, Carpenter & Just (1975) were rather evasive, stating that 'The internal representation of a sentence is not necessarily linguistic in nature' and that '...there may be a level of representation that is neither linguistic nor pictorial in nature...' (p. 47). In subsequent work however, it has been assumed that these processes are either linguistic, or draw heavily on people's abilities to use language, hence *verbal strategy*.

### *The spatial strategy*

An alternative strategy is to form a spatial representation of the sentence, and to compare this with the picture (hence *spatial strategy*). This requires additional processing during the comprehension phase, but this extra expenditure is made up for by a far simpler comparison process. This strategy will only work with problems in which negatives are the unambiguous equivalents of affirmative statements. In a more realistic context, *the star is not above the plus* could mean that the star is below the plus, next to it, or absent altogether. However, in the context of this highly constrained task, the only possible alternative meaning of the sentence is that *the star is below the plus.*

Other *spatial algorithms* may appear to be equivalent (in terms of performance) to the use of spatial representations. For example, it is possible to use the sentence to decide which shape should be at the top of the picture, thus the decision process may be considerably speeded as only one shape in the picture needs to be attended to.

### *The recoding verbal strategy*

The final alternative is the *recoding verbal strategy.* This involves converting the negatives, where they occur, to affirmatives, and encoding a recoded sentence as a verbal proposition which is compared with a propositionally encoded picture (as for the verbal strategy). Again, in real life negatives cannot be recoded unambiguously, but in the context of this task, a sentence such as *the star is not above the plus* may safely be converted to *the star is below the plus* or to *the plus is above the star.*

From a performance point of view, the spatial strategy and the recoding verbal strategy are indistinguishable. The effects of negatives are completely removed during the comprehension phase, so that during the decision phase TA trials are equivalent to TN trials, and FA trials equivalent to FN trials. There is thus no affirmative-negative difference overall. As false trials still take longer than true trials, plotting the expected pattern of decision times on the same axes as for the verbal strategy, should give a flattened, inverted U, hence these strategies may together also be described as *flat strategies* or *quadratic strategies.* The ideal patterns of decision times for the verbal and the flat strategies are shown in Fig. 1.

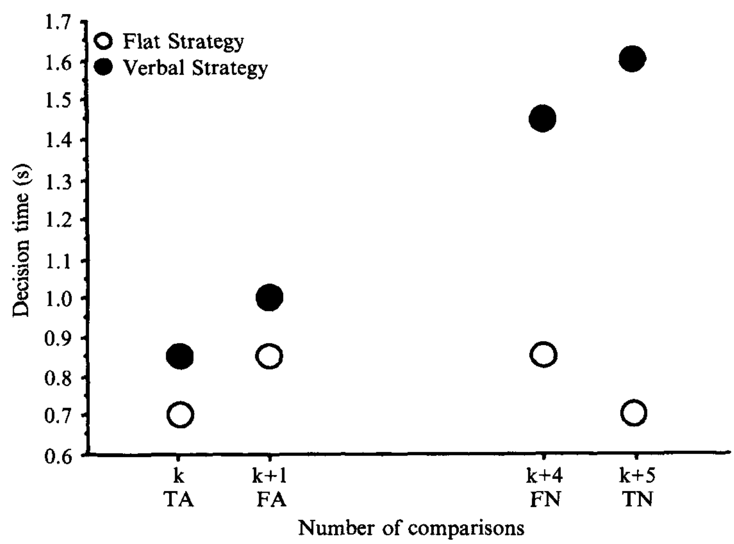


**Figure 1.** Ideal patterns of mean decision times for verbal and flat strategies.

## Theoretical approaches

### *Single models*

Initially, Clark & Chase (1972) and Carpenter & Just (1975) suggested that the vast majority of subjects used the verbal strategy (with this strategy drawing upon fundamental reasoning processes). They were aware that plausible alternative strategies could exist, but (for example) Clark & Chase dismissed their occurrence on two counts; (*a*) as they could not occur in real life, they were unlikely to be used for this task and (*b*) their data favoured the verbal strategy overall rather than the alternatives (i.e. possible individual differences were trivial). Neither of these dismissals is valid. There is no reason why subjects should not develop task specific strategies that enable them to solve problems more effectively, nor is it necessary for all subjects to conform to a single model. Siegler (1987) showed how it is possible for the performance characteristics of one strategy to conceal others if the measure to detect them is not sufficiently sensitive. One comment by Carpenter & Just (1975, p. 54) was particularly telling: 'The results were first analysed for only the first block of trials, since previous pilot studies... indicated a systematic change in the pattern of latencies across blocks of trials' (subjects received seven blocks of 36 trials). Hence they indicated that the verbal strategy is a rather transient, short-lived entity, that quickly changes into (what they considered to be) the recoding verbal strategy. This leads one to wonder exactly what the status should be of an all-encompassing model of performance that only applies to the first 36 trials of a 252 trial task.

### *Individual differences*

MacLeod *et al.* (1978) explicitly searched for individual differences by looking in detail at the performance of individual subjects. They found that the performance of some 25 per cent of their subjects fitted the verbal strategy very poorly indeed, matching the performance of the flat strategies instead. Psychometric test scores were used to validate the strategy groupings and to decide which of the alternative flat strategies was chosen by this group. Decision time performance of the flat strategy users was predicted by spatial ability but not verbal ability test score (the reverse being the case for those who used the verbal strategy). Moreover, the mean spatial score for the flat strategy subjects was higher than for the verbal strategy subjects (verbal ability being equally high in both groups). It was concluded from this that the flat strategy subjects were using the spatial strategy.

Further work by Mathews, Hunt & MacLeod (1980) replicated the aptitude–strategy-performance correlations (different abilities only predicting the performance of their associated strategies), but failed to replicate the aptitude–strategy relationship (high spatial subjects choosing the spatial strategy, low spatial subjects choosing the verbal strategy). One possible reason for this is that a different method was used to classify subjects into strategy groups. Their main finding was that it was possible to instruct subjects to use strategies other than the ones that they would have chosen naturally (with patterns of performance changing in predicted ways), though with mixed success as some subjects did have great difficulty in using these instructed strategies. The authors suggested that the ability for (most) subjects to change strategy on instruction meant that aptitude–strategy relationships were completely uninteresting (even if they could be found), as subjects could (fairly easily) be instructed to change. They thus ignored the question of whether some subjects were more able to change than others, and whether this also could be related to psychometric test scores, with people low in the ability required to execute a particular strategy being put in difficulty when instructed to use it. As it is never possible to instruct people as to how they should solve every task that they are likely to approach, the strategies that they adopt spontaneously can be extremely important. If some people have a natural preference, and perform poorly when instructed against this, then little has been gained in instructing them. For these reasons, the conclusions of Mathews *et al.* are incomplete and unnecessarily dismissive. Also, Coney (1988) has fully replicated MacLeod *et al.*'s aptitude–strategy relationship by classifying subjects into high or low psychometric ability groups and looking at the performance of these groups.

## Strategy classification systems

### *General issues concerning strategy classification*

For individual differences work to be successful, it is essential that a reliable, unambiguous strategy classification system exists. To this end, a number of criteria must be satisfied. These include the following:

1. Mis-classifications must be minimized.
2. When a group of people who supposedly used an identical strategy have been assembled, their overall performance should match the expected performance of that group.
3. As many different sources of converging evidence as possible should be drawn upon.
4. If there are strong disagreements between different sources of evidence, these need to be resolved.

Most systems of classification for this task have two components; one part uses decision time performance to distinguish verbal strategy subjects from flat strategy subjects (also creating a group of *mixed strategy* subjects, whose performance falls between the other strategies). The other part uses converging evidence to validate the grouping and to determine the exact strategy chosen by the flat strategy subjects. Examples of converging evidence that have been used include verbal protocols and psychometric test scores.

### *MacLeod, Hunt & Mathews* (1978)

MacLeod *et al.* measured how closely each subject's decision time data resembled the verbal strategy by finding the size of the correlation between the number of comparison steps for each trial type (0, 1, 4, and 5 steps for TA, FA, FN and TN trials respectively) and each corresponding mean decision time. This gives a measure of agreement with the verbal strategy with +1 being a perfect match, and 0 being complete disagreement – and therefore a flat strategy instead. Negative correlations are rare, and difficult to interpret where they do occur. There are two potential problems with this measure. It ignores the variance around each mean and it also removes information, thus, two people with very different patterns of data may be given exactly the same correlation.

Having calculated this measure, Fisher's clustering algorithm for one-dimensional data was then used to create the strategy groups. This involves ranking the set of correlations and then dividing subjects into two groups to maximize the between-group difference on this measure (using a *t* test). This typically gives one large cluster of values which matches the verbal strategy fairly well and a smaller group – typically 25 per cent of the subjects – that matches the verbal strategy very poorly (this consists of the flat strategy users). The larger group again has the algorithm applied to give a verbal strategy group – typically 50 per cent of all the subjects – and a mixed strategy group. The main problem with using cluster analysis is that a subject's classification depends upon all of the others rather than on absolute criteria. This will not be a problem if heterogeneous clusters really do exist. However, the algorithm will *always* find clusters, even when they do not obviously exist, and great care must be taken to ensure that the resulting strategy groups really do have some validity and that they are the best description of the data. Psychometric test data were used for this purpose, and the validity of this source of evidence and the assumptions upon which its use is based will be discussed later.

### *Mathews, Hunt & MacLeod (1980)*

Mathews *et al.* used the difference between mean affirmative (TA, FA) and mean negative (FN, TN) decision times as the basis of classification, the larger the difference, the closer the fit to the verbal strategy. Analogous methods have been used by Hartley & Coxon (1984) and Hartley, Dunne, Schwartz & Brown (1986). However, the expected relationship between affirmative–negative difference and strategy is only indirectly derived from the verbal strategy model. While the verbal strategy model predicts the ordering and expected relative differences between the means, all that follows from it is that the affirmative trials should be significantly slower than the negative trials, the difference does not need to be particularly large. It is quite possible for the means of a subject to fit the verbal model perfectly, but have a small affirmative–negative difference as is shown in Fig. 2. As it is also possible for a set of means with a very large difference not to fit the verbal strategy at all, there is plenty of scope for mis-classifying subjects. Mathews *et al.* (p. 538) state that using their method they obtained identical groups as when using MacLeod *et al.*'s method, which leads one to ask why they complicated the issue by introducing a new method?

### *Marquer & Pereira (1990)*

Marquer & Pereira suggested a completely different approach to strategy classification for this task based upon trend analysis of an individual's data. Each verbal strategy individual should have a significant linear trend and a non-significant quadratic trend, each flat strategy individual should have a non-significant linear

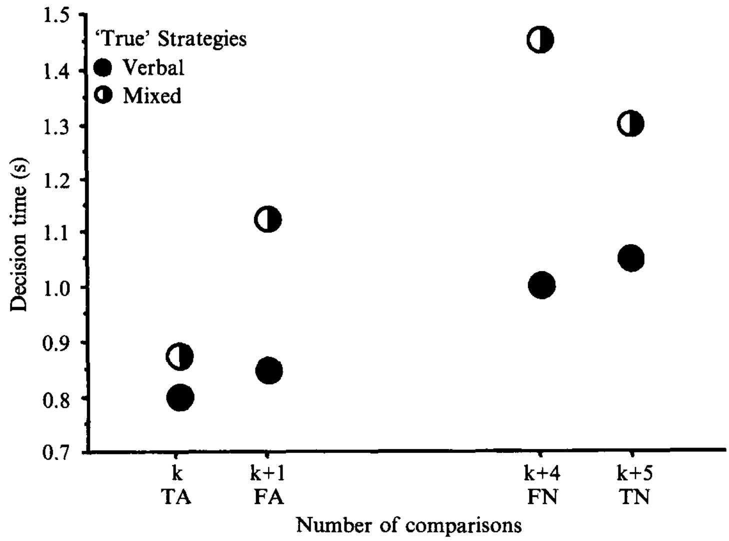


**Figure 2.** Possible mis-classifications due to using affirmative–negative differences as a means of classification.

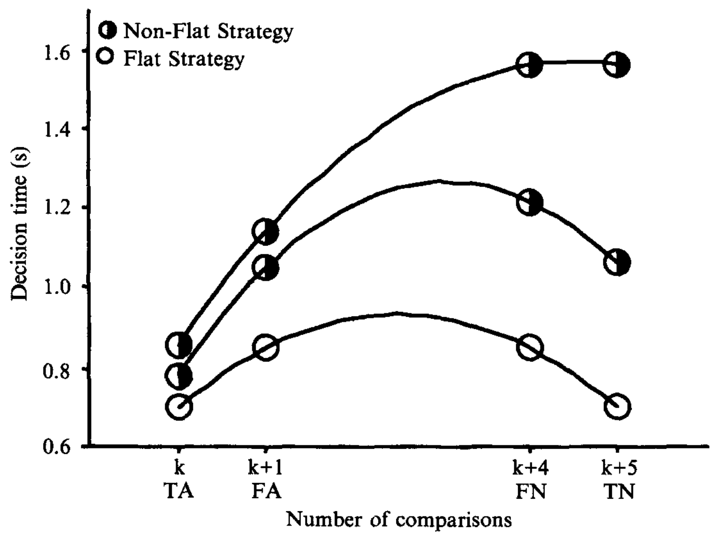


**Figure 3.** Examples of data points fitting quadratic curves.

trend and a significant quadratic trend (mixed strategy subjects would have both trends significant). Theoretically, this method has two advantages over that of MacLeod *et al.* All data points of each subject are considered rather than just the means, also, a more strict criterion than cluster analysis (statistical significance) is used to divide subjects into strategy groups. There is a disadvantage, however, which is that although a subject who is using a flat strategy *must* have a significant quadratic trend, it is possible for other patterns of data to do so too, as is shown in Fig. 3. One set of points clearly forms a flat strategy, the other two clearly do not, yet all fall exactly on quadratic curves.

Marquer & Pereira suggested that subjects should first be divided into strategy groups by using retrospective verbal reports of their strategy choice, and then the trend analysis should be used to see whether the verbal reports match the patterns of data, thus establishing the nature of the flat strategy group. Presumably, if this system is to be used, then if the verbal report of an individual disagrees with the performance based classification, then the data of that individual should be rejected as the strategy classification would therefore be ambiguous. If there is large-scale disagreement then verbal reports are not a valid source of converging evidence. Obviously, the utility of this strategy classification system depends upon the reliability with which subjects can report their reasoning processes, and this issue will be discussed later.

## Aims of this study

While at first sight, this appears to be a well-established paradigm associated with interesting findings, a closer look reveals that there may be problems with the strategy classification systems. The intention of this study was to collect data in order to assess the different methods of using response time performance to classify individuals into strategy groups (many methods exist but only the main three have been discussed here). If the methods disagree with each other, or any are found to be unsound (e.g. by failing the first two of the criteria for successful strategy classification mentioned earlier) then conclusions drawn from these studies may need to be reassessed.

## Method

### *Subjects*

Fifty-two undergraduates and postgraduates from the University of Nottingham were tested. Thirty-two of these were paid 50 pence for participation, the rest were unpaid volunteers. Eighteen were male and 34 female. All had normal or corrected-to-normal vision, and spoke English as a first language.

### *Apparatus*

Stimuli were presented on an Apple Macintosh Plus micro-computer running MacLab (Costin, 1988). Responses were made using the keyboard, with the computer recording latencies and responses.

### *Stimuli*

Stimuli were presented in four blocks of 32 trials, the first block being a practice block. All blocks contained identical stimuli, which consisted of two each of the eight possible combinations of:

The [plus/star] [is/is not] [above/below] the [star/plus]

followed by either a picture of a plus above a star or of a star above a plus (giving 16 possible combinations in total).

Each sentence always contained one instance of the word 'plus' and one of the word 'star'. Each block of trials thus consisted of eight each of TA, FA, FN and TN trials.

### *Design*

The four classes of trial form two within-subjects factors each of two levels (affirmation; affirmative or negative trials and answer; true or false trials). Order of trial presentation was randomized. Dependent variables were the comprehension time for each sentence and the decision time for each picture.

### *Procedure*

Subjects were given written instructions. They were asked to read the sentence in their own time, and then press a key that they would later press to make true or false responses. Following this, they were to respond true or false on being presented with the picture as quickly and as accurately as possible. For half of the subjects, comprehension and true responses were made with their dominant hand, and false responses with their non-dominant hand. For the other half, true responses were made with their non-dominant hand, and comprehension and false responses with their dominant hand. This counter-balancing of response collection is necessary to ensure that the overall results are not distorted by possible response biases caused by pressing one key three times as frequently than the other. It also prevents any biases due to comparing responses from the dominant hand and the non-dominant hand. Subsequently, keyboard layout was found not to interact with any other results.

The sequence of events for a trail was as follows: A blank screen was shown for 1.5 seconds followed by a fixation point for a further 1.5 seconds. The sentence was then shown, and was displayed until the subject had indicated that he or she had read it by pressing a key. This was followed by a 0.5 second gap, and then the picture, which remained upon the screen until the subject made his or her response. The cycle then began again. No feedback was given regarding performance.

## Results

### *Overall performance*

For each subject, six means were calculated from the three combined blocks of experimental trials (excluding the practice bloc). This gave a mean comprehension time for affirmative and negative sentences, and a mean decision time for the TA, FA, FN and TN trials. When an error had been made, the decision time and corresponding comprehension time were excluded from the analysis (as for MacLeod *et al.*). In addition, for each subject, response times that were greater than three standard deviations from their means were recursively removed (as for MacLeod *et al.*). Approximately 1 per cent of data points were thus excluded as outliers. Table 1 shows the overall performance for this task. Overall, performance fails to match the verbal strategy since the TN decision times are *not* slower than the FN decision times as would have been expected if the former required more processing steps than the latter. Although the pattern of data may imply that the verbal strategy dominates, it also indicates that other strategies are likely to be present too.

### *Individual differences and strategy classification*

*Comparing the classification system of MacLeod et al.* (1978) *with Mathews et al.* (*1980*). The overall means given above concealed a wide range of performance, from subjects who fitted the verbal strategy very closely, to those who bore no resemblance to it at all. Using the same measure of closeness of fit to the verbal strategy as MacLeod *et al.*, the four decision time means for each subject were correlated with the number

**Table 1.** Overall task performance (response times in seconds)

| | Comprehension times | | Decision times | | | |
|---|---|---|---|---|---|---|
| | AFF | NEG | TA | FA | FN | TN |
| Response time | 1.21 | 1.55 | 0.81 | 1.01 | 1.27 | 1.21 |
| (SD) | (0.41) | (0.56) | (0.28) | (0.37) | (0.50) | (0.56) |
| % error | — | — | 3 | 4 | 7 | 6 |

of comparisons predicted for verbal strategy users for each trial type. These correlations were found to range from 1.00 to −.71. The affirmative–negative difference for each subject was also calculated so that Mathews *et al.*'s method could be compared.

Using Fisher's clustering algorithm as described earlier on both sets of measures, verbal, mixed and flat groups were formed. Visually (as shown in Figs 4 and 5), no such clusters are present. Mathews *et al.* claimed that three clusters of values were clearly visible in their data, this is not the case for these data which appear to be unsuited for such a treatment.

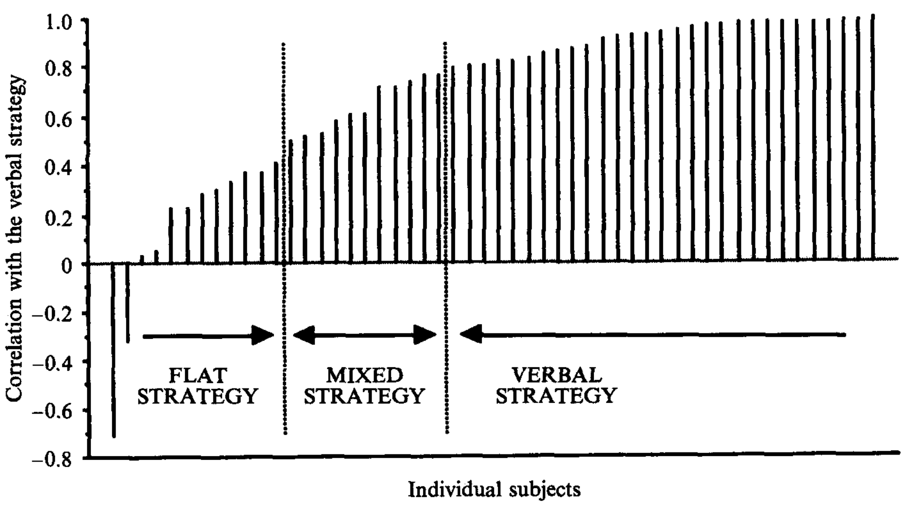


**Figure 4.** Individual correlations.

Table 2 compares the strategy classification of MacLeod *et al.* with that of Mathews *et al.* There *is* considerable disagreement between the two systems. Only 62 per cent of subjects are classified in agreement and this is a long way from the claim of Mathews *et al.* that the two methods produced identical groups. This perhaps accounts for the different findings of the two studies.

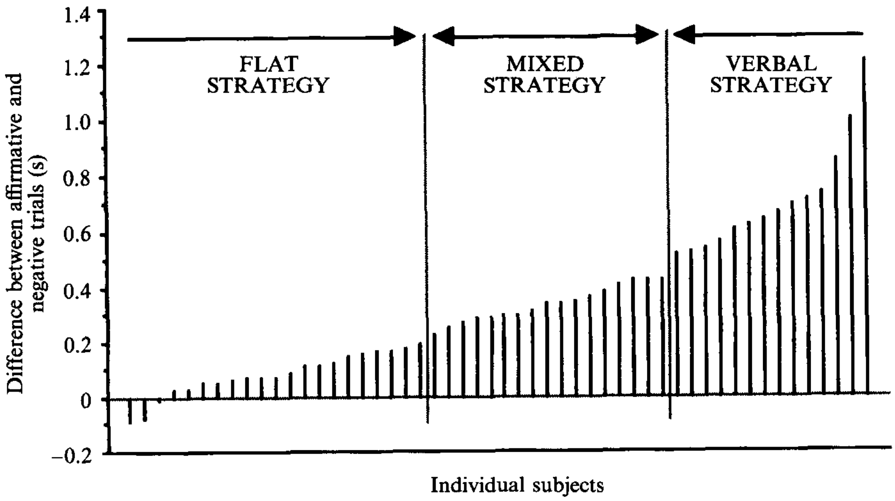


**Figure 5.** Individual affirmative–negative differences.

**Table 2.** Comparing classification systems: MacLeod *et al.* and Mathews *et al.*

| | MacLeod *et al.* | | |
|---|---|---|---|
| Mathews *et al.* | Verbal | Mixed | Flat |
| Verbal | 14 | — | — |
| Mixed | 11 | 6 | — |
| Flat | 4 | 5 | 12 |

*Investigating the classification system of MacLeod et al. (1978).* Using MacLeod *et al.*'s classification method, the three strategy groups were as follows: 29 subjects were classified as using the verbal strategy (correlations ranging from .79 to 1.00 with a median of .94), 11 subjects were classified as using mixed strategies (correlations ranging from .50 to .77 with a median of .61) and 12 subjects were classified as using the flat strategies (correlations ranging from −.71 to .41 with a median of .26). Table 3 and Fig. 6 show the mean performance for each of the groups.

While various methods may be used to classify individuals into strategy groups on the basis of their decision time performance, analysis of variance should be used to test group performance to ensure that an assembled strategy group is performing in line with expectations. It may be the case that although no clusters are apparent visually, the classification method used nevertheless places the boundaries in the correct places. If this is not the case however, group performance will not correspond with expected performance. It is not necessary to test the mixed strategy group however, since an expected pattern of performance cannot be specified for these subjects, performance is simply expected to fall somewhere between the other two

**Table 3.** Mean performance within each strategy group (response times in seconds)

| | Comprehension times | | Decision times | | | |
|---|---|---|---|---|---|---|
| Response time (SD) | AFF | NEG | TA | FA | FN | TN |
| Verbal strategy | 1.11 (0.34) | 1.42 (0.49) | 0.84 (0.33) | 1.04 (0.38) | 1.42 (0.55) | 1.45 (0.57) |
| Mixed strategy | 1.17 (0.31) | 1.48 (0.50) | 0.81 (0.22) | 1.11 (0.42) | 1.26 (0.40) | 1.13 (0.42) |
| Flat strategy | 1.47 (0.53) | 1.90 (0.67) | 0.72 (0.19) | 0.85 (0.23) | 0.92 (0.26) | 0.72 (0.20) |

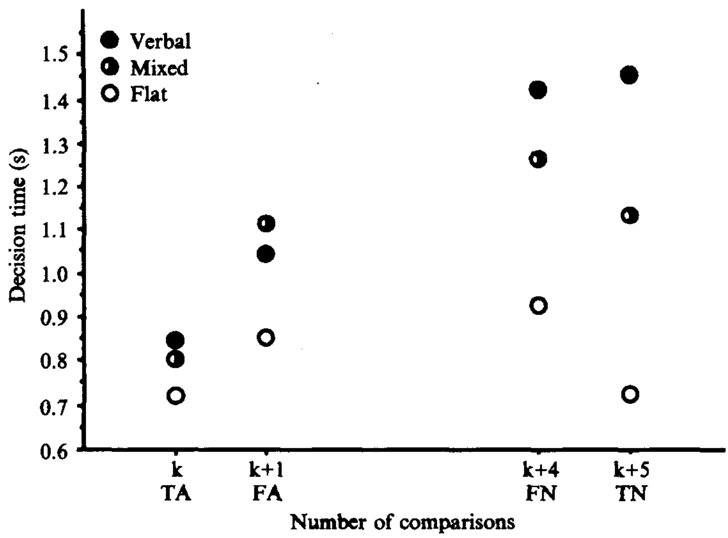


***Figure 6.*** Graph of mean decision times for each strategy group.

groups. MacLeod *et al.* and Mathews *et al.* also used analysis of variance to test their strategy groups, but failed to analyse interactions where they occurred, asserting that where a particular model predicted an interaction, and one was found to occur, that was sufficient.

In a $2 \times 2$ design (affirmation $\times$ answer) in which decision times are analysed, a group in which all members are using flat strategies would be expected to show no significant affirmation effect (affirmative trials the same speed as negative trials), a significant answer effect (true trials faster than false trials), and have no significant interaction (see Fig. 1). This was confirmed by the analysis; the affirmation factor was not significant ($F(1, 11) = 3.65$, $p > .05$), while the answer factor was ($F(1, 11) = 50.84, p < .01$), with no significant interaction ($F(1, 11) = 3.02, p > .05$).

A group in which all members are using the verbal strategy would be expected to show a significant affirmation effect (affirmative trials faster than negative trials), no significant answer effect (true trials the same overall speed as false trials), but with a significant interaction. However, here the verbal strategy group did not match the expected performance. There was a significant effect of affirmation ($F(1,28) = 103.46$, $p < .01$) *and* of answer ($F(1,28) = 7.55$, $p = .01$), the interaction was also significant ($F(1,28) = 16.77$, $p < .01$). Looking at the simple main effects, all were significant (all $F$s were greater than 20, $p < .01$) except for the answer factor at negative trials ($F(1,28) = 0.5$, $p > .05$), thus the required difference of TN trials being slower than FN trials was not to be found, despite the fact that TN trials require additional processing to FN trials. This failure to obtain a verbal strategy group that corresponded to the predictions suggests that other strategies (probably mixed) are present. It may even be the case that subjects who genuinely conform to the verbal strategy are very rare, with most taking some action to reduce the relative difficulty of the TN trials (perhaps by recoding negatives from time to time). Thus, the use of correlation followed by cluster analysis is not a sufficiently sensitive device to separate mixed strategy users from those who actually do only use the verbal strategy, because there is no clear difference between the two and hence no distinct clusters.

*Investigating the classification system of Marquer & Pereira (1990).* Contrast coefficients were derived that specified the (linear) verbal strategy (−.625, −.375, +.375, +.625) and the flat strategies (a quadratic curve, using the contrasts −.5, +.5, +.5, −.5). Subjects were re-classified to see whether their data contained a significant linear trend, a significant quadratic trend, or both (which would correspond to a mixed strategy). These classifications are compared with MacLeod *et al.*'s system in Table 4.

**Table 4.** Comparing classification systems: MacLeod *et al.* and Marquer & Pereira

| | MacLeod *et al.* | | |
|---|---|---|---|
| Marquer & Pereira | Verbal | Mixed | Flat |
| Linear | 22 | — | — |
| Mixed | 7 | 8 | 2 |
| Quadratic | — | 2 | 9 |
| None | — | 1 | 1 |

*Note.* Linear corresponds to the verbal strategy, while quadratic corresponds to the flat strategies. Subjects with significant quadratic trends only will continue to be described as using flat strategies.

As may be seen, although there is broad agreement between the two methods, there are some disagreements. In particular, quadratic trends were found in seven subjects who would have otherwise been classified as using the verbal strategy using

MacLeod *et al.*'s system, indicating that trend analysis may be more sensitive at detecting verbal strategy users. The mean decision times of the new groups were again tested using analysis of variance and are shown in Table 5.

**Table 5.** Mean decision times within each new strategy group (decision times in seconds)

| Decision time (SD) | TA | FA | FN | TN |
|---|---|---|---|---|
| Verbal strategy | 0.84 (0.37) | 1.01 (0.39) | 1.36 (0.53) | 1.43 (0.59) |
| Mixed strategy | 0.80 (0.23) | 1.08 (0.43) | 1.34 (0.57) | 1.24 (0.57) |
| Flat strategy | 0.78 (0.16) | 0.94 (0.21) | 1.05 (0.25) | 0.81 (0.18) |

At first sight, the verbal strategy group still did not come up to expectations. There was a significant effect of affirmation ($F(1,21) = 88.53, p < .01$) *and* of answer ($F(1,21) = 7.07$, $p < .05$), with the interaction also being significant ($F(1,21) = 16.51$, $p < .01$). However, looking at the simple main effects, all the expected differences were now to be found (all $F$s were greater than 20, $p < .01$), except for the answer factor at negative trials ($F(1,21) = 7.41$, $p = .01$), thus the required difference between FN and TN has at last been found.

Unfortunately, the flat strategy group did not perform as expected as the affirmation factor was now significant ($F(1,10) = 8.07$, $p < .05$), as well as the answer factor ($F(1,10) = 45.40, p < .01$), with no significant interaction ($F(1,10) = 3.17$, $p > .05$). Thus, although these subjects have no significant linear trend, affirmative trials are still slower than negative trials overall, and so this group of subjects cannot all be using purely flat strategies. This potential problem was highlighted earlier (see Fig. 3); a significant quadratic trend is not necessarily the same as the ideal pattern of performance that would be expected for someone using a flat strategy. Thus, this classification method forms a satisfactory verbal strategy group at the expense of the flat strategy group.

### *Summary*

The various classification systems evaluated have great difficulty in assembling homogenous strategy groups which perform as would be expected overall. The system used by MacLeod *et al.* runs into particular difficulty here. These difficulties highlight the necessity for converging evidence to be used, not only to identify the strategies used by those in the flat strategy groups (spatial or recoding verbal) but also to show that all strategy groups have been correctly assembled.

## Converging evidence used for strategy classification

*Psychometric test scores*

The conclusions that MacLeod *et al.* derived from the correlations between psychometric test score and strategy selection and between test score and performance were based upon two assumptions:

1. If subjects performing a task can be divided into two distinct strategy groups on the basis of their performance, and one group has higher spatial ability than the other, then the high ability group is likely to be using a spatial strategy, and the low ability group a non-spatial strategy.
2. If performance correlates with spatial ability and not verbal ability, then the processes used in performing the task necessarily involve spatial representations.

The first assumption is derived from a common-sense belief that people who have high spatial ability prefer to reason using spatial representations, while people with low spatial ability prefer to use verbal representations (avoiding the use of representation that draws upon ability that they do not have); and thus that people use a *make best use of resources* meta-strategy when deciding how to solve problems. This has been reiterated elsewhere (e.g. Carpenter & Just, 1986; Cooper & Regan, 1982; Onken, Hastie & Revelle, 1985; Pezdek, Simon, Stoeckert & Kiely, 1987). However, it needs to be asked whether either of these assumptions really do hold.

The ambiguousness of the nature of spatial ability (particularly high ability) has been highlighted by many researchers in the past (for example Barratt, 1953; French, 1965; Just & Carpenter, 1985; Kyllonen, Lohman & Snow, 1984; Kyllonen, Lohman & Woltz, 1984). For example, Kyllonen, Lohman and colleagues found that subjects who scored highest on spatial ability test type items were those who were the most flexible in strategy choice, yet they did not necessarily generate detailed spatial representations to solve the problems. Roberts, (1991) also found this counter-intuitive relationship between spatial ability and strategy selection. Thus, the first assumption does not hold because it has been shown that people with high spatial ability do not necessarily use spatial representations to obtain these high scores and therefore test score is *not* predictive of representation choice.

The second assumption does not hold because even if performance on a task is highly correlated with a spatial ability test score, no conclusions can be drawn about the use of spatial representations for the task, until the strategies used to solve the psychometric test items used to measure spatial ability have also been identified. High spatial ability appears to reflect the size of strategy repertoire and the ability to find efficient strategies for spatial tasks as much as (or more than) it reflects the ability and preference to generate and manipulate spatial representations, with the consequence that any correlation between test score and reaction times/error rates is completely uninterpretable.

Thus, psychometric test score correlations do not provide converging evidence as to the validity of the strategy groups except of a very circumstantial nature. While the spatial strategy group of MacLeod *et al.* is quite likely to contain some subjects using this strategy, it is also likely to contain subjects using the recoding verbal strategy, especially given that this is a simple and effective strategy for the task. As long as any given strategy is effective and workable, and not too obscure, then there is no reason why it should not be used by at least some subjects.

### *Verbal reports*

Marquer & Pereira suggest that retrospective verbal reports are a useful source of converging evidence for this task. The use of such reports to discover reasoning processes has attracted much controversy particularly since Nisbett & Wilson (1977) suggested that we may not necessarily have verbal access to our reasoning processes. Views as to the utility of reports range from '... verbal reports, elicited with care and interpreted with full understanding of the circumstances under which they are obtained, are a valuable and thoroughly reliable source of information about cognitive processes' (Ericsson & Simon, 1980, p. 247) to 'We should not comfort ourselves with the vain thought that if all our objective methods fail, we can, after all, just ask the subjects what they are thinking' (Evans, 1990, p. 170). For a brief review of the issues, see Gilhooly (1988).

Verbal reports may be considered to be a form of response data, and like all data they may be subject to error. This does not rule out the possibility that from time to time subjects may report correctly, indeed there are no hard and fast rules concerning whether or not a report is likely to be accurate, only degrees of safety. Thus, for example, it is not the case that concurrent reports are accurate and retrospective reports are not, retrospective reports are merely less likely to be accurate, so the question becomes: how much less likely? Generally, it is accepted that only the contents of working memory that have been fully attended to can be verbalized, and also processes if they have not been automated. If retrospective reports are to be taken, then it is essential that this information is transferred to long-term memory sufficiently intact to enable it to be described later. Subjects who report inaccurately may omit information; for example they may only report one strategy when they used two, or they may add spurious information perhaps by reporting using two strategies when only one was used, or in the absence of any awareness at all of how they solved the problems, inventing a plausible strategy to explain their behaviour.

Obviously, whether or not subjects can accurately report depends not only upon how the reports are collected but also upon the nature of the task itself, the processes that are being used to perform it and the use that is being made of working memory. Until more is known about cognition and metacognition, it cannot be known in advance with certainty whether or not a particular task will be associated with inaccurate reporting although some rough predictions can be made.

If retrospective reports are to be used to identify the strategies chosen for the sentence–picture verification task, two questions must be asked: (*a*) is it reasonable to suggest that people *could* have verbal access to their reasoning processes? and (*b*) is it possible to gather evidence to support this? The usefulness of verbal reports for casting light on the reasoning processes that have taken place cannot be dismissed out of hand until the nature of this task has been considered.

Verbal reports may serve two functions as converging evidence for this task:

1. If reported strategy corresponds to strategy classification on the basis of response

times for the vast majority of subjects, then we have evidence that the methods of classification discussed earlier, despite all their problems, have produced homogeneous groups of subjects which really do differ in terms of the strategies that they use, since they differ not only quantitatively in terms of performance but also qualitatively in terms of verbal reports. This would leave a small minority of ambiguous subjects who showed a disagreement between the two. Even if subjects cannot be expected to give precise reports of their strategies (and can only say, for example, whether or not imagery was present, or if they remembered recoding negatives), their verbal reports should be regarded as the products of the processes that they used, with different processes being expected to result in different reports.

2. If (1) above is the case, the flat strategy group can be investigated to see if the reports indicate who have used the recoding verbal strategy and who have used the spatial strategy.

Is it reasonable that people performing this task could have verbal access to their reasoning processes? There are some pointers. Aspects of the task are novel, so the processes will not be automated (at least to begin with). Those subjects who are classified as being in one of the pure strategy groups (verbal or flat, not mixed) should not be able to make the error of omitting to report one of several strategies that they used since they appear to have only used one. Trials were of moderate duration (the mean time for each trial was about three seconds including the time delay between comprehension and decision phases), certainly not so short as to rule out verbal access and transfer to long-term memory.

It is easy to identify the contents of working memory that could be reported. These contents must be attended to if the problems are to be solved correctly. With the spatial strategy, the sentences are converted into images. It would be reasonable to expect these subjects to be able to report that they were aware of imagery. If a person is classified as using a strategy in which imagery forms part of the reasoning processes, but denies any knowledge of using it, then clearly the classification has to be in doubt as one of the definitions of imagery is that people are aware of its presence. The reverse of this does not follow; people may report using imagery but this may be epiphenomenal and play no part in the reasoning processes. Hence, subjects should be classified by performance first, and then have this classification corroborated with their verbal reports rather than the other way around as suggested by Marquer & Pereira.

With the recoding verbal strategy and the verbal strategy users, the pictures must be converted to verbal representations. It would be reasonable to expect these subjects to be able to report that they were aware of pictures being represented in this way in working memory. It would also be reasonable to expect recoding verbal strategy users to report the process of recoding negative sentences, since this process cannot usually be applied in real life and so is unlikely to be automated. It would be less certain that people would be able to accurately report using the verbal strategy, especially given the uncertain nature of its representational status. As normal reading processes are heavily proceduralized and relatively rapid, they may well take place without subjects' awareness and so not be available for introspection.

Thus, those subjects classified as using flat strategies would be expected to report using flat strategies, while those classified as using the verbal strategy may be expected to report using a variety of strategies, not necessarily the verbal strategy. This is precisely the pattern of data obtained by Marquer & Pereira. Retrospective verbal reports were also collected for this study with identical results to Marquer & Pereira being obtained (see Roberts, 1991, for full details). Unfortunately, this means that for the group classified as using the verbal strategy on the basis of decision times, there are considerable numbers of subjects who give reports that disagree with performance, so there is no converging evidence that this strategy grouping is valid let alone homogeneous.

Marquer & Pereira attempt to get around this difficulty in two ways:

1. They suggest that subjects may carry over processing from the comprehension phase to the decision phase, thus they make the first key-press before forming the image or recoding the negative. This would make the recoding verbal and spatial strategies resemble the verbal strategy.
2. They suggest that some subjects may use idiosyncratic strategies in which normal strategies are distorted to resemble others (or none at all) due to the systematic execution of unnecessary extra processing steps.

There is little that experimenters can do about either of these problems if they wish to investigate problem-solving processes that are spontaneously chosen rather than those that they direct subjects to use. The difficulty is that if these explanations for disagreement between verbal reports and performance data are to be invoked, then every possible combination of verbal report and decision times becomes a valid strategy and it becomes no longer possible to determine whether (*a*) this is a task in which people generally are able to verbalize accurately or (*b*) whether particular individuals have failed to verbalize accurately. Thus either it must be accepted that current classification methods are unable to produce homogeneous strategy groupings or it must be accepted that verbal reports *are not* a useful source of converging evidence for this particular task. The evidence from Marquer & Pereira (1990) that retrospective verbal reports can correspond with subject performance was based upon a study in which subjects were first instructed to use a particular strategy before they embarked upon the task. It is likely that rather than reporting their memory of the strategy which they used, subjects were reporting their memory of the instructions, or at the very least using their memory of them to fill in gaps in their knowledge of how they actually performed. Because of this there is no evidence that verbal reports are accurate when subjects are not instructed to perform in any particular way.

## Conclusions

A number of problems with the sentence–picture verification task have been identified. If any of them had existed in isolation, then there would have been less cause for complaint, but the combination of these problems indicates that the use of this paradigm is fraught with difficulty. It has been found to be very difficult to take a group of subjects, and classify them as to the strategies that individuals are using with any degree of confidence. Problems were found with all the main classification systems based upon response time performance and also sources of converging evidence based upon psychometric test scores and verbal reports. It is difficult to see how more satisfactory alternatives could be developed although the possibility of using interference tasks does not appear to have been investigated.

Despite the problems with this paradigm, there is still the finding by MacLeod *et al.* (replicated by other researchers) that those whose data resembled flat strategies had higher mean spatial ability than those whose data resembled the verbal strategy (verbal ability being equally high in both groups). The principle that people choose strategies to make the best use of their resources does seem to be intuitively sound. It would be a very foolhardy subject indeed who had low spatial ability, yet persisted in using a spatial strategy (at which he or she would perform poorly) if other, better strategies were available. Yet, to the extent that they can be trusted, verbal reports suggest that the flat strategy group does not consist entirely of spatial strategy users, the majority reporting recoding verbal strategies. If further use is to be made of this paradigm for the investigation of individual differences in cognition, then it will be necessary explicitly to consider the nature of possible individual differences in strategies used for solving psychometric tests, and consequently to consider the exact meaning of a correlation between test score and task performance and/or strategy selection. Without this consideration, correlational studies investigating the relationship between task performance and cognitive abilities are likely to be uninterpretable.

These findings should by no means be taken as to suggest that individual difference work is inevitably unreliable as a whole, or too difficult to contemplate. Roberts (1993) suggests that in the field of human reasoning, individual difference work may be all that is possible. The problem is simply one of accurately identifying the different processes that are being used. Where tasks cannot yield this information except with great difficulty, it would be far better, if possible, to find other tasks subject to individual differences that have more reliable strategy classification systems available.

## Acknowledgement

This work was supported by a studentship from the Science and Engineering Research Council while the first author was at the University of Nottingham.

## References

Barratt, E. S. (1953). An analysis of verbal reports of solving spatial problems as an aid in defining spatial factors. *Journal of Psychology*, **36**, 17–25.

Carpenter, P. A. & Just, M. A. (1975). Sentence comprehension: A psycholinguistic model of verification. *Psychological Review*, **82**, 45–73.

Carpenter, P. A. & Just, M. A. (1986). Spatial ability: An information processing approach to psychometrics. In R. J. Sternberg (Ed.), *Advances in the Psychology of Human Intelligence*, vol. 3. Hillsdale, NJ: Erlbaum.

Clark, H. H. & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, **3**, 472–517.

Coney, J. (1988). Individual differences and task format in sentence verification. *Current Psychological Research and Reviews*, **7**(2). 122–135.

Cooper, L. A. & Regan, D. T. (1982). Attention, perception and intelligence. In R. J. Sternberg (Ed.), *A Handbook of Human Intelligence*. Cambridge: Cambridge University Press.

Costin, D. (1988). MacLab: A Macintosh system for psychology labs. *Behavior Research Methods, Instruments and Computers*, **20**, 197–200.

Ericsson, K. A. & Simon, H. A. (1980). Verbal reports as data. *Psychological Reviews*, **87**(3), 215–251.

Evans, J. St B. T. (1990). Verbal reports of cognitive strategies: A note on Marquer and Pereira. *Quarterly Journal of Experimental Psychology*, **42**A, 169–170.

French, J. W. (1965). The relationship of problem-solving styles to the factor composition of tests. *Educational and Psychological Measurement*, **25**, 9–28.

Gilhooly, K. J. (1988). *Thinking, Directed, Undirected and Creative.* London: Academic Press.

Hartley, L. & Coxon, L. (1984). Effect of alcohol on a sentence–picture verification task. *Ergonomics*, **27**, 195–203.

Hartley, L., Dunne, M., Schwartz, S. & Brown, J. (1986). Effects of noise on cognitive strategies in a sentence verification task. *Ergonomics*, **29**, 607–617.

Just, M. A. & Carpenter, P. A. (1985). Cognitive coordinate systems: Accounts of mental rotation and individual differences in spatial ability. *Psychological Reviews*, **92**, 137–172.

Kyllonen, P. C., Lohman, D. F. & Snow, R. E. (1984). Effects of aptitudes, strategy training and task facets on spatial task performance. *Journal of Educational Psychology*, **76**, 130–145.

Kyllonen, P. C., Lohman, D. F. & Woltz, D. J. (1984). Componential modeling of alternative strategies for performing spatial tasks. *Journal of Educational Psychology*, **76**, 1325–1345.

MacLeod, C. M., Hunt, E. B. & Mathews, N. N. (1978). Individual differences in the verification of sentence–picture relationships. *Journal of Verbal Learning and Verbal Behavior*, **17**, 493–507.

Marquer, J. M. & Pereira, M. (1990). Reaction times in the study of strategies in sentence-picture verification: A reconsideration. *Quarterly Journal of Experimental Psychology*, **42**A, 147–168.

Mathews, N. N., Hunt, E. B. & MacLeod, C. M. (1980). Strategy choice and strategy training in sentenced-picture verification. *Journal of Verbal Learning and Verbal Behaviour*, **19**, 531–548.

Nisbett, R. E. & Wilson, T. DeC. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, **84**, 231–259.

Onken, J., Hastie, R. & Revelle, W. (1985). Individual differences in the use of simplification strategies in a complex decision-making task. *Journal of Experimental Psychology: Human Perception and Performance*, **11**, 14–27.

Pezdek, K., Simon, S., Stoeckert, J. & Kiely, J. (1987). Individual differences in television comprehension. *Memory and Cognition*, **15**, 428–435.

Roberts, M. J. (1991). Individual differences and strategy selection in problem solving. Unpublished doctoral dissertation, University of Nottingham.

Roberts, M. J. (1993). Human reasoning: Deduction rules or mental models, or both? *Quarterly Journal of Experimental Psychology*, **45**A, 569–589.

Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General*, **116**, 250–264.

Wason, P. C. (1972). In real life, negatives are false. *Logique et Analyse*, **57–58**, 19–38.

Wason, P. C. (1980). The Verification Task and Beyond. In D. R. Olsen (ed.), *Social Foundations in Language and Thought: Essays in Honour of Jerome S. Bruner*. New York: Norton.

Wason, P. C. & Jones, S. (1963). Negatives: Denotation and connotation. *British Journal of Psychology*, **54**, 299–307.

*Received 21 August 1992; revised version 6 October 1993*