The effect of stimulus size on stereoscopic fusion limits and response criteria

Philip M Grove1, Nonie J Finlayson1, Hiroshi Ono2
1 School of Psychology, The University of Queensland, St Lucia, QLD 4072, Australia;
2 Centre for Vision Research and Department of Psychology, York University, Toronto, Canada;
e-mail: p.grove@psy.uq.edu.au
Received 22 April 2013, in revised form 14 February 2014

Abstract. The stereoscopic fusion limit denotes the largest binocular disparity for which a single fused image is perceived. Several criteria can be employed when judging whether or not a stereoscopic display is fused, and this may be a factor contributing to a discrepancy in the literature. Schor, Wood, and Ogawa (1984 Vision Research, 24, 661–665) reported that fusion limits did not change as a function of bar width, while Roumes, Plantier, Menu, and Thorpe (1997 Human Factors, 39, 359–373) reported higher fusion limits for larger stimuli than for smaller stimuli. Our investigation suggests that differing criteria between the studies could contribute to this discrepancy. In experiment 1 we measured horizontal and vertical disparity fusion limits for thin bars and for the edge of an extended surface, allowing observers to use the criterion of either diplopia or rivalry when evaluating fusion for all stimuli. Fusion limits were equal for thin bars and extended surfaces in both horizontal and vertical disparity conditions. We next measured fusion limits for a range of bar widths and instructed observers to indicate which criterion they employed on each trial. Fusion limits were constant across all stimulus widths. However, there was a sharp change in criterion from diplopia to rivalry when the angular extent of the bar width exceeded about twice the fusion limit, expressed in angular terms. We conclude that stereoscopic fusion limits do not depend on stimulus size in this context, but the criterion for fusion does. Therefore, the criterion for fusion should be clearly defined in any study measuring stereoscopic fusion limits.

Keywords: binocular fusion, stereopsis, diplopia, rivalry

1 Introduction
Wheatstone’s (1838) invention of the stereoscope showed that disparate points in the two eyes could be stimulated with similar images and still be perceived as single. The range of disparities for which single vision is experienced is called Panum’s fusional range (Grove, 2012; Panum, 1858; Steinman, Steinman, & Garzia, 2000). This range is usually determined by increasing the disparity of a target and noting the magnitude at which its images just appear unfused, called the fusion limit; and reducing an initially large disparity and noting the magnitude at which initially unfused images appear fused, called the re-fusion limit. Panum’s fusional range is taken as the mean of the fusion and re-fusion limits (see Howard & Rogers, 2012).(1)

Binocular single vision is crucial for optimal stereoscopic performance and visual comfort. Diplopia is associated with visual discomfort after prolonged viewing of three-dimensional media (Tam, Speranza, Yano, Shimono, & Ono, 2011; Wopking, 1995). Stereoscopic discrimination is poorer for diplopic targets than for fused targets (Ogle, 1952). Lastly, visual direction judgments are compromised for diplopic images, as there are two direction estimates for a single object (Ono & Mapp, 1995). Therefore, it is important to specify the stimulus properties that affect the range of fusible horizontal and vertical disparities as well as specifying the criteria used by observers to determine whether or not the binocular inputs are fused.

(1) We cite the most recent volume of Howard and Rogers’s book. However, they raised the issues described in this paper in their original volume in 1995.
Observers can employ several criteria when judging the fusion and re-fusion limits of a given stimulus. Considering horizontal disparities, in simple line or bar stimuli apparent depth accompanies the introduction of disparity and could be used as a cue for fusion. For example, as horizontal disparity is increased beyond the range for which precise metrical depth is experienced, referred to as patent stereopsis by Ogle (1952), stereoscopic depth is noticeably degraded. This transition could be interpreted as a loss of fusion. Conversely, the emergent perception of depth when decreasing the disparity of initially unfused stimuli with no apparent depth may be interpreted as a return of fusion, though the stimuli may still be diplopic. The criterion of apparent depth does not apply for vertical disparity, however, as vertical disparities do not result in the perception of relative depth.\(^{(2)}\) Rather, vertical disparities signal absolute viewing distance (Rogers & Bradshaw, 1993) and eccentricity (Gillam & Lawregren, 1983; Mayhew & Longuet-Higgins, 1982).

A second criterion for evaluating fusion is a thickening of the initially fused stimulus [which could be a thin line, extended bar, or more specialized stimulus such as a difference of Gaussian (DoG) patch]. As disparity (horizontal or vertical) is increased past the fusion limit, an initially fused stimulus may first thicken perceptually as the contours of identical contrast polarity move apart. Conversely, an initially diplopic stimulus may perceptually merge into a thickened image as disparity is reduced, and this percep may be interpreted as a fused stimulus. Perceived thickening is commonly reported in studies employing one-dimensional (1-D) DoG patches (eg Schor, Wood, & Ogawa, 1984; Wilson, Blake, & Halpern, 1991), which appear as a bright luminance band flanked by dark bands, or two-dimensional (2-D) DoG patches (eg Roumes, Plantier, Menu, & Thorpe, 1997), which appear as a bright circular dot surrounded by a dark circular region, but less so for bar or line stimuli. Additionally, 2-D DoG patches can perceptually distort into an oval shape at the limits of fusion, or appear as two distinct circles (Roumes et al., 1997).

It is possible that perceived thickening coincides with another criterion, binocular rivalry. For extended objects or surfaces, increasing disparity moves contours of the same contrast polarity away from correspondence until fusion breaks. Surpassing the fusion limit results in the perception of a form of area–contour rivalry (Howard & Rogers, 2012) along one of the disparate edges as it competes with the central region of the bar in the other eye’s image. The last and most common criterion is diplopia, in which, with increasing disparity, thin lines are increasingly removed from correspondence such that they perceptually separate in cyclopean perception into two distinct images.

An additional complexity in the literature on binocular fusion of simple stimuli is the interaction between stimulus type and the criterion employed by observers. Howard and Rogers (2012) pointed out that the majority of experiments on binocular fusion limits employed lines, bars, or dots as stimuli presented against a background of opposite luminance polarity (ie white bars on a black background or vice versa). They observed that, when disparity is increased in these stimuli, the contours with the same contrast polarity are moved further from correspondence, while contours of opposite polarity are first brought closer to correspondence, fall on corresponding points, and then separate further with increasing disparity. In this context, using the criteria of diplopia as the fusion limit is a measure of when contours of opposite polarity perceptually break. In order to measure diplopia thresholds for contours of the same contrast polarity, Howard and Rogers (2012) suggested using extended bars with observers monitoring the fusion status of one of the vertical edges (see figure 1b for an example). This ensures that thresholds are measured for contours with the same contrast polarity. However, using the criterion of diplopia would likely inflate fusion thresholds for

\(^{(2)}\)This is true in normal vision. However, vertical magnification of one eye’s image in the laboratory or with a meridional lens yields the perception of slant. This is referred to as the induced effect (Ogle, 1938).
wider bars and extended surfaces because the transition from a single image to two distinct images, or vice versa, would only be visible at larger disparities. A better criterion indicating fusion would be the presence or absence of area–contour rivalry as a function of disparity.

Given that observers can employ at least four different criteria when evaluating stereoscopic fusion, it is possible that threshold measurements vary depending on the combination of stimulus type and criterion or criteria used. Furthermore, discrepancies between reported fusion limits may be attributable to different criteria employed across studies. For example, Schor et al. (1984) reported that, for 1-D DoG patterns, fusion thresholds were lower for patches of high spatial frequency than those for patches of low spatial frequencies. Additionally, they reported that fusion thresholds were unchanged for sharp edge bar stimuli of various widths equivalent to the bright bar of the set of DoG patches they employed, though the bars contained a broad spectrum of spatial frequencies. Schor et al. concluded that, as fusion thresholds increased with decreasing spatial frequency of the DoG stimuli but remained constant for broadband bar stimuli (and at similar values observed for the high spatial frequency DoGs), fusion thresholds are determined by the highest spatial frequency component in the stimulus.

Conversely, Roumes et al. (1997) reported that fusion thresholds were increased for compound 2-D DoG patches containing both high and low spatial frequencies compared with 2-D DoGs containing only high spatial frequencies. They concluded that the low spatial frequency content expands the fusional range of those stimuli, contrary to Schor et al.’s (1984) claim that high spatial frequency content determines the upper fusion limit. In a second experiment Roumes et al. measured fusion limits for 2-D DoG patterns with either high or low spatial frequency content and for solid disk stimuli equal in diameter to the width of the positive phase of the DoGs. They found that fusion limits were similar for both DoG and solid disk stimuli, and that fusion limits were smaller for higher spatial frequency DoGs and smaller disks than for lower spatial frequency DoGs and larger disks, inconsistent with Schor et al.’s report that fusion limits were unchanged across a range of bar widths.

There are several methodological differences between Schor et al. (1984) and Roumes et al. (1997). For example, Schor et al. required strict fixation during stimulus presentation, whereas Roumes et al. did not restrict eye movements. Another difference is that Roumes et al. did not state what criteria were used to evaluate fusion for the sharp edged disk stimuli, while Schor et al. (1984) stated that “disparity was increased by method of adjustment until the subject reported a slight doubling, an increase in width or lateral displacement” (page 662). The later criterion is consistent with suppression of one eye’s image, indicating that their observers were free to choose diplopia, widening, or rivalry. If observers in Roumes et al.’s study used a strict criterion of diplopia for large and small disks, such that they ensured that two distinct images were visible before indicating that fusion had broken, this practice would result in higher fusion limits for larger stimuli since disparity settings on fused to unfused trials would be inflated for larger disks relative to smaller ones. The apparent discrepancy between Roumes et al. and Schor et al. motivated us to explore how an observer’s criterion for evaluating fusion might influence his or her fusion thresholds.

Therefore, this report has two goals. The first is to compare fusion limits for thin lines and extended surfaces to evaluate Howard and Rogers’s (2012) criticism of previous studies, which have failed to assess fusion thresholds for contours of the same contrast polarity.

(3) Wilcox and Hess (1995) concluded that the upper disparity limit was dependent on stimulus size and not spatial frequency content. However, they used the criterion of loss of perceived depth as the upper disparity limit. Presumably, all stimuli appeared diplopic at these large disparities regardless of their size.
We measured fusion limits for thin lines and extended bars for both horizontal and vertical disparities. The second goal was to systematically explore the effects of stimulus size on fusion limits and on the criterion used to evaluate fusion. Elucidating any relationship between the criterion for fusion and fusion thresholds could suggest a reason for the discrepancy between Schor et al. (1984) and Roumes et al. (1997). We pursued the second goal by employing a wide range of line/bar widths, measuring diplopia thresholds and recording observers’ criteria for when fusion broke or returned, to determine if criterion changes as a function of stimulus size, and to see if a change in criterion is accompanied by a change in the magnitude of fusion limits.

2 Experiment 1
This experiment measured and compared fusion thresholds for thin bars and extended bars when horizontal or vertical disparity was manipulated. Importantly, observers were informed of two criteria, diplopia and rivalry, to base their judgments on. We limited the criteria to these two choices because each involved a qualitative change in perception. In the case of diplopia, the qualitative change occurs when an initially fused object is perceived as single, but perceived as two objects when fusion breaks. Alternatively, the initial perception of two objects fuses to one when re-fusion occurs. In the case of rivalry, the qualitative change occurs when an initially fused edge is perceived as sharp and single, but rivals with the corresponding region in the other eye when fusion breaks. Rivalry ceases when re-fusion occurs. We reasoned that the qualitative changes associated with diplopia and rivalry are more compelling, perceptually, than the criterion based on a gradual change in width of the fused/unfused image. We assumed that observers’ criterion for evaluating fusion of thin lines would be the presence or absence of diplopia. For extended bars, we assumed observers’ criterion would be the presence or absence of area–contour rivalry. These criteria should remain stable for thin lines and extended bars.

2.1 Method
2.1.1 Observers. Two of the authors (PG and NF) plus five observers naive to the purpose of the experiment participated. All reported normal or corrected-to-normal binocular vision.

2.1.2 Apparatus and stimuli. Stimuli were drawn and scripted using Psychophysics Toolbox (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997) for Matlab and presented on two 24-inch Macintosh Cinema displays in a mirror stereoscope. The optical distance was 65 cm, and one pixel subtended 1.46 arcmin. Observers sat in a dark room with their heads stabilized by a chin-and-forehead rest and viewed the stimuli via two front-silvered mirrors oriented at ±45° to the sagittal plane. Observers were positioned such that the centres of their pupils were at the same elevation as the centre of the display. Their eyes were converged at 65 cm.

The fixation stimulus and all test stimuli were drawn in black (0.06 cd m⁻²) on a white (110.3 cd m⁻²) background. The fixation stimulus, positioned in the centre of the display, consisted of a black dot subtending 12 arcmin surrounded by a ring (width 2.9 arcmin; diameter 24 arcmin). Vertical Nonius lines (width 2.9 arcmin; height 15 arcmin) extended above and below the fixation ring to monitor horizontal vergence, and horizontal Nonius lines extended to the left and right of the fixation ring to monitor vertical vergence. We chose to use only a central fixation stimulus and no surrounding fusion lock to accommodate the test stimuli, some of which extended 12 deg into the periphery either horizontally or vertically. We reasoned, first, that either the configuration of the fusion lock would need to be changed for different test stimuli, introducing a potential confound in our experiment. Second, another option, to use a single fusion lock in the periphery, was discounted because it would be too eccentric to aid in fusion.
Test stimuli were of two general types. The first type measured horizontal disparity limits and consisted of a thin bar (4.4 arcmin wide and 1 deg tall) and an extended bar (12 deg wide (4) and 1 deg tall), positioned 1.2 deg above the fixation stimulus. The midline of the thin bar was centred in the median plane above the fixation stimulus at zero disparity. For the extended bar, one edge was positioned in the median plane above the fixation stimulus at zero disparity. Disparity was introduced by equal and opposite shifts in the left and right eyes’ images of the thin bar or the extended bar. Therefore, the thin bar or the edge of the extended bar was perceived straight ahead for fused disparities. The wide bar extended to the left side of the display on half the trials and to the right of the display on the other half.

For measurements of vertical disparity limits, the thin bar and an extended bar were rotated 90° and positioned 1.2 deg to the left or right of the fixation stimulus. For these measurements, the midline of the thin horizontal bar was centred on the horizontal plane of regard intersecting the middle of the fixation stimulus. For the extended bar, one edge was centred on the horizontal plane of regard and extended towards the top of the display on half the trials and towards the bottom of the display on the other half. The test stimulus appeared to the left of the fixation point on half the trials and to the right on the other half. Examples of fixation and test stimuli are shown in figure 1.

Figure 1. Stimuli used in experiment 1. (a) Stimulus for horizontal disparity fusion limits for a thin bar; (b) same as in (a) but for an extended bar (note the horizontal extent of the bar in the figure is less than in the actual experiment); (c) stimulus for vertical disparity fusion limits for a thin bar; (d) same as in (c) but for an extended bar (note the vertical extent of the bar in the figure is less than in the actual experiment). Stereograms are for cross-fusion. Diplopia should be apparent in (a) and (c); rivalry should be apparent in (b) and (d).

(4) We chose an extended bar to ensure that fusion judgments were made based on the perceptual status of the edge located in the median plane and not on the more peripheral edge.
2.1.3 Procedure. We used the method of adjustment, which has been used in similar investigations when fixation is required (Schor et al., 1984). Prior to the formal experiment, observers were shown examples of the line stimuli and the extended bar stimuli. They were encouraged to make a few practice settings to experience both types of stimuli when fused and unfused. Observers were instructed to use the appearance or disappearance of diplopia or rivalry as their criteria to indicate the images had become unfused or fused, respectively. Diplopia was described as the experience of two distinct images in perception. Rivalry was described as the appearance of lustre along the edge of the bar. Observers were told that suppression of one image was possible and this would be perceived as a lateral displacement of one of the thin lines or of the vertical edge of the bar stimulus. Observers were further instructed that rivalry may take some time to appear and that they were to scrutinize stimuli for one or two seconds to ensure that stimuli that appeared fused initially did not start to rival. This practice continued until the experimenter was confident that the observer had a clear understanding of the perceptions of diplopia and rivalry.

In the formal experiment observers were instructed to hold their gaze on the fixation point and monitor the horizontal and vertical Nonius lines to ensure their alignment while adjusting the disparity of the target stimuli. Observers either increased the disparity of initially fused targets until binocular fusion broke or reduced the disparity until initially unfused targets were perceptually fused. Disparity was increased and decreased by pressing the ‘up’ and ‘down’ arrow keys, respectively, on a computer keyboard. Observers recorded their responses by pressing the space bar. Although responses involved increasing or decreasing disparity, observers were allowed to make fine adjustments in their settings involving small reversals in cases when their settings missed the exact point when fusion broke or returned. A warning tone sounded whenever settings were made in the direction opposite to what was intended for the trial to discourage large reversals in the settings. We arbitrarily signed uncrossed horizontal disparities positive and crossed disparities negative. We signed vertical disparities arising from upward displacements of the right eye’s image and downward displacements of the left eye’s image as positive and the opposite displacements negative. The experiment was self-paced, and viewing time was unlimited.

At the start of each trial, targets could have zero disparity relative to the plane of the display, or an initial crossed or uncrossed disparity of 72 arcmin. The initial disparity of 72 arcmin was sufficient to induce a perception of diplopia (for thin lines) or rivalry (for extended bars) for all observers. In half of the increasing trials, observers were cued to increase the crossed disparity of the target with an initial disparity of zero. In the other half of the increasing trials, observers were cued to increase uncrossed disparities. In the other half of the trials, they were cued to reduce the disparity, crossed or uncrossed, until fusion returned. For horizontal disparity measurements, observers completed at least 10 trials for each permutation of bar size (thin line, extended bar), disparity (crossed, uncrossed), direction of adjustment (increasing, decreasing), and side of the display (extending to the left, extending to the right), for a minimum of 160 randomized trials. Some observers with less stable responses completed 12 or 15 trials per condition for a total of 192 or 240 trials, respectively. For vertical disparity measurements all observers completed 6 trials for each permutation of bar size (thin line, extended bar), disparity (positive, negative), direction of adjustment (increasing, decreasing), extension of the bar (upwards, downwards), and side of the display (left of fixation, right of fixation), for a total of 192 randomized trials. Fewer trials were required in the vertical disparity condition because observers’ responses were less variable than when responding to horizontal disparities.
2.2 Results

2.2.1 Horizontal disparities. For horizontal disparities we first tabulated the mean settings for which the bar extended to the left versus the right and tested for a difference between these two sets using a paired $t$-test. This was nonsignificant ($t_6 = 0.38, p > 0.05$), and so we collapsed across this factor, bringing the total observations in each condition up to at least 10 per observer. The total number of observations was higher for some observers whose responses were more variable. We next computed the average single to double adjustments and compared them with the average double to single adjustments for crossed and uncrossed disparities, respectively. Hysteresis was evident for each stimulus type with increasing trials yielding larger disparities than decreasing trials (see figures 2a and 2c).

We next computed the overall mean of increasing and decreasing trials for thin bars for crossed and uncrossed disparity, respectively, and likewise for extended bars. These means, based on at least 20 observations per condition for each observer, served as the units in our subsequent analyses.

Figure 2 illustrates the group mean fusion thresholds for horizontal and vertical disparities for each condition. Inspection of figure 2a shows fusion thresholds for each stimulus for both single to double and double to single adjustments. Hysteresis is apparent when comparing adjacent solid and hatched bars. Figure 2b illustrates the group mean fusion thresholds collapsed across the direction of adjustment. Inspection of the graph reveals no systematic differences in fusion thresholds between line stimuli and extended bar stimuli. However, overall fusion thresholds are greater for uncrossed disparities than for crossed disparities.

Figure 2. Group mean ($n = 7$) stereoscopic fusion limits. (a) Horizontal disparity fusion thresholds for single to double (solid bars) and double to single (hatched bars) conditions are plotted separately to show hysteresis. Bars with white backgrounds indicate data for thin bars; bars with grey backgrounds indicate data for extended bars. (b) Mean horizontal disparity fusion thresholds collapsed across the direction of adjustment. (c) Same as (a) but for vertical disparity. (d) Same as (b) but for vertical disparity. In (a) and (b) positive values represent uncrossed disparities, negative values represent crossed disparities. Error bars represent ±1 SEM.
A 2 bar width (thin line, extended bar) × 2 disparity (crossed, uncrossed) ANOVA with mean fusion threshold as the repeated measure supported the observations described above for horizontal disparities. We removed the sign of the disparity for this analysis in order to examine the difference in magnitude in fusion thresholds for crossed and uncrossed disparities. This analysis revealed no significant effect for width (line versus extended bar) \((F_{1,6} = 0.2, p = 0.6)\). Sign of disparity yielded a significant difference \((F_{1,6} = 7.1, p = 0.04)\), with larger absolute fusion thresholds observed for uncrossed disparities (12.7 arcmin) than for crossed disparities (7.7 arcmin). There was no significant interaction between width and sign of disparity \((F_{1,6} = 2.6, p = 0.2)\).

We speculate that the significant difference observed for sign of horizontal disparity is because all measurements occurred in the upper visual field. Such a pattern is expected if Panum’s fusional range is centred on the backwards-tilted empirical vertical horopter (Grove, Kaneko, & Ono, 2001; Nakayama, 1977; Schreiber, Hillis, Filippini, Schor, & Banks, 2008). We specifically tested this in experiment 2 and return to this issue in the discussion (section 4).

### 2.2.2 Vertical disparities

We carried out similar analyses for vertical disparity measurements. We first tested for a difference in settings between conditions where the bar extended upwards versus downwards using a paired \(t\)-test. This was nonsignificant \((t_6 = 0.26, p = 0.8)\). We next tested for differences between conditions when the bar appeared on the left or right side of fixation. Again, the paired \(t\)-test was nonsignificant \((t_6 = 0.3, p = 0.8)\), and so we collapsed across these factors for our formal analyses. Fusion thresholds for each stimulus for both single to double and double to single adjustments, are shown in figure 2c. Hysteresis is apparent when comparing adjacent solid and hatched bars. Group mean fusion thresholds, collapsed across the direction of adjustment, are shown in figure 2d. Inspection of the graph reveals a slightly elevated threshold for extended bar stimuli over line stimuli, though this difference was not significant.

A 2 bar width (thin line, extended bar) × 2 disparity (positive, negative) ANOVA with mean fusion threshold as the repeated measure supported the observations described above for vertical disparities. We removed the sign of the disparity for this analysis in order to examine the difference in magnitude in fusion thresholds for positive and negative vertical disparities. This analysis revealed no significant effect for width (line versus extended bar) \((F_{1,6} = 4.2, p = 0.1)\), nor for sign of disparity \((F_{1,6} = 0.8, p = 0.4)\). There was no significant interaction between width and sign of disparity \((F_{1,6} = 0.007, p = 0.9)\).

We found no systematic differences in the magnitude of fusible horizontal or vertical disparities between thin lines and extended bars when observers were free to choose the appearance/disappearance of either diplopia or rivalry as their criterion to indicate the loss or acquisition of stereoscopic fusion. Because we found no systematic difference in fusion thresholds between thin lines and extended bars for either disparity type, we computed the overall mean fusion threshold for all stimuli for horizontal and vertical disparities, respectively. Consistent with previous data (eg Schor et al., 1984), fusion thresholds for horizontal disparities \((M = 9.8 \text{ arcmin}, SD = 4.3)\) were larger, on average, than for vertical disparities \((M = 7.1 \text{ arcmin}, SD = 1.5)\), though this difference was not statistically significant.

In accordance with our assumptions, we noted informally from postexperiment debriefings that observers primarily reported the presence/absence of diplopia for thin bars and the presence/absence of rivalry for extended bars. We speculated that one’s criterion changes from diplopia to rivalry or vice versa at some point as bar width changes. It is possible that diplopia thresholds are affected during this transition. Our coarse manipulation in experiment 1 of just a line and an extended bar masked this possible effect. Therefore, in experiment 2 we systematically manipulated bar width and measured observers’ diplopia thresholds and their criterion for the threshold.
3 Experiment 2
In experiment 1 we instructed observers to use the presence or absence of either diplopia or rivalry when evaluating stereoscopic fusion, but we did not formally ascertain when observers employed diplopia or rivalry as their criterion. In this experiment we examined how bar width affects observers’ selection of criterion more closely. Using similar methods to experiment 1, we measured diplopia thresholds for a range of line/bar widths for both horizontal and vertical disparities. We also instructed observers to use either the criterion of diplopia or rivalry, whichever was the more compelling perception, and we required them to indicate which criterion they used on every trial. Therefore, we obtained data for both the magnitude of disparity at which fusion broke or returned, and for which criterion was used on each trial.

3.1 Method
3.1.1 Observers. PG and NF plus two naive observers from experiment 1 participated.

3.1.2 Apparatus and stimuli. The apparatus was the same as for experiment 1. Stimuli were similar to experiment 1, except that we used a range of bar widths from 2.9 to 36 arcmin. The specific range of bar widths was determined for each observer based upon pilot data. This was to ensure we employed wide enough bars to elicit exclusive rivalry responses and narrow enough bars to elicit exclusive diplopia responses indicating a break in fusion. A black circular annulus (width 9.6 arcmin, diameter 5 deg) surrounded the entire display to act as an additional vergence lock.

For horizontal disparity measures and for all bar widths, one of the vertical edges was aligned with the median plane of the head. In half the trials the bar extended to the left, and in the other half the bar extended to the right. Moreover, to test our speculation about the difference in fusion limits for crossed and uncrossed disparities in experiment 1, half of the stimuli were positioned 1.2 deg above the fixation and half 1.2 deg below the fixation point. Disparity was introduced with equal and opposite shifts in the left and right eye’s images. The method for measuring fusion limits for vertical disparities was the same as for experiment 1.

3.1.3 Procedure. On a given trial, observers adjusted the disparity of the target, either increasing disparity from zero or reducing an initially large disparity (72 arcmin) and indicated either when fusion broke or when the images re-fused. Additionally, observers were prompted to indicate which criterion they used for their judgment, diplopia or rivalry. Prior to testing, observers were instructed that for increasing trials they should indicate the first and most compelling perception that indicated a break in fusion. For descending trials, observers were instructed to indicate their last perception (diplopia or rivalry) before fusion returned. Increasing/decreasing disparity trials, bar widths, and bar side were randomized within each block. Visual field (upper/lower for horizontal disparities; left/right for vertical disparities) was blocked, with block order counterbalanced across observers. Observers completed four increasing and four decreasing adjustments for crossed disparities and the same for uncrossed disparities. Therefore, observers completed 16 disparity adjustments for each bar width. Correspondingly, there were 16 reports of which criterion was used for fusion at each bar width. Observers completed 144 trials for horizontal disparities and 144 trials for vertical disparities.

3.2 Results
3.2.1 Fusion limits for horizontal disparities. We signed horizontal and vertical disparities in the same manner as experiment 1 and combined the horizontal disparity settings for each individual as in experiment 1. Individual mean fusion thresholds are plotted for horizontal disparities as a function of bar width, sign of disparity, and visual field in figure 3a. Inspection of the figure reveals no systematic change in fusion thresholds as a function of bar width, with thresholds approximately constant across the range of tested bar widths for all observers.
Figure 3. (a) Individual stereoscopic fusion limits for horizontal disparity as a function of bar width. Closed symbols represent settings in the upper visual field; open symbols represent settings in the lower visual field. Squares represent settings for uncrossed disparity, and circles for crossed disparity. Error bars in the left panels represent ±1 SEM. (b) Percentage of trials in which each observer reported that the targets appeared diplopic as a function of bar width. Curves are cumulative Gaussians fitted using maximum likelihood estimation. Fitted parameters are displayed for each observer.
Notably, the pattern of fusion thresholds shows that fusion limits for uncrossed disparities are larger in the upper visual field than in the lower visual field for every observer. Similarly, fusion limits for crossed disparities are larger in the lower visual field than in the upper visual field for every observer. This pattern of results is consistent with our speculation in experiment 1 that the differences in absolute fusion limits for uncrossed and crossed disparities were because the fusional volume for horizontal disparities is biased such that it is centred on the empirical vertical horopter.

We next calculated the mean fusion threshold across all bar widths for crossed and uncrossed disparities for each individual and for the group. These data are shown in figure 4. Inspection of the graph reveals that fusion limits were larger overall for crossed disparities than for uncrossed disparities for three of our four observers. Paired $t$-tests on the individual data revealed significant differences for two observers (PG: $t_8 = 21.4, p < 0.05$; JR: $t_6 = 11.1, p < 0.05$) and nonsignificant differences for the other two (NF: $t_8 = -1.7, p > 0.05$; WH: $t_6 = 1.59, p > 0.05$). An analysis on the group mean data revealed nonsignificant results ($t_3 = 1.96, p > 0.05$). These results are most consistent with Roumes et al.’s (1997) data from experiment 1 (figure 4 in their paper) that indicate no difference in fusion limits between crossed and uncrossed disparities for what they termed neural fusion—that is, fusion limits with minimal eye movements, which is a closer comparison to our procedure. We return to this point in section 4.

### 3.2.2 Tabulating response criteria for horizontal disparity fusion limits

For each observer we tabulated the number of trials in which diplopia or rivalry was reported as the criterion used to evaluate the fusion status of the stimulus. To ascertain if there was a systematic transition between the two criteria as a function of bar width, we plotted the number of reported diplopia responses as a function of bar width. Inspection of figure 3b shows that, for thin bars, all four observers employed diplopia as their criterion for fusion exclusively. As bar width increased, observers shifted almost exclusively to rivalry as their criterion.

We determined the point at which observers changed from the criterion of diplopia to rivalry by fitting a sigmoid curve to the data using maximum likelihood estimation (Wichmann & Hill, 2001) and took the 50% point of the curve as marking the transition from a criterion of diplopia to rivalry. We fitted the data with a cumulative Gaussian curve using four parameters: $\alpha$ and $\beta$ describe the inflection point and the slope of the fitted function. The parameters $\gamma$ and $\lambda$ indicate the guess rates and lapse rates, respectively, and were both constrained within $[0, 0.05]$, expressed as a proportion here and in figures 3 and 5. Confidence intervals were calculated with a bootstrap procedure ($n = 999$) (Effron, 1987) (see table 1).

### Figure 4

Individual mean fusion thresholds collapsed across bar width. The last pair of bars is the mean of the four observers. Error bars are ±1 SEM.

For each observer we tabulated the number of trials in which diplopia or rivalry was reported as the criterion used to evaluate the fusion status of the stimulus. To ascertain if there was a systematic transition between the two criteria as a function of bar width, we plotted the number of reported diplopia responses as a function of bar width. Inspection of figure 3b shows that, for thin bars, all four observers employed diplopia as their criterion for fusion exclusively. As bar width increased, observers shifted almost exclusively to rivalry as their criterion.

We determined the point at which observers changed from the criterion of diplopia to rivalry by fitting a sigmoid curve to the data using maximum likelihood estimation (Wichmann & Hill, 2001) and took the 50% point of the curve as marking the transition from a criterion of diplopia to rivalry. We fitted the data with a cumulative Gaussian curve using four parameters: $\alpha$ and $\beta$ describe the inflection point and the slope of the fitted function. The parameters $\gamma$ and $\lambda$ indicate the guess rates and lapse rates, respectively, and were both constrained within $[0, 0.05]$, expressed as a proportion here and in figures 3 and 5. Confidence intervals were calculated with a bootstrap procedure ($n = 999$) (Effron, 1987) (see table 1).
We next discarded the sign of the horizontal disparity and calculated the overall mean fusion threshold for each observer. These values are listed in table 1. Comparing the mean horizontal fusion limits with the bar widths at which the criterion switched, it is apparent that observers switch their criterion from diplopia to rivalry at bar widths close to twice the overall fusion threshold.

**Table 1.** Individual mean horizontal disparity fusion limits and the 50% transition point marking the transition from diplopia responses to rivalry.

| Observer | Horizontal disparity fusion limit/arcmin | 50% transition point from diplopia to rivalry/arcmin |
|----------|----------------------------------------|---------------------------------|
| PG       | 14.3                                   | 28.9 (28.1–29.6)                |
| NF       | 10.2                                   | 32.8 (31.3–33.8)                |
| JR       | 10.3                                   | 23.5 (20.3–25.6)                |
| WH       | 11.7                                   | 23.3 (19.7–26.5)                |

Note: 95% confidence intervals are given in parentheses.

3.2.3 *Fusion limits for vertical disparities.* We combined the vertical disparity settings for each individual as in experiment 1. Mean fusion thresholds are plotted for vertical disparities as a function of bar width in figure 5a. Inspection of the figure shows that fusion thresholds remain approximately constant across the range of tested bar widths.

3.2.4 *Tabulating response criteria for vertical disparity fusion limits.* As with the fusion limits for horizontal disparities, we tabulated the number of trials in which diplopia and rivalry were reported as the criterion used to make a judgment about the fusion of the stimulus for each observer. We plotted the number of reported diplopia responses as a function of bar width. Inspection of figure 5b shows that, for thin bars, all four observers employed diplopia as their criterion for fusion exclusively. As bar width increased, 3 of 4 observers shifted to rivalry as their exclusive criterion. The fourth observer, WH, markedly changed criterion as bar width increased but did not adopt rivalry exclusively. We fitted cumulative Gaussians to the vertical disparity data with the 50% point of the curve marking the transition from a criterion of diplopia to rivalry using the same methods as for the horizontal disparity data. These values are listed in table 2, with the associated 95% confidence intervals in parentheses.

We next discarded the sign of the vertical disparity and calculated the overall mean fusion threshold for each observer. These values are also listed in table 2. Comparing the mean vertical disparity fusion limits with the bar widths at which the criterion switched, we see that observers switch their criterion from diplopia to rivalry for bar widths that were at least twice the overall fusion threshold. The fusion limits for both horizontal and vertical disparities were approximately constant across all bar widths. However, there was a clear change in criterion from diplopia to rivalry as bar width increased.

**Table 2.** Individual mean vertical disparity fusion limits and the 50% transition point marking the transition from diplopia responses to rivalry.

| Observer | Vertical disparity fusion limit/arcmin | 50% transition point from diplopia to rivalry/arcmin |
|----------|----------------------------------------|---------------------------------|
| PG       | 6.3                                    | 11.8 (11.3–13)                  |
| NF       | 5.4                                    | 16.4 (15.7–17.2)                |
| JR       | 8.0                                    | 21.1 (20.2–22)                  |
| WH       | 5.6                                    | 27.1 (25.2–29.7)                |

Note: 95% confidence intervals are given in parentheses.
Discussion

Our results indicate that, when observers are free to use either the criterion of diplopia or rivalry to evaluate stereoscopic fusion of thin lines or extended bars, thresholds are independent of stimulus size. In both the experiments we found that stereoscopic fusion thresholds remained constant despite changes in the size of the bars in our displays for both horizontal and vertical disparities. On the other hand, experiment 2 showed that the criterion to evaluate fusion switches

Figure 5. (a) Individual stereoscopic fusion limits for vertical disparity as a function of bar width. Error bars in the left panels represent ±1 SEM. (b) Percentage of trials in which each observer reported that the targets appeared diplopic as a function of bar width. Curves are cumulative Gaussians fitted using maximum likelihood estimation. Fitted parameters are displayed for each observer.

4 Discussion

Our results indicate that, when observers are free to use either the criterion of diplopia or rivalry to evaluate stereoscopic fusion of thin lines or extended bars, thresholds are independent of stimulus size. In both the experiments we found that stereoscopic fusion thresholds remained constant despite changes in the size of the bars in our displays for both horizontal and vertical disparities. On the other hand, experiment 2 showed that the criterion to evaluate fusion switches
swiftly from diplopia to rivalry as bar width increases past about twice the fusion threshold, expressed in angular terms. Observers’ criteria were stable at both extremes of our bar width manipulation, with diplopia the exclusive criterion for evaluating fusion of thin lines and rivalry the dominant criterion for evaluating fusion of extended bars. However, there was a clear change in criterion from diplopia to rivalry across intermediate bar widths.

Howard and Rogers (2012) observed that measurements of fusion limits employing thin bars or dots as stimuli and diplopia as the response criterion measure fusion limits when contours of opposite polarity perceptually break. They argued that a measure of fusion limits for same contrast polarity contours would need to employ extended surfaces, as we did in experiments 1 and 2. However, as we have shown, the extent of the stimulus has no effect on fusion limits provided observers are free to choose either diplopia or rivalry as their criterion to evaluate fusion.

These results are most consistent with Schor et al. (1984), who concluded that diplopia thresholds are dependent on the highest spatial frequency present in the stimulus and not on the actual size of the binocular images. Moreover, our methodology of allowing observers choose either diplopia or rivalry when evaluating fusion is similar to Schor et al.’s method, in which their observers, who were experienced psychophysical observers, used any indication that fusion broke, presumably including diplopia, and rivalry. Roumes et al. (1997) reported conflicting data showing, in one experiment employing small and large solid disks, that larger diplopia thresholds were observed for larger disks than for smaller disks. One possible explanation for these results is that observers maintained a strict criterion of diplopia to evaluate fusion for both small and large stimuli. Therefore, larger disparities would be required to elicit diplopia for the larger disks than for the smaller disks. It is not known whether or not the observers in Roumes et al.’s study experienced rivalry for the larger disks. Given the results from the two experiments reported here, we predict that fusion thresholds would be similar for small and large disks if observers were free to choose either diplopia or rivalry as a criterion to evaluate fusion.

All our test stimuli were black lines or bars on a white background. One might ask whether reversing the polarity of the stimuli might influence fusion limits. A comparison of our results with those of Schor et al. (1984) indicates that polarity is not an important factor. Mean overall horizontal fusion limits in the present study were ~10 arcmin, comparable with those reported by Schor et al.’s two observers, which were ~8 and 12 arcmin, respectively. Our mean vertical disparity fusion limits were 7.1 arcmin in experiment 1 and 6.3 arcmin in experiment 2, again comparable to Schor et al.’s results, which ranged from 2.5 to 10 arcmin.

We observed a pattern of hysteresis in all our measurements consistent with the early reports by Fender and Julesz (1967) and the subsequent replication by Piantanida (1986). The data from experiment 1 show that hysteresis manifests similarly for thin lines and extended bars when adjustment of disparity is under the control of the observer.

One might ask why the fusion limits for horizontal disparities are larger than those for vertical disparities. There are at least two reasons why this is so. One is that vertical disparities are typically very small compared with horizontal disparities in 3-D scenes. They arise from vertical image differences between the two eyes when objects are positioned away from the median plane. For example, a 2 cm tall line viewed at the extreme edge of our stimulus display at a viewing distance of 65 cm, 13.8 cm from the median plane, generates binocular images with a vertical disparity of 0.9 arcmin, well within the range of fusible vertical disparities. A second reason is that vertical disparities can be thought of as calibration information for the binocular system to ensure optimal alignment of the eyes (Howard & Rogers, 2012). Vertical disparities arise when the vertical gaze elevations are not equal in the two eyes, introducing a vertical offset between corresponding features similar to what was simulated in the present experiments. Ogle (1955) reported decrements in stereo acuity in response to
horizontal disparities when the test targets also had a standing vertical disparity greater than 5 arcmin. Torsional misalignment of the eyes, when the horizontal meridians are not parallel, introduces vertical shear disparities across the visual field. The binocular system responds to vertical shear disparities with compensatory cyclorotary eye movements (Rogers & Howard, 1991). Therefore, it would be advantageous if the binocular system had a low tolerance for vertical disparity and was calibrated to keep it to a minimum. The larger range of fusible horizontal disparities, on the other hand, is due to the fact that they code for relative depth. Therefore, to process a wide range of disparities/depth, it is advantageous to experience single vision for a wide range of values.

In experiment 1 we observed a significant difference in absolute values of fusion limits between crossed and uncrossed disparities. This is likely due to the fact that all measurements were taken above the fixation point. Recently, Grove and Ono (2012) and Grove and Harrold (2013) demonstrated that, for locations above and below the fixation point, fusional ranges are shifted such that Panum’s fusional range is centred on the vertical horopter. Because the empirical vertical horopter is inclined top far (Grove et al., 2001; Nakayama, 1977; Schreiber et al., 2008), this would predict larger fusion thresholds for uncrossed disparities in the upper visual field than in the lower visual field and the opposite pattern would be predicted for crossed disparities. Collectively, the pattern of fusion limits for horizontal disparities in experiments 1 and 2 is consistent with this prediction.

The larger fusion limits for crossed disparities than for uncrossed disparities observed in experiment 2 are consistent with our account of the data in relation to the backwards inclination of the empirical vertical horopter. Cooper, Burge, and Banks (2011) reported large individual differences in the shape of the empirical vertical horopter, with some individuals showing differences in inclination between the upper and lower hemifields. The pattern of results reported here is consistent with the empirical vertical horopter that is steeply inclined in the lower hemifield but closer to vertical in the upper hemifield for the four observers in experiment 2.

In summary, we have shown that fusion thresholds for both horizontal and vertical disparities are independent of stimulus extent. However, in our experiments we found that observers evaluated fusion in thin bars using the criterion of presence or absence of diplopia and the presence/absence of rivalry to evaluate fusion for extended bars.

Although the discrepancy between Schor et al. (1984) and Roumes et al. (1997) was one motivation for us to explore the relationships among stimulus size, response criteria, and fusion limits, it was not our intention to decisively resolve the apparent conflict between the two studies. Observers’ response criterion is one of many differences between the studies, and a conclusive resolution is beyond the scope of this paper. Nevertheless, our investigation generates a number of questions related to Roumes et al.’s study. For example, how would the results reported here generalize to Roumes et al.’s stimuli? Another possible question is how would fusion limits be affected if observers were required to choose diplopia or rivalry exclusively as their response criterion? We have explored these questions in additional experiments described in the appendix. To summarize those investigations, we found that fusion thresholds are dependent on stimulus size when diplopia is enforced as the exclusive criterion but not when rivalry is the exclusive criterion.

**Acknowledgment.** PG was supported by ARC Discovery DP0878140.

**References**
Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 437–442.
Cooper, E. A., Burge, J., & Banks, M. S. (2011). The vertical horopter is not adaptable but it may be adaptive. *Journal of Vision, 11*(3):20, 1–19.
Effron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association, 82*, 171–200.
Fender, D., & Julesz, B. (1967). Extension of Panum’s fusional area in binocularly stabilized vision. Journal of the Optical Society of America, 57, 819–830.

Gillam, B. J., & Lawregren, B. (1983). The induced effect, vertical disparity, and stereoscopic theory. Perception & Psychophysics, 34, 121–130.

Grove, P. M. (2012). The psychophysics of binocular vision. In C. Zhu, Y. Zhao, L. Yu, & M. Tanimoto (Eds.), 3DTV system with depth-image-based rendering: Architectures, techniques and challenges. Hong Kong: Springer.

Grove, P. M., & Harrold, A. (2013). The range of fusible horizontal disparities around the empirical horopters. Paper presented at 2013 International Conference on 3D Imaging (IC3D), Liège, 3–5 December. Piscataway, NJ: IEEE. Available from doi:10.1109/IC3D.2013.6732084

Grove, P. M., Kaneko, H., & Ono, H. (2001). The backwards inclination of a surface defined by empirical corresponding points. Perception, 30, 411–429.

Grove, P. M., & Ono, H. (2012). Horizontal/vertical differences in range and upper/lower visual field differences in the midpoints of sensory fusion limits of oriented lines. Perception, 41, 939–949.

Howard, I. P., & Rogers, B. J. (2012). Perceiving in depth. Volume 2. Stereoscopic vision. Oxford: Oxford University Press.

Kleiner, M., Brainard, D. H., & Pelli, D. G. (2007). What’s new in Psychotoolbox-3? Perception, 36 (ECVP Supplement) 14 (Abstract).

Mayhew, J. E. W., & Longuet-Higgins, H. C. (1982). A computational model of binocular depth perception. Nature, 297, 376–378.

Nakayama, K. (1977). Geometric and physiological aspects of depth perception. Proceedings of Society of Photo-Optical Instrumentation Engineers, 120, 2–9.

Ogle, K. N. (1938). Induced size effect I. A new phenomenon in binocular space perception associated with the relative sizes of the images of the two eyes. Archives of Ophthalmology, 20, 604–623.

Ogle, K. N. (1952). On the limits of stereoscopic vision. Journal of Experimental Psychology, 44, 253–259.

Ogle, K. N. (1955). Stereopsis and vertical disparity. Archives of Ophthalmology, 53, 495–504.

Ono, H., & Mapp, A. P. (1995). A restatement and modification of Wells–Hering’s laws of visual direction. Perception, 24, 237–252.

Panum, P. L. (1858). Physiologische Untersuchungen über das Sehen mit zwei Augen. Kiel: Schwers.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. Spatial Vision, 10, 437–442.

Piantanida, T. P. (1986). Stereo hysteresis revisited. Vision Research, 26, 431–437.

Rogers, B. J., & Bradshaw, M. F. (1993). Vertical disparities, differential perspective and binocular stereopsis. Nature, 361, 253–255.

Rogers, B. J., & Howard, I. P. (1991). Differences in the mechanisms used to extract 3-D slant from disparity and motion parallax cues. Investigative Ophthalmology & Visual Science, 32, 695 (Abstract).

Roumes, C., Plantier, J., Menu, J. P., & Thorpe, S. (1997). The effects of spatial frequency on binocular fusion: From elementary to complex images. Human Factors, 39, 359–373.

Schor, C., Wood, I., & Ogawa, J. (1984). Binocular sensory fusion is limited by spatial resolution. Vision Research, 24, 661–665.

Schreiber, K. M., Hillis, J. M., Filippini, H. R., Schor, C. M., & Banks, M. S. (2008). The surface of the empirical horopter. Journal of Vision, 8(3):7, 1–20.

Steinman, S., Steinman, R. A., & Garzia, R. P. (2000). Foundations of binocular vision. Toronto, ON: McGraw Hill.

Tam, W. J., Speranza, S., Yano, S., Shimono, K., & Ono, H. (2011). 3D-TV: visual comfort. IEEE Transactions on Broadcasting, 57, 335–346.

Wheatstone, C. (1838). Contributions to the physiology of vision—Part the first. On some remarkable, and hitherto unobserved, phenomena of binocular vision. Philosophical Transactions of the Royal Society of London, 128, 371–394.

Wichmann, F., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. Perception & Psychophysics, 63, 1293–1313.
APPENDIX A

A1 Experiment A1

In order to generalize our findings to the stimuli used by Roumes et al. (1997), we first needed to replicate their basic experiment. We measured the upper fusion limit for small and large disks.

A1.1 Method

A1.1.1 Observers. One author (PG) plus three observers naive to the purpose of the experiment participated. All reported normal or corrected-to-normal binocular vision.

A1.1.2 Stimuli. Test stimuli were small (diameter 5.84 arcmin) or large (diameter 87.6 arcmin) black disks, presented on a white background. The fixation point was identical to experiments 1 and 2. Disks were presented with a range of horizontal disparities ranging from 0 to 116.8 arcmin, crossed and uncrossed.

A1.1.3 Procedure. We used the method of constant stimuli for this replication. Observers initially viewed the fixation point to ensure alignment of the Nonius lines. When ready, they pressed a button, and the disk stimulus replaced the fixation point as the only visible feature on the display. The disks were visible for a maximum of 5 s, after which the test stimulus was erased and the fixation stimulus returned. Observers were instructed not to make voluntary eye movements but to allow the disks to fuse if they perceptually drifted together during the viewing period. Observers were instructed to press the space bar on the keyboard at the moment the disks appeared to fuse. If the disks did not fuse in that time, the stimulus was extinguished. Alternatively, if the stimuli appeared unfused and did not appear to drift towards one another, the observer was allowed to indicate that the stimulus was not fused by pressing the appropriate key on the keyboard. Response times were recorded on every trial measuring the interval between the button press that elicited the test stimulus and the response button press or termination of the trial after the 5 s viewing window.

Observers PG, AH, and CC completed 10 trials for each of two disk sizes (5.84 and 87.6 arcmin) × 21 disparities (0–116.8 arcmin) × 2 directions (crossed and uncrossed) for a total of 840 trials over four blocks. Observer JR completed a smaller set of trials including 13 disparities (2.92–116.8) for a total of 520 trials.

A1.2 Results and discussion

Individual data for the upper limit of fusion and the associated response times are plotted in figures A1 and A2. In figures A1a and A2a the percentages of fused responses are plotted on the y-axis and disparity is plotted on the x-axis. In keeping with our data presentation in experiments 1 and 2, crossed disparities are negative in figure A1 and uncrossed disparities are signed positive in figure A2. It is clear from the figures that, as both uncrossed and crossed disparity increased, the percentage of fused responses decreased. Importantly, the percentage of fused responses decreases at smaller disparities for small disks than for larger disks, replicating Roumes et al.’s (1997) general findings.
Inspection of figures A1b and A2b indicate that, for three of four observers’ mean response times and response time variability increased with increasing disparity, consistent with eye movements being involved with the fusional response of larger disparities as reported by Roumes et al. One observer, CC, did not show a marked increase in response times. However, there is a marked increase the variability of her response times, suggesting that she too was employing eye movements to fuse the larger disparities.

**Figure A1.** (a) Percentage of fused responses as a function of crossed disparity for small (open circles) and large disks (filled circles). (b) Response times to indicate fusion as a function of crossed disparity for small and large disks. Error bars indicate ±1 SD.
This experiment verifies that we have established stimuli and an experimental protocol similar to Roumes et al. (1997). However, this experiment does not address the role of response criterion in evaluating the fusion status of the disks as a function of their size. Our next supplemental experiment addressed this issue.

A2 Experiment A2
In experiment 2 we found that, when observers are free to choose their criterion, fusion thresholds are independent of stimulus size. Here we conducted the converse of that investigation, measuring fusion limits for small and large disks and restricting observers’ response criterion to either diplopia or rivalry, exclusively. This experiment employed the same stimulus used by
Roumes et al. (1997). Observers made fusion judgments on the same stimulus set, once when using diplopia as their criterion and once using rivalry as their criterion.

A2.1 Method

A2.1.1 Observers. One author (PG) plus three observers experienced in stereoscopic psychophysical experiments but were naive to the purpose of the experiment. All reported normal or corrected to normal binocular vision.

A2.1.2 Stimuli. Test stimuli were the same as experiment A1. However, in order to allow observers enough time to make rivalry judgments, we used the same fixation protocol as in experiments 1 and 2, in which observers maintained strict fixation throughout the trial. The test stimuli were presented 1.2 deg above the fixation point. Disks were presented with a range of horizontal disparities ranging from 0 to 46.72 arcmin, crossed and uncrossed.

A2.1.3 Procedure. Prior to formal data collection, observers were shown samples of the stimuli and instructed how to evaluate them using either diplopia or rivalry as their criterion. To satisfy the criterion of diplopia, observers were told that the stimulus must be perceived as two distinct disks. The disks could still be overlapping perceptually, but there must be the impression of two distinct disks. To satisfy the criterion of rivalry, one of the edges of the black disks must appear to rival with the inner area of the other disk. Instructions and examples of stimuli generating each of the perceptions were presented until the experimenter was satisfied that the participant understood the task.

We used the method of constant stimuli for this replication. Observers viewed the fixation stimulus throughout each trial and ensured the alignment of the Nonius lines. When ready, they pressed a button and the disk stimulus appeared 1.2 deg above the fixation point. The disks remained visible until a response was made. In one block observers were instructed to evaluate the fusion status of the disks using diplopia as their criterion and report unfused stimuli by pressing the ‘D’ key on the keyboard. In another block they were instructed to use rivalry as their criterion and report unfused stimuli by pressing the ‘L’ key.

A different block order was used for each observer. Observers were instructed to press the space bar on the keyboard if the disks were fused.

Observer PG completed 20 trials for each of two disks sizes (5.84 and 87.6 arcmin) × 11 disparities (0 – 46.7 arcmin) × 2 directions (crossed and uncrossed) for a total of 1760 trials over four blocks. AH, BC, and JR completed 20 trials for each of two disks sizes (5.84 and 87.6 arcmin) × 8 disparities (0 – 29.2 arcmin) × 2 directions (crossed and uncrossed) × 2 criteria (diplopia and rivalry) for a total of 1280 trials over four blocks.

A2.2 Results and discussion

Individual graphs plotting the percentage of fused responses as a function of disparity for small and large disks using diplopia as the criterion are plotted in figure A3. Similar plots based on the criterion of rivalry are illustrated in figure A4. In keeping with our data presentation in experiments 1 and 2, crossed disparities are negative and uncrossed disparities are signed positive in the figures. It is clear in figure A3 that, as both crossed and uncrossed disparity increased, the percentage of fused responses decreased with the percentage of fused responses decreasing at smaller disparities for small disks than for larger disks. Therefore, when observers are required to maintain the criterion of diplopia when evaluating the fusion status of small and large disks, a clear pattern emerges where larger fusion limits are reported for large disks than for small disks.

Inspection of figure A4 shows a striking difference in the pattern of results from figure A3. Like figure A3, as both crossed and uncrossed disparity increased, the percentage of fused responses decreased with the percentage of fused responses decreasing. However, the percentage of fused responses decreases at the same disparity values for small and large
Figure A3. Individual plots illustrating the percentage of fused responses as a function of crossed disparity (left column) and uncrossed disparity (right column) for small and large disks when observers are forced to use the criterion of diplopia.
Figure A4. Individual plots illustrating the percentage of fused responses as a function of crossed disparity (left column) and uncrossed disparity (right column) for small and large disks when observers were forced to use the criterion of rivalry.
disks for three of the four observers, yielding the same fusion limit estimate for small and large disks when observers must use rivalry as the criterion. Using rivalry as a criterion for evaluating fusion resulted in a more subtle change in fusion responses for one observer (BC). Nevertheless, this experiment demonstrates that fusion limits can be altered depending on which criterion is employed and generalizes the effects shown in experiments 1 and 2 to stimuli that are comparable with those used by Roumes et al. (1997).