



The basis of report-difference superiority in delayed perceptual comparison tasks

Michael Pilling¹ · Douglas J.K. Barrett² · Angus Gellatly¹

Published online: 20 February 2020
© The Psychonomic Society, Inc. 2020

Abstract

A major role for visual short-term memory (VSTM) is to mediate perceptual comparisons of visual information across successive glances and brief temporal interruptions. Research that has focused on the comparison process has noted a marked tendency for performance to be better when participants are required to report a *difference* between the displays rather than report the absence of a difference (i.e. a *sameness*). We refer to this performance asymmetry as *report-difference superiority* (RDS). It has been suggested that RDS reflects the operation of a reflexive mechanism that generates a mismatch signal during the comparison of visual input with information maintained in VSTM. This bottom-up mechanism therefore gives evidence for the presence of a feature change but not for the absence of such a change; consequently, a sameness is harder to detect than a difference between two displays. We test this explanation, and determine whether by itself it is a sufficient explanation of the RDS. In a delayed comparison task we find the RDS effect is most prevalent when items retain the same display locations; however, the effect does persist even when compared item locations were scrambled across memory and test arrays. However, with a conjunction task this scrambling of locations was effective in wholly abolishing the RDS effect. We consider that the RDS effect is a consequence of local comparisons of features, as well as global statistical comparisons.

Keywords Perceptual comparisons · Visual short-term memory · Location · Global statistics

Introduction

Visual short-term memory (VSTM) is a capacity-limited store of recently viewed visual information (Cowan, 2001; Phillips, 1974). VSTM performs a role in mediating perception across consecutive glances or brief visual disruptions (Luck, 2006; Hollingworth, 2006). Consequently, VSTM is involved in tasks that require perceptual comparison of visual information that is temporally separated in some way or which is derived across successive views (Markman & Gentner, 2000). This role can be seen in tasks such as the detection of change (Luck & Vogel, 1997; Pashler, 1988), simultaneous comparisons (Huang, 2010; Scott-Brown, Baker, and Orbach, 2000), matching-to-sample (Mangini, Villano & Crowell, 2010), probe-matching (Griffin & Nobre, 2003), and the detection

of sameness/feature repetitions (Hyun, Woodman, Vogel, Hollingworth & Luck, 2009; Wilson & Goddard, 2011).

There is a curious asymmetry in the efficiency with which VSTM comparisons are made. Specifically, participants tend to be markedly better at reporting a *difference* between two displays, i.e. reporting that a feature or object has changed, than they are at noticing a *sameness* between displays, i.e. reporting that a feature or object has *not* changed (Hyun et al., 2009; Pilling & Barrett, 2016).

This finding of *report-difference superiority* (hereafter, RDS) was noted some time ago in work looking at perceptual comparisons (Egeth, 1966; Farrell, 1985; Taylor, 1976). However, this earlier work treated perceptual comparisons as something only involving current visual processing and did not consider the likely involvement of VSTM.

Hyun and colleagues (Hyun, et al., 2009) were the first to directly explore this asymmetry from the perspective of VSTM. In a critical experiment they presented two displays, each containing four disks arranged as a notional square, in sequence interleaved by a blank interval. In their task the number of colours that differed across the two displays was parametrically varied, on some trials all the disks retained the same colour across the displays, in others one to four of the

✉ Michael Pilling
mpilling@brookes.ac.uk

¹ Psychology, Faculty of Health and Life Sciences, Oxford Brookes University, Oxford OX3 0BP, UK

² School of Psychology, University of Leicester, Leicester, UK

disks were given a different colour, one previously not present in the array. The experiment presented participants with two displays each with coloured disks. Participants had to perform one of two types of comparison task, each of which required a binary decision. Under *any difference* instruction they had to report if any (i.e. one or more) of the colours was *different* between the two displays, under *any sameness* instruction they had to report if any (i.e. one or more) of the colours was *the same* across the displays.

Hyun et al. (2009) found that performance was markedly better for the *any difference* task. This was most clearly exemplified in the condition where two of the colours were different and two the same, a positive response was required under both task instructions, and there were the same number of critical features present for each task instruction. Despite this, participants were faster and more accurate in detecting the difference than when required to detect the absence of a difference.

Hyun et al. (2009) explained the RDS as a consequence of the manner in which information is compared between a VSTM-held representation and a current visual input. It was suggested that difference judgements are supported by the existence of a reflexive mechanism in the brain which automatically compares the locations of a VSTM-held object with those derived from current input. In this explanation, the locations in the test array are automatically compared in parallel with the VSTM-held representation of the memory array. Where the features are different, a transient signal is generated that tends to attract focal attention and a subsequent confirmatory active comparison of the location contents. In the case of a sameness, the absence of a mismatch doesn't generate any such transient in the initial reflexive comparison; the consequence of this is that sameness detection is wholly reliant on active serial comparisons between memory-held and viewed representations, meaning that performance tends to be poorer. Hyun et al. supported this interpretation with evidence from eye movements and EEG recordings as well as behavioural data, all of which indicated that there is automatic registration of a perceptual mismatch event and that this tends to influence internal attention. Others in the literature have similarly suggested that detection of change is influenced by an initial automatic comparison stage that guides a second confirmatory comparison process (Gilchrist & Cowan, 2014; Pilling & Barrett, 2018; Yin, Gao, Jin, Ye, Shen, and Shui, 2011)

The purpose of the current experiments was to evaluate whether this proposed mechanism is by itself *sufficient* explanation for the report-difference advantage. We suspected, a priori, that there may be at least one additional process underlying the RDS. Specifically, it is known that observers are sensitive to *global statistical properties* (Ariely, 2001; Chong & Treisman, 2003). It is known that such statistical information is generated automatically from viewed displays (Chong & Treisman, 2005), and that observers are sensitive to the changes in statistical properties across temporally

separated displays (Brady & Alvarez, 2011; Haberman & Whitney, 2011). The original claims were that such global statistical information was extracted in an entirely pre-attentive manner; however, more recent work suggests that this process utilises attention in the same way as does attending to object features individually (e.g. Huang, 2015; Jackson-Nielsen, Cohen & Pitts, 2017)

To reiterate, Hyun et al. propose that RDS is explained wholly as a consequence of a putative reflexive mechanism that makes local comparisons between spatial locations in VSTM-held representations and those of current input. If this is the case, then RDS should depend entirely on the compared items being ones that maintain their spatial configuration between memory and test. If the item locations are scrambled between memory and test then a local comparison mechanism could not provide any useful information about the presence or absence of a feature change. Consequently, under these conditions, RDS should no longer occur.

If, however, detection of difference is, in part, supported by comparison of global statistics (e.g. comparison of mean hue values), then the report-difference advantage might persist even when required to compare items with scrambled locations. Experiment 1 tested this possibility.

Experiment 1

A paradigm based on that described by Hyun et al. (2009) was given in which participants are presented with displays consisting of four coloured disks in a memory display and a following test display.¹ On half the trials all the disks retained the same spatial locations they had across the memory and test displays (unscrambled trials), on the other half of trials (scrambled trials) all the disks changed locations between the memory and test display; on such trials the disks were presented at locations that were previously unoccupied in the memory array.

Either all test disks are in the same colours as the memory display, or some or all of them are a different colour. In one set of trials participants have to report whether any of the four given colours were *different* across the two displays (hereafter, *report difference task*), in another set of trials, participants have to report if any given colours were *the same* across the two displays (hereafter, *report sameness task*). Following Hyun et al. (2009), the manipulation of the number of colour differences across the memory and test displays was expressed in both tasks as the *number of*

¹ We opted for a task where observers made unsped responses to the displays, as is the case in most comparison tasks (e.g. Luck & Vogel, 1997), rather than one that emphasised speed of responding, as Hyun et al. (2009) did. We opted for this because our pilot work in which both speed and accuracy of responding was emphasised produced data that contained differential speed-error trade-offs that made it difficult to validly compare across the conditions.

critical features (nCF) with respect to the task instructions. For the difference task, each colour that was *different* between memory and test was viewed as a critical feature. For the sameness task, each colour that was *the same* between memory and test was viewed as a critical feature. This meant that identical trial sequences diverged in terms of the number of critical features for the two tasks. For example, if one of the four colour items changed between memory and test, there would be one critical feature (1-CF) under any difference instructions, but three critical features (3-CF) under any sameness instructions (because three of the four features are the same). Note that the 2-CF condition is possibly the most interesting in this respect. It is in this condition alone that there is an equal number of same and different coloured items across the memory and test displays.

The unscrambled trials mimic the circumstances in which Hyun et al. found a RDS. It was expected that this effect would replicate here. The crucial question concerned relative performance on the scrambled trials. Here local comparison mechanisms are of no value in detecting differences because the memory and test stimuli are in different locations and in a different configuration. If RDS is sustained solely by this local comparison mechanism then it should be abolished under these conditions. If, however, the advantage is also supported by a mechanism that makes global statistical comparisons of the two displays (Brady & Alvarez, 2011; Haberman & Whitney, 2011), then an amount of RDS should persist in these scrambled conditions.

Methods

Participants

There were 24 participants (14 female). All were staff or students recruited from Oxford Brookes University. All had normal or corrected-to-normal vision. The number of participants was determined by the conventions of sample sizes for experiments within this field.

Stimuli

Stimuli were shown on a 15-in. Sony Trinitron CRT monitor controlled by a Quad-core PC fitted with an NVIVO graphics card. The monitor was viewed from an approximate distance of 450 mm in a darkened and sound-deadened room. The stimuli consisted of coloured disks (diameter subtending a viewing angle of 3.9°). These colours are each listed with their descriptive name, DAC (RGB) values and then their luminance (cd/m^2), and CIE (1932) chromaticity coordinates (x, y). The colours are: *black* (0, 0, 0; 0.08, .338, .361), *red* (255, 0, 0; 20.07, .625, .341), *green* (0, 255, 0; 66.57, .284, .611), *yellow*

(255, 255, 0; 86.41, .404, .516), *blue* (0, 0, 255; 7.91, .150, .072), *orange* (255, 140, 0; 32.66, .538, .410), *pink* (255, 192, 203; 55.27, .360, .333), *brown* (139, 69, 19; 4.12, .560, .389), *purple* (128, 0, 128; 3.09, .313, .166), *sky-blue* (161, 224, 255; 60.09, 235, .299), *peach* (255, 204, 153; 58.50, .405, .412) and *magenta* (255, 0, 255; 27.97, .316, .166). All stimuli were presented on a *grey* background (170,170,170; 30.79, .302, .354). The fixation cross presented on all stimulus frames was *white* (255, 255, 255; 94.58, .304, .341)

Procedure

Participants did both the *report difference* and *report sameness* tasks. These were done in counterbalanced order. Participants were given a demonstration and did 30 practice trials before beginning each task. A short break was given between the tasks. The whole experimental session took approximately 50 min to complete.

For both tasks a trial began with a fixation screen for 250 ms. This consisted of a white fixation point (of size subtending 0.2° of visual angle at the approximate viewing distance.) The fixation point remained on-screen until the instruction screen following the test display. The memory display followed the initial fixation and was presented for 200 ms. The memory display contained four coloured disks presented at random locations on-screen with the constraint that the centre of the disk was within a radius of 5.5° (min.) to 11.5° (max.) from the screen centre, and was also at least 6.9° from the centre of any other disk in the display. The colours were randomly drawn without replacement from the set of 12 colours earlier described. The memory display was followed by a blank inter-stimulus interval (ISI) for 500 ms, and then the test display also for 200 ms. The test display always consisted of four coloured disks. On *unscrambled trials* the four disks were presented in the same locations as the disks in the memory display. On *scrambled trials* the four disks were each presented at new randomly selected locations. The test items were also within a 5.5° to 11.5° radius from the screen centre. Items were also each at least 6.9° from the centre of any other disk in the test display and also at least 6.9° from the centre of any item in the previous memory display.

For both report conditions half the trials had 0-CFs. On the other half of trials there were 1-, 2-, 3- or 4-CF trials, given in equal amounts. On trials where there were one or more different colours presented in the memory display (i.e. 1- to 4-CF for the difference task, 0- to 3-CF for the sameness task) these colours were randomly selected without replacement from the set of remaining eight possible colours from the set. A schematic depiction of an example trial is given in Fig. 1A.

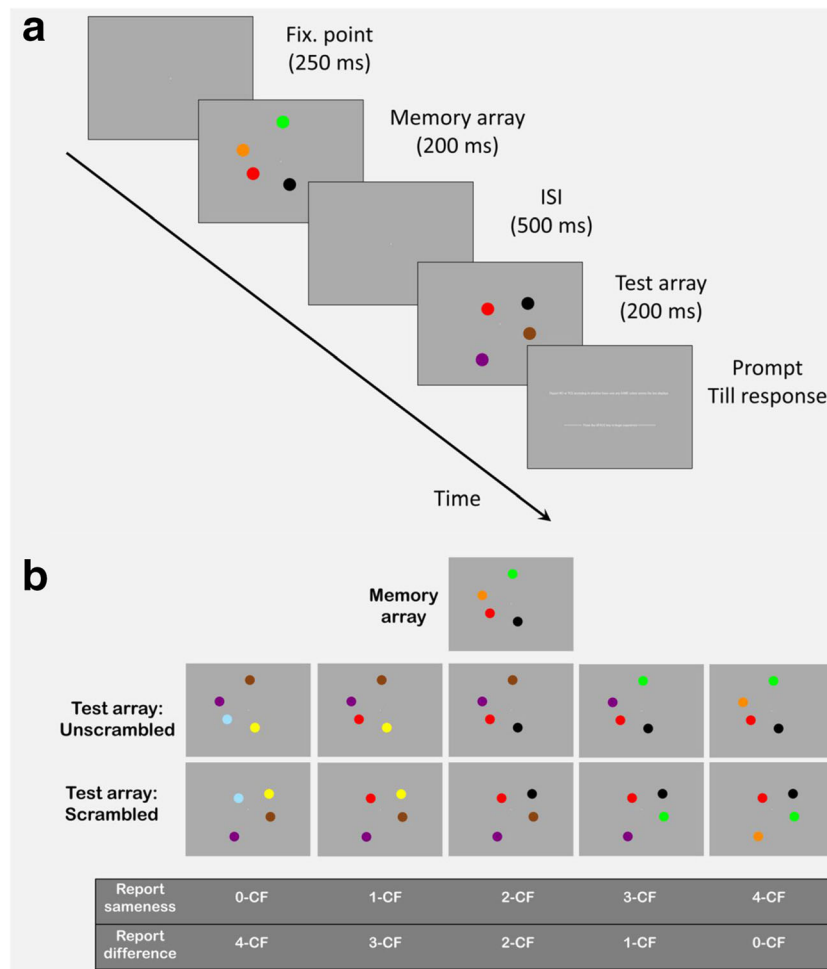


Fig. 1 Upper half of figure (A) gives a schematic diagram of a single trial in Experiment 1; lower-half of figure (B) shows example depictions of the different trial examples for each critical feature (CF) condition on

unscrambled and scrambled trials. Note how the CFs are different under the two different task instructions except for CF-2. See online version for rendering of the colour image

Figure 1B shows an example of the various CF conditions under the two task instructions.²

The instruction screen immediately followed the test display and reminded participants of the critical feature they were responding to. Participants indicated their decision using the left and right trigger-keys on a joystick. They were instructed to press the right key if any of the critical features was present and the left key if they were absent. The keys were appropriately labelled. Participants were instructed to emphasise accuracy not speed. They were informed that half the trials required a *yes* and half a *no* response. Immediate auditory

feedback was given on response. The participant’s response instigated the next trial after a 500-ms blank inter-trial interval.

Results

The mean hit rate (*pHIT*) was calculated for each of the critical feature conditions that required a positive response (CF-1 to CF-4), and the mean false-alarm rate (*pFA*) for the critical feature condition that required a negative response (CF-0). These are shown in Table 1.

A signal-detection analysis (Macmillan and Creelman, 2005) was performed on the hit and false-alarm data. This produced four *d*-prime scores (1-CF to 4-CF), respectively, for the unscrambled and scrambled conditions for the two tasks. This was calculated from the appropriate *pHIT* rate and the false-alarm rate (*pFA*) for the 0-CF trials. This yielded eight *d*-primes per participant for each task. These *d*-primes (*d'*) are shown in Fig. 2. Note that in this graph there are no data points for the 0-CF condition. This is because this

² To give further clarification to the CF manipulation, note that, under report difference instructions, the CF value is effectively a direct reflection of the number of colour changes that occur between memory and test. Under report-sameness instructions, however, the CF value is reversed with respect to this metric. To express the CF value in terms of the number of colour changes one simply needs to subtract the CF value from the number of overall items (i.e. 4). Thus, for example, under report-sameness instructions CF-1, expressed in number of changes is: 4 minus 1 = 3.

Table 1 Accuracy rates in Experiment 1 for Report-Difference (Diff.) and Report-Same (Same.) tasks. This is expressed as a false-alarm rate (pFA) for 0-CF, and as a hit rate (pHit) for 1-CF to 4-CF. Values are stated as probabilities as a proportion of 1. Standard deviations are given in brackets

Locations	Report task	0-CF (pFA)	1-CF (pHit)	2-CF (pHit)	3-CF (pHit)	4-CF (pHit)
Unscrambled	Diff.	0.06 (0.04)	0.66 (0.12)	0.84 (0.12)	0.91 (0.09)	0.9 (0.08)
	Same.	0.15 (0.07)	0.43 (0.14)	0.68 (0.11)	0.87 (0.1)	0.97 (0.04)
Scrambled	Diff.	0.16 (0.09)	0.6 (0.14)	0.78 (0.13)	0.88 (0.11)	0.91 (0.07)
	Same	0.19 (0.08)	0.44 (0.13)	0.67 (0.15)	0.86 (0.11)	0.96 (0.05)

CF critical feature

condition is incorporated in the d-prime scores for all conditions. All analyses are performed on these d-prime scores.

A calculation of the response bias criterion (C) was also computed for the unscrambled and scrambled conditions of the two tasks. On this measure positive values indicate conservative responding and negative, liberal responding (respectively a tendency to report ‘no’ and ‘yes’ under conditions of uncertainty). These values are given in Table 2.

Two separate two-way repeated-measures ANOVAs were performed to analyse the d-prime scores. These were done separately for the unscrambled and scrambled conditions. For both ANOVAs the factors were *Task* (report sameness, report difference) and nCF (1,2,3,4). The bias scores were also analysed by t-test to compare the difference between the two task conditions.

Unscrambled trials: For the d-prime analysis, both main effects were significant, *Task* ($F[1,23]=35.1$, $MS_{err}=0.775$, $p<.001$, $\eta_p^2=0.604$), and nCF ($F[3,57]=149.4$, $MS_{err}=0.128$, $p<.001$, $\eta_p^2=0.867$). The *Task* \times nCF interaction was also significant ($F[3,57]=30.7$, $MS_{err}=0.098$, $p<.001$, $\eta_p^2=0.572$).³

Post hoc comparisons explored the significant interaction. These and subsequent reported post hoc comparisons are all calculated using estimated-marginal means (EMMs) with the Holm-Bonferroni procedure used to correct for multiple comparisons. It was found that performance was significantly better for the report-difference task compared to the report-sameness task for 1- to 3-CF, ($t\geq 4.89$, $P_{Holm}<.001$). In the 4-CF condition performance did not differ between the two conditions ($t\leq 0.038$, $P_{Holm}=.999$).

The analysis of the bias data found no significant difference between the two task conditions $t(23)=1.64$, $p=.115$. However, one-tailed t -tests showed that for both tasks the C values were significantly different from zero ($t\geq 5.19$, $p<.001$), indicating a significant conservative bias.

Scrambled trials: The main effect of *Task* did not reach significance ($F[1,23]=3.47$, $MS_{err}=0.917$, $p=.075$). There

was a significant effect of nCF ($F[3,69]=169.01$, $MS_{err}=0.122$, $p<.001$, $\eta_p^2=0.88$). The *Task* \times nCF interaction was significant ($F[3,69]=18.09$, $MS_{err}=0.091$, $p<.001$, $\eta_p^2=0.44$).

Post hoc comparisons of the interaction showed a significant advantage for report difference for 1-CF and 2-CF ($t\geq 3.01$, $P_{Holm}\leq .038$). There was no significant difference between the report conditions for the 3-CF and 4-CF conditions ($t\leq 1.72$, $P_{Holm}\leq .414$).

The t-test analysis of the bias data again found no significant difference between the two task conditions $t(23)=0.596$, $p=.557$. Both values were significantly different from zero ($t\geq 2.68$, $p\leq 0.014$), indicating that both tasks had a significant conservative response bias.

Discussion

For the unscrambled trials there was a clear RDS in the 1- to 3-CF conditions but not the 4-CF condition. For the scrambled trials the RDS was evident in the 1- and 2-CF conditions but not those where more than half the items contained a critical feature (3-CF, 4-CF). In those conditions where the RDS was found it was clearly diminished compared to in the unscrambled trials. We can take an example in the 2-CF condition, where participants were viewing the same displays with the same number of CFs, across the two tasks for the unscrambled trials. Performance was more than 70% higher under report-difference instructions compared to report sameness. For the scrambled trials performance was only 34% higher for report difference than it was for report sameness.

Another notable aspect of the Experiment 1 results concerned the performance change with respect to the nCF variable. For the sameness task performance showed a linear improvement as nCF was increased. For the difference task this improvement was monotonic, but not linear. This was true for both the unscrambled and scrambled conditions. It was for this reason that an interaction was found for both the scrambled and unscrambled trials. The effect of the nCF variable indicates a further way in which difference and sameness judgements diverge in performance.

³ This particular interaction should be treated with some caution. Some participants were near or at ceiling in the 3-CF and 4-CF conditions, at least for the report difference trials. This may have inflated the extent of this interaction.

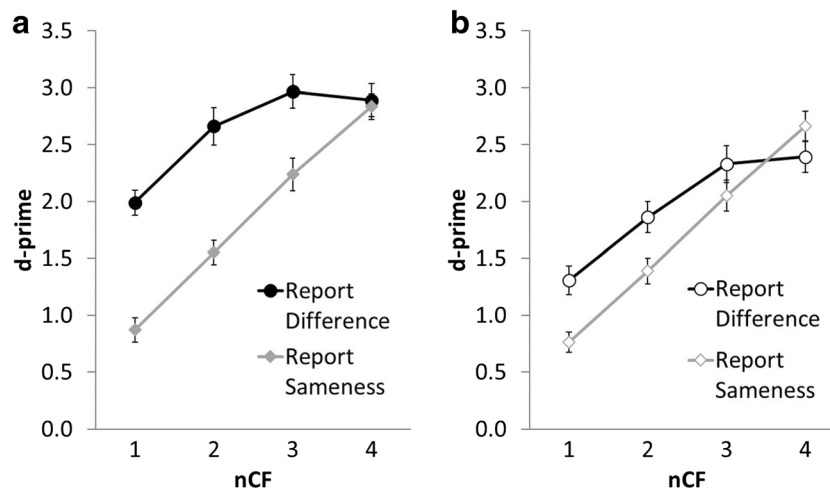


Fig. 2 Accuracy (d-prime) in Experiment 1 for the report difference (black lines) and report sameness (grey line) report conditions. Left figure (A) shows unscrambled condition. Right figure (B) shows scrambled condition. Error bars give ± 1 standard error

This interaction between task and *nCF* is something to which we will return later in the paper.

Importantly the predictions for Experiment 1 were supported. A substantial RDS was found in the (standard) unscrambled condition. This condition illustrates the basic RDS that has been reported in numerous behavioural studies that directly compare accuracy of sameness and difference (Farrell, 1985, Hyun et al., 2009). Importantly, the scrambled condition, where items were spatially displaced across memory and test arrays, also produced an RDS. The effect was smaller in magnitude than in the unscrambled condition, but it was still very much present.

The persistence of the RDS effect under scrambled conditions means that a local comparison mechanism (Hyun et al., 2009) is not, by itself, sufficient to account for the RDS. We postulated an additional mechanism that automatically compares global statistics of the two displays (Brady & Alvarez, 2011; Haberman & Whitney, 2011) that would also asymmetrically support difference over sameness judgements. The results from Experiment 1 are at least consistent with this possibility. However, this is somewhat indirect evidence. Experiment 2 further tested the possibility of an involvement of a global comparison mechanism in the RDS. It did this by giving a task in which comparison of global feature information would be unhelpful in making difference judgements.

Table 2 Mean criterion values (*C*) for Experiment 1. Standard deviations are given in brackets

Locations	Report difference	Report sameness
Unscrambled	0.30 (0.17)	0.21 (0.20)
Scrambled	0.11 (0.20)	0.15 (0.21)

Experiment 2

In Experiment 2 participants had to report about difference or sameness in colour-shape pairings, not just in colour alone as in the previous experiment. On trials where there was a difference between memory and test arrays this was always manifest in terms of changes to the colour-shape pairings. The memory and test displays always contained the same features, just in different combinations on trials involving a change. What this meant is that changes to the display objects never affected the overall global statistics of the display with regards to the features present.

As in Experiment 1 there was an unscrambled and a scrambled condition. In the unscrambled condition local comparisons can still be used to detect differences in the feature conjunctions. Consequently, we should expect a similar RDS to that found in Experiment 1.

The crucial question concerns performance in the scrambled condition. Here local comparisons cannot be used, and, because of the conjunction nature of the stimuli global, statistical comparisons would also be unrevealing. Because of this we predicted that the RDS would be entirely abolished in the scrambled condition. We varied *nCF*, as we did in the previous experiment. Note that it is logically impossible to have a 1-CF condition for the difference condition, since a swap requires a minimum of two items to be affected. For the same reason it is logically impossible to have a 3-CF condition for the sameness condition. Therefore, we presented only the three *nCF* conditions that were logically possible under both task instructions, 0-CF, 2-CF and 4-CF. The arguably most important condition from an analysis perspective was again the 2-CF condition. An RDS effect was found in Experiment 1 for both scrambled and unscrambled conditions. The question was whether this would be retained or abolished in the scrambled condition of Experiment 2.

Methods

Participants

There were 24 participants (18 female). This number was chosen to match the sample size for Experiment 1. Participants were recruited using the same criteria as Experiment 1. None had taken part in Experiment 1.

Stimuli

The stimuli were displayed in the same way and using the same computer equipment as per Experiment 1. The same colour set was used. The colours were in the form of one of four shapes (each approximately 3.8° in height or width): rectangle, oval, diagonal cross and triangle.

Procedure

The procedure was the same as for Experiment 1. The critical difference between the experiments was in the viewed stimuli. Participants were presented with a memory set of four distinct shapes, each in a separate colour. The colours were all randomly drawn without replacement from the set of 12 colours.

Each trial began with the presentation of a fixation point for 250 ms; this remained present until the response screen. The memory display was shown for 200 ms then replaced after a 500-ms blank ISI by the test display, which was shown for 200 ms. The test display also contained the four distinct shapes. On unscrambled trials these shapes retained the same positions in the test display as in the memory display. On scrambled trials the shapes were presented in new locations, determined with the same constraints described for Experiment 1. Participants did both the report-difference task and the report-sameness task in counterbalanced order. Participants responded using trigger keys on a joystick. This resulted in immediate auditory feedback on the response and instigated the next trial after a 500-ms blank inter-trial interval. As in Experiment 1, practice trials were given for each task before embarking on the experimental trials.

For both the report-difference and report-sameness tasks half the trials were 0-CF, the other half were 2- or 4-CF each with equal frequency. For the difference task, on 0-CF trials all four of the shapes kept the same colours as in the memory display. On 2-CF trials two of the shapes, selected at random swapped their colours. On 4-CF trials each shape swapped its colour with one of the other shapes. The particular colour-shape pairings that were subject to change were randomly determined on each trial. For the sameness task, the 0-CF, 2-CF and, 4-CF trials were identical to, respectively, the 4-CF, 2-CF and, 0-CF trials of the difference task. It is important to note that irrespective of the task, and irrespective of *nCF*, the memory and test displays always contained the same colour

and shape features, just in different combinations in some trials. Figure 3 depicts an example schematic of a single trial (A) and shows example displays in the given CFs for the scrambled and unscrambled conditions (B).

Results

The false alarms from the 0-CF condition and the hits from the 2-CF and 4-CF conditions are shown in Table 3. The analysis took the same form as described for Experiment 1, by calculating *d*-primes from the above described hit and false-alarm data. The resulting *d*-primes are shown in Fig. 4. Note again that only the 2- and 4-CF conditions are shown. The 0-CF condition is used in calculating the *d*-primes for CF conditions. Response bias criterion values (*C*) were also calculated. These are given in Table 4.

The *d*-primes for unscrambled and scrambled trials were analysed separately each using two-way repeated-measures ANOVAs. Each had two factors: *Task* (report-difference, report-sameness); *nCF* (2-CF, 4-CF). The response bias values *C* were analysed using *t*-tests, as described for Experiment 1.

Unscrambled There was a main effect of both factors: *Task* ($F[1,23]=8.65$, $MS_{err}=0.319$, $p=.007$, $\eta_p^2=.0.273$); *nCF* ($F[1,23]=195.76$, $MS_{err}=0.149$, $p<.001$, $\eta_p^2=.0.895$). The *Task* \times *nCF* interaction was also significant ($F[1,23]=18.17$, $MS_{err}=0.130$, $p<.001$, $\eta_p^2=.0.441$). Post hoc comparisons of the interaction showed a report-difference advantage for the 2-CF ($t=4.77$, $P_{Holm}<.001$), but not the 4-CF ($t=0.18$, $P_{Holm}=.856$) condition.

Analysis of the bias (*C*), found a significantly greater conservative bias for the difference task than for the sameness task, $t(23)=4.06$, $p<.001$. One-sample *t*-tests compared against zero showed that there was a conservative bias for the report-difference task, $t(23)=8.12$, $p<.001$, but no bias on the report-sameness task, $t(23)=1.58$, $p=.129$.

Scrambled There were main effects of *Task* ($F[1,23]=6.46$, $MS_{err}=0.2553$, $p=.018$, $\eta_p^2=.0.219$) and *nCF* ($F[1,23]=68.32$, $MS_{err}=0.1659$, $p<.001$, $\eta_p^2=.0.748$). The main effect of task is because performance is slightly worse in the report-difference task than in the report-sameness task, i.e. there is a slight inversion of the RDS effect. The *Task* \times *nCF* interaction was non-significant ($F[1,23]=1.68$, $MS_{err}=0.0969$, $p=0.207$).

Analysis of the bias scores found a significantly greater conservative shift in the report-difference compared to the report-sameness task, $t(23)=2.76$, $p=.011$. One-sample *t*-tests found that while the report-difference task produced a conservative bias, $t(23)=2.58$, $p<.017$, the sameness task had no bias, $t(23)=1.47$, $p=.15$.

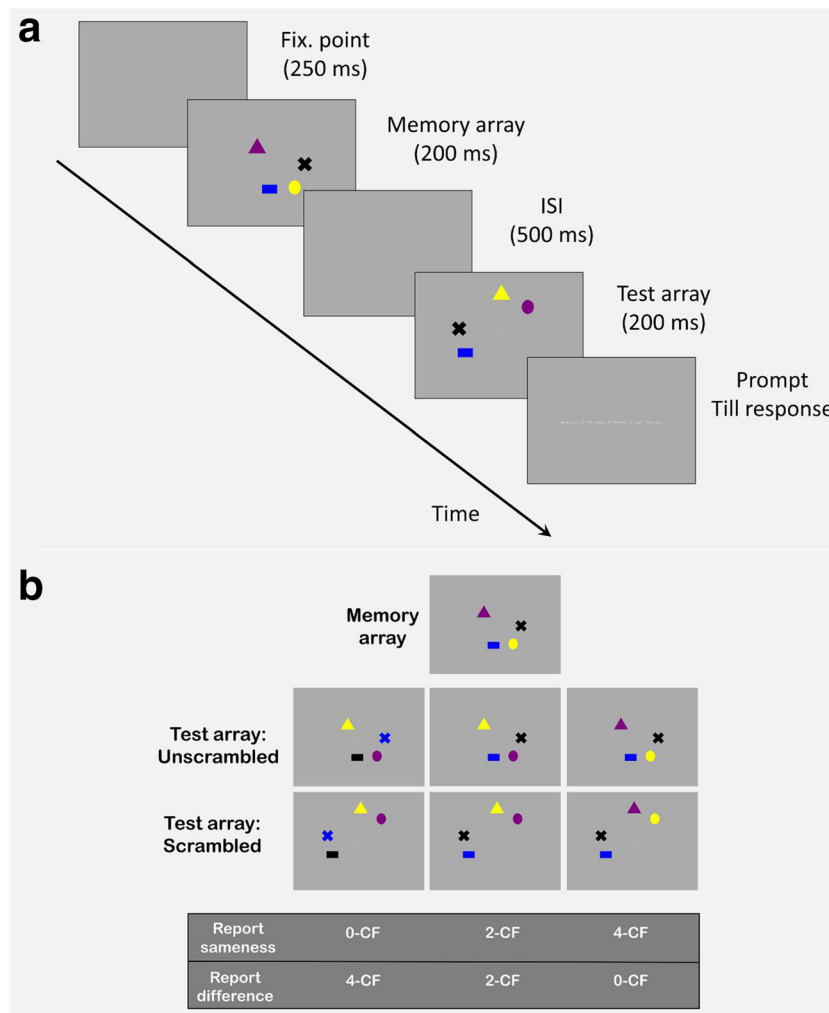


Fig. 3 Upper half of figure (A) gives a schematic diagram of a trial in Experiment 1; lower half of figure (B) gives example depictions of the different trial examples for each critical feature (CF) condition on

unscrambled and scrambled trials. Note how the CFs are different under the two different task instructions except for CF-2. See online version for rendering of the colour image

Comparison of Experiments 1 and 2 A further analysis was done on the d-prime scores to compare the extent of the RDS across comparable conditions of the two experiments. The RDS scores were calculated by subtracting the equivalent report-

sameness d-prime score from each report-difference d-prime (those shown in Figs. 2 and 4). These values ($\Delta d'$) are shown in Table 5. Note that these difference values can be positive or negative. A negative value indicates better performance in the equivalent report-sameness condition than in report difference.

Table 3 Accuracy rates in Experiment 2 for Report-Difference (Diff.) and Report-Same (Same.) tasks. This is expressed as a false-alarm rate (pFA) for 0-CF, and as a hit rate for 2-CF and 4-CF. Values are stated as probabilities as a proportion of 1. Standard deviations are given in brackets

Locations	Report task	0-CF (pFA)	2-CF (pHIT)	4-CF (pHIT)
Unscrambled	Diff.	0.1 (0.07)	0.58 (0.12)	0.82 (0.14)
	Same.	0.23 (0.12)	0.55 (0.15)	0.9 (0.12)
Scrambled	Diff.	0.31 (0.14)	0.49 (0.14)	0.71 (0.11)
	Same	0.35 (0.11)	0.6 (0.15)	0.82 (0.15)

CF critical feature

Two-way mixed ANOVAs with *Experiment* (Exp. 1, Exp. 2) as an independent-measures factor and CF (2-CF, 4-CF) as a repeated-measures factor were conducted on these RDS scores. These ANOVAs were performed separately for the unscrambled and scrambled trials.

For the unscrambled condition there was no main effect of *Experiment* ($F[1,46]=1.70, MS_{err}=0.839, p=.199$), a main effect of *nCF* ($F[1,46]=69.41, MS_{err}=0.245, p<.001, \eta_p^2=0.601$), and an *Experiment* \times *nCF* interaction ($F[1,46]=4.46, MS_{err}=0.245, p=.040, \eta_p^2=0.088$). The post hoc comparisons between Experiment 1 and Experiment 2 just escaped significance for 2-CF ($t=2.151, P_{Holm}=.07$), and was clearly non-significant for the 4-CF condition ($t=0.143, P_{Holm}=.887$). The

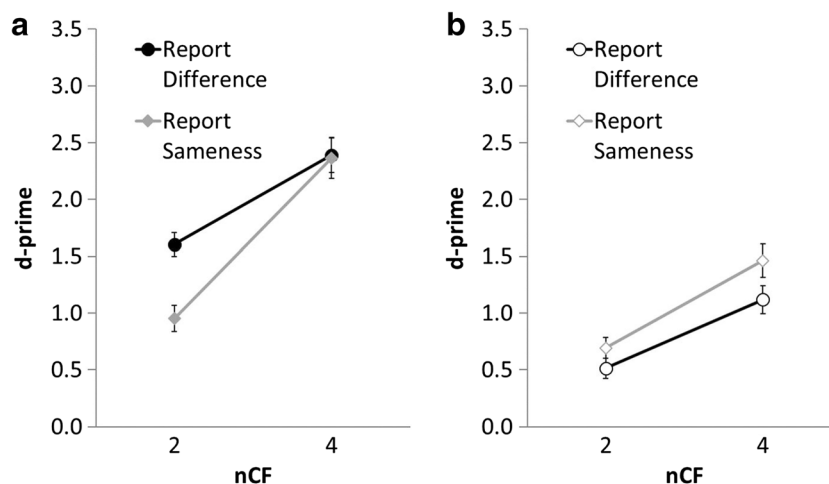


Fig. 4 Accuracy (d-prime) in Experiment 2 for the report difference (black lines) and report sameness (grey line) report conditions. Left figure (A) shows unscrambled condition. Right figure (B) shows scrambled condition. Error bars give ± 1 standard error

interaction occurs because the RDS scores are marginally higher in Experiment 1 than in Experiment 2 for the 2-CF condition, but not the 4-CF condition.

For the scrambled condition there were significant main effects of *Experiment* ($F[1,46]=4.65$, $MS_{en}=0.0.682$, $p=.036$, $\eta_p^2=.092$), and *nCF* ($F[1,46]=28.3$, $MS_{err}=0.175$, $p<.001$, $\eta_p^2=0.381$). There was also an *Experiment* \times *nCF* interaction ($F[1,46]=11.50$, $MS_{err}=0.175$, $p=.001$, $\eta_p^2=0.2$). Post hoc comparisons showed that this was because RDS in the scrambled condition was significantly larger in Experiment 1 than in Experiment 2 for the 2-CF condition ($t=3.456$, $P_{Holm}=.004$), but not 4-CF ($t=0.390$, $P_{Holm}=0.999$). In essence, the interaction reflects the fact that, of the statistical comparisons made across the experiments, the 2-CF condition of Experiment 1 is the only scrambled condition in which RDS is found.

Discussion of Experiment 2

For Experiment 2, as with Experiment 1, a clear RDS was found for the unscrambled conditions. This particular result can again be explained by a local comparison mechanism. Under these conditions the spatial invariance of the displays means that it is unnecessary to notice the particular colour-shape pairings to detect the change. Instead, it is strategically possible to do either report task by just comparing the colour features at the unchanging memory and test locations. That said, overall performance even in the unscrambled condition was poorer in comparison

with Experiment 1. This fact may indicate that participants did not generally adopt this simple feature comparison strategy on unscrambled trials. Since they could not predict on which trials this strategy could be usefully applied, this is not surprising.

We can reasonably assume that the RDS is supported, as with Experiment 1, by transients that accompany local feature mismatches between the two displays (Hyun et al., 2009). The interesting question concerned the scrambled condition. Here neither a putative local nor global comparison mechanism would be effective in supporting an RDS on all trials. Consistent with this, the RDS effect was completely abolished for the scrambled trials of Experiment 2. The decisive comparison was with the 2-CF condition that produced a clear RDS in Experiment 1. Comparison across the experiments found the RDS obtained under the scrambled 2-CF condition of Experiment 1 was reversed in the scrambled 2-CF condition of Experiment 2. The superiority for reporting difference became, with both a conjunction task and with scrambling, a (weak) superiority for reporting sameness.

General discussion

The two experiments revealed factors on which the RDS effect depends. As expected, the relative locations of the

Table 4 Mean criterion values (*C*) for Experiment 2. Standard deviations are given in brackets

Locations	Report difference	Report sameness
Unscrambled	0.43 (0.26)	0.09 (0.29)
Scrambled	0.14 (0.27)	-0.08 (0.27)

Table 5 Report-difference superiority values ($\Delta d'$) for comparable conditions of Experiments 1 and 2. Note that negative scores indicate a report-sameness advantage. Standard deviations are given in brackets

	Unscrambled		Scrambled	
	2-CF	4-CF	2-CF	4-CF
Experiment 1	1.11 (0.80)	0.06 (0.79)	0.47 (0.67)	-0.27 (0.75)
Experiment 2	0.65 (0.66)	0.03 (0.68)	-0.18 (0.50)	-0.35 (0.68)

CF critical feature

compared items between memory and test was crucial. RDS was most in evidence when compared items maintained their spatial locations in the memory and test display. Scrambling markedly reduced the RDS effect in Experiment 1, and in Experiment 2 abolished it. Secondly, RDS was more in evidence where the task required comparison of simple features (Exp. 1), than conjunctions of features (Exp. 2). Finally, the report difference advantage was most evident when there were relatively few critical features (1 or 2), compared to when the majority of items contained a critical feature (3 or 4).

Local and global comparison processes

Both experiments found that location invariance was an important factor driving RDS. Hyun et al. (2009) did not do such a manipulation. However, our finding is consistent with their interpretation of the RDS as one supported by a reflexive local comparison mechanism. This mechanism produces an RDS because the transient signals facilitate the detection of a difference, but are rather less diagnostic about the presence or absence of a sameness.

These transient signals associated with local change are possibly sufficient for the participant to report that a change has occurred. Research has shown that participants can sense the presence of, or broadly locate a change in, the test display before they can identify what it is (Busch, Dürschmid & Herrmann, 2010; Rensink, 2004; Watanabe, 2003). Local comparison processes of the kind demonstrated in our experiments are likely to underlie such effects discovered in earlier work on change detection.

Importantly, a local comparison mechanism can only support difference judgements under conditions where the spatial structure of compared items is maintained. Spatial scrambling across memory and test would remove any advantage that such a local comparison mechanism supports. The fact that this manipulation did not obliterate the RDS suggests that this local mechanism did not solely account for it. It suggests one or more additional factors are at play.

We suggest that where possible difference judgements are also supported by a mechanism involving a global statistical comparison of the memory and test displays. As was mentioned earlier, other work has indicated that the brain tends to generate statistical means of held VSTM content and that such statistical averaging can be used to make global comparisons of feature content between temporally separated displays (Brady, & Alvarez, 2011; Chong & Treisman, 2005; Dubé, Zhou, Kahana & Sekuler, 2014; Haberman & Whitney, 2011; Maule, Witzel & Franklin, 2014; Oriet & Brand, 2013).

We think that the visual system routinely compares these statistical means extracted from the contents of VSTM and from current viewed input and that the information derived from this can sometimes inform our perceptual judgements.

The information about differences from global statistical comparisons is likely to be noisy, and the shift in mean hue might only be modest even when all display items are different between memory and test. Where the comparison revealed a substantial shift in the mean hue across memory and test, this would serve as an unequivocal alert to the presence of a new colour. Importantly, however, such statistical comparisons would be less informative of whether any items had retained their colour, as required by the sameness tasks.

The only case where such signals would assist would be in the 4-CF condition, where – under ideal conditions – a global comparison would reveal the absence of a difference. Consequently, global comparisons of stimuli, like local ones, would under most circumstances tend to asymmetrically support difference judgements. A global comparison mechanism would operate irrespective of whether the spatial structure of the individual display items was retained or not. However, global comparison information would be uninformative to difference decisions when these differences consisted of manipulations in which the feature content of the displays was always the same. This, we argue, is why the RDS is completely eliminated for the conjunction stimuli in the scrambled conditions of Experiment 2.

Number of critical features

Results shows that the number of critical features was the main variable determining performance. This is unsurprising in itself: the more different (or same) the displays were, the easier it was to detect that there was a difference (or sameness) present. More interesting is the fact that the CF variable interacted with type of report. The RDS effect was most strongly in evidence on trials with one or two CFs. With three or four CFs the RDS was diminished or abolished.

As CF increased in amount the disparity between detecting sameness and difference; when all items contain a CF then the disparity is no longer found. This interaction between the tasks cannot be dismissed as a performance ceiling artefact. It occurs in both unscrambled and scrambled conditions of Experiment 1 and in the unscrambled condition of Experiment 2. Thus, the interaction occurred across a range of performance values.

We had not specifically predicted the RDS and nCF interaction.⁴ However, the finding, we think, is easily explained. To recap, we have been essentially assuming that the

⁴ Hyun et al. (2009) also found an interaction between these variables. They took reaction time, rather than accuracy, as the primary measure. They did not find the same pattern that we report with respect to the critical feature variable. However, they do not give enough detail in their paper about if and how accuracy varied across the conditions for us to make any comparison with their results. The demands of the task are also rather different in our task because the items are spatially scrambled on half the trials so it isn't clear how comparable data would be expected to be in any case.

asymmetry between report-difference and report-sameness performance is associated with the efficiency with which a designated CF can be identified. Any divergence in efficiency will be most marked for cases where there is just a single to-be-detected CF involving just a single memory and test stimulus. Correspondingly this divergence would be least marked when all display items are involved, because the presence of a CF should be efficiently identified for both tasks. Given this, it is unsurprising that the performance disparity for report-difference and report-sameness was highest in the 1-CF condition and declined to nothing in the 4-CF condition. Note that though comparison processes would be most effective at identifying critical features in the 4-CF condition this would not necessarily mean that a performance ceiling would be reached. This is because task performance is not just limited by the effectiveness with which perceptual comparison can be made between memory-held items and the test display, but also by other factors, such as VSTM encoding and storage; if items are not encoded or retained in memory then they cannot be compared and consequently performance will be constrained.

It should be noted that most perceptual comparison experiments usually only have two types of CF conditions, one in which there is a single critical feature present (1-CF) and one in which it is absent (0-CF) (e.g. Burmester & Wallis, 2011; Davis & Leow, 2005; Delvenne & Dent, 2008; Luck & Vogel, 1997; Pashler, 1988; Pilling & Barrett, 2016; Sligte, Sholte & Lamme, 2008; Wilson & Goddard, 2011; but cf. Hyun et al., 2009; Rhodes, Cowan, Hardman, & Logie, 2018). What our results show is that these conditions, in which there is a single, to-be-detected CF, are actually those where the RDS is most likely to be in evidence.

Spatial organisation and perceptual comparisons

The difference and sameness tasks both showed poorer performance when the displays were scrambled compared to when the spatial structure was retained. Our findings give further evidence of the spatial nature of VSTM representations (Aginsky & Tarr, 2000; Golomb, Kupitz & Thiemann, 2014; Huang, 2010; Jiang, Olsen & Chun, 2000; Simons, 1996). Previous evidence has indicated that visual features and objects are automatically bound to spatial locations, leading to reduced performance on comparison tasks when the spatial organisation of a display is altered (Jiang et al., 2000; Olivers and Schreij, 2014; Perzov & Husain, 2014; Rajsic & Wilson, 2014; Treisman & Zhang, 2006).

There is evidence that this tendency to make local comparisons across all items is not entirely obligatory and is somewhat dependent on task conditions and individual strategy (Bodoroglu & Shah, 2009; Udale, Farrell, & Kent, 2018). In our paradigm, location was nominally task-irrelevant; however, the experimental contingencies

of our tasks made it strategically useful to take account of location information in making comparisons: Half the trials were always unscrambled ones, trials where the spatial organisation of the display was unaltered. Furthermore, the scrambled trials were always ones in which test items were placed at locations which were unoccupied ones in the memory display. This meant that the two types of trial were easily distinguishable purely based on the spatial structure of the locations.

Given this it would make strategic sense to compare across locations to detect a difference or sameness on trials that were identified as unscrambled. It is possible that if a much smaller proportion of unscrambled trials was given in the experiment then this would limit the value of such a local comparison strategy. Consequently, such a manipulation may result in a smaller same-location advantage compared to the levels we found. The extent to which this is a strategic effect rather than a purely automatic one remains an open question. However, irrespective of this, what is clear from our data is that scrambling affected the ability to report difference much more than it did sameness. As stated earlier we attribute this to the partial reliance of difference judgements on local comparison processes (Hyun et al., 2009). We consider that such a mechanism is no longer able to diagnose the presence of a difference when items are scrambled.

It would be interesting to know exactly how tolerant such automatic comparisons are to spatial displacement. We compared only two spatial conditions, unscrambled and scrambled. In the unscrambled condition the memory items retained both their absolute and their relative spatial positions on screen. It is an open question whether the former is important or whether similar report-difference advantages would be found for unscrambled trials if only the relative positions were ever maintained across memory and test. Some work on VSTM has found a performance advantage in tasks where the relative spatial structure of items is maintained though the absolute positions are changed when compared against conditions in which there is no correspondence (Hollingworth, 2007; Hollingworth & Rasmussen, 2010). Other work has shown that allocentric and egocentric spatial information can be independently encoded in VSTM (Aagten-Murphey & Bays, 2019). It is not clear what this work tells us about how the underlying comparison processes themselves operate with regard to such spatial displacements. It is possible that the reflexive local comparisons between VSTM and current vision (Hyun et al., 2009) can only indicate feature changes when compared items retain the same egocentric positions. Future research could address this question by looking at the extent to which the same-locations advantage for report-difference is maintained over report-sameness instruction, when comparing displays that retain the same spatial position

or only the same relative structure between memory and test.

Memory search for sameness and difference

Our manipulation of the spatial organisation of the displays brings up another issue. In most VSTM studies of perceptual comparison the participant is only required to compare items in the test display with the same locations in the memory display (Hyun et al. 2009; Luck & Vogel, 1997; Pashler, 1988; Pilling & Gellatly, 2013; Wilson & Goddard, 2011). This was true in our unscrambled condition. In our unscrambled condition, however, spatial location was no longer a useful basis for comparison for either task. Consequently, this may require participants to actively search other VSTM locations in order to determine the presence or absence of the required CF. A body of work has investigated memory search processes in VSTM (Gilchrist & Cowen, 2014; Kong & Fougny, 2019; Kuo, Rao, Lepsien & Nobre, 2009; Magen, 2017). However, these studies typically only present a single test item and have never tried to tease apart the issue of search for difference versus sameness.

As we stated previously, in the case of Experiment 1 we think that difference judgements are supported by comparisons of global statistics where stimulus displays are scrambled. However, statistical information alone will always be noisy and imprecise in indicating the presence or absence of a difference. There may be occasions on scrambled trials when this is insufficient to inform a decision of whether there is a difference or not. Under such circumstances it may be that participants have to resort to an active search of memory to identify whether any of the items in the test display are new.

Though it has not yet been specifically investigated, there are reasons to assume that such a memory search for difference CFs would be less efficient than the equivalent search for a sameness CF. This is because for sameness the participant is simply looking for a match with the same type of item. To take an example, if search begins with a *yellow* test item the aim under sameness instructions is to locate in memory an item that is also *yellow*. This could perhaps be done by serially comparing the items in VSTM with the test item. If a match is found the search is terminated and a decision is made that a sameness is present. If it is not then another test item may be selected. If none of the searches yield a match then a decision is made that a sameness is absent. It may be that internal attention can guide the search in memory towards the matching same colour. Certainly, it is known from retro-cueing studies that internal attention can be guided towards items of a specified colour in VSTM (Griffin & Nobre, 2003; Heuer & Schubö, 2016; Li & Saiki, 2015). Memory search for a feature sameness might benefit from such a process.

For the difference task things are more complicated when it comes to active search. In identifying a difference from

memory search one is not looking for a match, but rather for the *absence* of a match with the test item. This is an inherently harder thing to search for. It might therefore be impossible to guide attention towards any difference in the same way as for a sameness, because a difference is not defined by the presence of any specific feature in memory. Consequently, we might expect that memory search alone is poorer at identifying differences between the test display and memory-held items than it is at identifying sameness with those items.

We recognise that our suggestions about memory search with respect to identification of sameness and difference are tentative. We cannot determine the extent to which memory search played a role on our respective tasks. We can, however, suggest is that whatever extent memory search processes contributed on our experiments, they would tend to favour the sameness task and therefore tend to diminish rather than augment the RDS. We might also speculate that the finding of a small but significant *sameness advantage* in the scrambled condition of Experiment 2 might be attributable to such search processes. It is possible that when task conditions remove the advantages of local and global comparisons, as they do in that condition of Experiment 2, then the search advantage for sameness is expressed in the small performance advantage for sameness that we found.

Response bias in reporting sameness and difference

We should also comment on the pattern of response biases we found on our tasks. Our tasks were designed to minimise response bias, by having equal frequencies of trials requiring a *yes* and *no* response, and by giving immediate feedback. Despite this, Experiment 1 displayed a general conservative tendency in responding across all tasks and conditions. That is, participants tended to err towards reporting that no critical feature present when uncertain. This was true for both tasks, under unscrambled and scrambled conditions. For Experiment 2 the same bias was found, but here only for the report-difference task. The general tendency towards conservative responding might arise because participants tend to give a negative response on trials where the VSTM memory item representations of the memory display are either unavailable or are of insufficient quality to support a judgement about the critical feature. Such circumstances might occur on trials where the memory display was not adequately attended to, encoded or retained.

Why then is this conservative bias found for all cases in Experiment 1 not found for Experiment 2 under sameness instructions? There must be some aspect of the sameness task in Experiment 2 that induces a liberal criterion shift, mitigating against the conservative response tendencies found in the Experiment 1 version of this task. One candidate possibility is the fact that in Experiment 2 sameness is present on every trial, in the sense that the same individual features are always

seen repeated across memory and test. The only difference is ever in the correspondence of these features. This repeated presence of sameness may induce participants to adopt a more liberal responding strategy under report sameness instruction when uncertain.

The fact that response bias did not significantly differ across the difference and sameness tasks in Experiment 1 is reassuring. However, the fact that this does diverge in Experiment 2 does not – in our view – mitigate against the interpretation of the *d*-primes that we make. Sensitivity and response bias are generally considered to be completely orthogonal aspects of responding, and are influenced by different aspects of the task (Macmillan & Creelman, 2004; Stanislaw & Todorov, 1999). If the instructions or conditions of Experiment 2 were altered in some way to induce comparable levels of bias across the two tasks, we are confident that the pattern of *d*-primes that we obtained here would still replicate. However further research would be required to investigate this.

Conclusion

Hyun and colleagues (Hyun et al., 2009) concluded in their paper that they would be unsurprised if future studies led to refinements and revisions of their proposal regarding the nature of comparison processes between VSTM representations and current vision. Based on our experiments we conclude these authors were basically correct in their original stated interpretation of the RDS. It mainly seems to be a consequence of automatic local comparisons that either guides a later confirmatory comparison process or provides direct information that contributes to the decision process. Sameness lacks any such bottom-up directed cues supporting the decision, and therefore performance in reporting sameness tends to be less accurate.

Our results, however, suggest at least one additional factor at play. Specifically, we suspect that comparison of global averaging statistics (Haberman & Whitney, 2011) contributes to decision processes, particularly when judging difference, and thus also underlies RDS. We also speculate that – under some conditions – sameness and difference judgements may be differently affected due to the extent to which memory search processes support the identification of their corresponding critical features.

Ultimately our experiments show that RDS will be most evident under the conditions that were given in Hyun and colleagues' original study (Hyun et al., 2009). That is to say, conditions in which participants must compare only the single feature of an object to detect a difference and sameness, and under conditions where those objects are spatially aligned in the two compared arrays.

Open Practices Statements The data, analysis and programme source code for all experiments are available on the Open Science Framework (URL link: <https://osf.io/bep8q/>). None of the experiments were preregistered.

Author note We would like to thank Liqiang Huang and one anonymous reviewer for comments on an earlier version of this manuscript.

References

- Aginsky, V., & Tarr, M. J. (2000). How are different properties of a scene encoded in visual memory?. *Visual Cognition*, 7(1-3), 147-162.
- Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, 12(2), 157-162.
- Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological science*, 22(3), 384-392.
- Boduroglu, A., & Shah, P. (2009). Effects of spatial configurations on visual change detection: An account of bias changes. *Memory & cognition*, 37(8), 1120-1131.
- Busch, N. A., Dürschmid, S., & Herrmann, C. S. (2010). ERP effects of change localization, change identification, and change blindness. *Neuroreport*, 21(5), 371-375.
- Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, 43(4), 393-404.
- Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision research*, 45(7), 891-900.
- Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24 (1), 87.
- Davis, G., & Leow, M. C. (2005). Blindness for unchanging targets in the absence of motion filtering: A response to Theeuwes (2004). *Psychological Science*, 16(1), 80-82.
- Delvenne, J. F., & Dent, K. (2008). Distinctive shapes benefit short-term memory for color associations, but not for color. *Perception & psychophysics*, 70(6), 1024-1031.
- Dubé, C., Zhou, F., Kahana, M. J., & Sekuler, R. (2014). Similarity-based distortion of visual short-term memory is due to perceptual averaging. *Vision research*, 96, 8-16.
- Egeth, H. E. (1966). Parallel versus serial processes in multidimensional stimulus discrimination. *Perception & Psychophysics*, 1(4), 245-252.
- Farrell, B. (1985). " Same"- " different" judgments: A review of current controversies in perceptual comparisons. *Psychological Bulletin*, 98(3), 419-456.
- Gilchrist, A. L., & Cowan, N. (2014). A two-stage search of visual working memory: Investigating speed in the change-detection paradigm. *Attention, Perception, & Psychophysics*, 76(7), 2031-2050.
- Golomb, J. D., Kupitz, C. N., & Thiemann, C. T. (2014). The influence of object location on identity: A "spatial congruency bias". *Journal of Experimental Psychology: General*, 143(6), 2262.
- Griffin, I. C., & Nobre, A. C. (2003). Orienting attention to locations in internal representations. *Journal of Cognitive Neuroscience*, 15(8), 1176-1194.
- Haberman, J., & Whitney, D. (2011). Efficient summary statistical representation when change localization fails. *Psychonomic Bulletin & Review*, 18(5), 855.

- Heuer, A., & Schubö, A. (2016). Feature-based and spatial attentional selection in visual working memory. *Memory & cognition*, 44(4), 621–632.
- Hollingworth, A. (2006). Visual memory for natural scenes: Evidence from change detection and visual search. *Visual cognition*, 14(4–8), 781–807.
- Hollingworth, A. (2007). Object-position binding in visual memory for natural scenes and object arrays. *Journal of Experimental Psychology: Human Perception and Performance*, 33(1), 31.
- Hollingworth, A., & Rasmussen, I. P. (2010). Binding objects to locations: The relationship between object files and visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, 36(3), 543–564.
- Huang, L. (2010). Characterizing the nature of visual conscious access: The distinction between features and locations. *Journal of Vision*, 10(10), 24–24.
- Huang, L. (2015). Statistical properties demand as much attention as object features. *PLoS one*, 10(8), e0131191.
- Hyun, J. S., Woodman, G. F., Vogel, E. K., Hollingworth, A., & Luck, S. J. (2009). The comparison of visual working memory representations with perceptual inputs. *Journal of Experimental Psychology: Human Perception and Performance*, 35(4), 1140.
- Jackson-Nielsen, M., Cohen, M. A., & Pitts, M. A. (2017). Perception of ensemble statistics requires attention. *Consciousness and Cognition*, 48, 149–160.
- Jiang, Y., Olson, I. R., & Chun, M. M. (2000). Organization of visual short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(3), 683.
- Kong, G., & Fougny, D. (2019). Visual search within working memory. *Journal of Experimental Psychology: General*. In press.
- Kuo, B. C., Rao, A., Lepsien, J., & Nobre, A. C. (2009). Searching for targets within the spatial layout of visual short-term memory. *Journal of Neuroscience*, 29(25), 8032–8038.
- Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, 390(6657), 279.
- Luck, S. J. (2006). Visual short-term memory. In S. J. Luck & A. Hollingworth (Ed.) *Visual Memory*, Oxford University Press, Oxford, UK.
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: A User's Guide*. 2nd. Lawrence Erlbaum Associates. New York.
- Magen, H. (2017). The role of central attention in retrieval from visual short-term memory. *Psychonomic bulletin & review*, 24(2), 423–430.
- Mangini, M., Villano, M., & Crowell, C. (2010). Visual short term memory for one item. *Journal of Vision*, 10(7), 620.
- Markman, A. B. & Gentner, D. (2000). Structure mapping in the comparison process. *American Journal of Psychology*, 113(4), 501–538.
- Maule, J., Witzel, C., & Franklin, A. (2014). Getting the gist of multiple hues: Metric and categorical effects on ensemble perception of hue. *JOSA A*, 31(4), A93–A102.
- Olivers, C. N., & Schreij, D. (2014). Visual memory performance for color depends on spatiotemporal context. *Attention, Perception, & Psychophysics*, 76(7), 1873–1884.
- Oriet, C., & Brand, J. (2013). Size averaging of irrelevant stimuli cannot be prevented. *Vision Research*, 79, 8–16.
- Pashler, H. (1988). Familiarity and visual change detection. *Perception & psychophysics*, 44(4), 369–378.
- Pertsov, Y., & Husain, M. (2014). The privileged role of location in visual working memory. *Attention, Perception, & Psychophysics*, 76(7), 1914–1924.
- Phillips, W. A. (1974). On the distinction between sensory storage and short-term visual memory. *Perception & Psychophysics*, 16(2), 283–290.
- Pilling, M., & Barrett, D. J. (2016). Dimension-based attention in visual short-term memory. *Memory & Cognition*, 44(5), 740–749.
- Pilling, M., & Barrett, D. J. (2018). Change perception and change interference within and across feature dimensions. *Acta psychologica*, 188, 84–96.
- Pilling, M., & Gellatly, A. (2013). Task probability and report of feature information: What you know about what you ‘see’ depends on what you expect to need. *Acta psychologica*, 143(3), 261–268.
- Rajicic, J., & Wilson, D. E. (2014). Asymmetrical access to color and location in visual working memory. *Attention, Perception, & Psychophysics*, 76(7), 1902–1913.
- Rensink, R. A. (2004). Visual sensing without seeing. *Psychological Science*, 15(1), 27–32.
- Rhodes, S., Cowan, N., Hardman, K. O., & Logie, R. H. (2018). Informed guessing in change detection. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(7), 1023.
- Scott-Brown, K. C., Baker, M. R., & Orbach, H. S. (2000). Comparison blindness. *Visual cognition*, 7(1–3), 253–267.
- Simons, D. J. (1996). In sight, out of mind: When object representations fail. *Psychological Science*, 7(5), 301–305.
- Sligte, I. G., Scholte, H. S., & Lamme, V. A. (2008). Are there multiple visual short-term memory stores? *PLOS one*, 3(2), e1699.
- Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior research methods, instruments, & computers*, 31(1), 137–149.
- Taylor D. A. (1976) Effect of identity in the multiletter matching task. *Journal of Experimental Psychology: Human Perception & Performance*. 2(3), 417–428.
- Treisman, A., & Zhang, W. (2006). Location and binding in visual working memory. *Memory & cognition*, 34(8), 1704–1719.
- Udale, R., Farrell, S., & Kent, C. (2018). Task demands determine comparison strategy in whole probe change detection. *Journal of Experimental Psychology: Human Perception and Performance*, 44(5), 778.
- Watanabe, K. (2003). Differential effect of distractor timing on localizing versus identifying visual changes. *Cognition*, 88(2), 243–257.
- Wilson, S., & Goddard, P. A. (2011). The effect of cueing on change blindness and same blindness. *Visual Cognition*, 19(8), 973–982.
- Yin, J., Gao, Z., Jin, X., Ye, L., Shen, M., & Shui, R. (2011). Tracking the mismatch information in visual short term memory: an event-related potential study. *Neuroscience letters*, 491(1), 26–30.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.