Some psychophysical tasks measure ocular dominance plasticity more reliably than others

Seung Hyun Min McGill Vision Research, Department of Ophthalmology and Visual Sciences, McGill University, Montreal, Canada

Ling Gong School of Ophthalmology & Optometry and Eye Hospital, and State Key Laboratory of Ophthalmology, Optometry and Vision Science, Wenzhou Medical University

Alex S. Baldwin McGill Vision Research, Department of Ophthalmology and Visual Sciences, McGill University, Montreal, Canada

Alexandre Reynaud McGill Vision Research, Department of Ophthalmology and Visual Sciences, McGill University, Montreal, Canada

Zhifen He School of Ophthalmology & Optometry and Eye Hospital, and State Key Laboratory of Ophthalmology, Optometry and Vision Science, Wenzhou Medical University

Jiawei Zhou School of Ophthalmology & Optometry and Eye Hospital, and State Key Laboratory of Ophthalmology, Optometry and Vision Science, Wenzhou Medical University

Robert F. Hess McGill Vision Research, Department of Ophthalmology and Visual Sciences, McGill University, Montreal, Canada

In the recent decade, studies have shown that short-term monocular deprivation strengthens the deprived eye's contribution to binocular vision. However, the magnitude of the change in eye dominance after monocular deprivation (i.e., the patching effect) has been found to be different between different methods and within the same method. There are three possible explanations for the discrepancy. First, the mechanisms underlying the patching effect that are probed by different measurement tasks might exist at different neural sites. Second, the test–retest variability of the same test can produce inconsistent results. Third, the magnitude of the patching effect itself within the same observer can vary across separate days or experimental sessions. To explore these possibilities, we assessed the test–retest reliability of the three most commonly used tasks (binocular rivalry, binocular combination, and dichoptic masking) and the repeatability of the shift in eye dominance after short-term monocular deprivation for each of the task. Two variations for binocular phase combination were used, at one and many contrasts of the stimuli. Also, two variations for dichoptic masking were employed; the orientation of the mask grating was either horizontal or vertical. Thus, five different tasks were evaluated. We hoped to resolve some of the inconsistencies reported in the literature concerning this form of visual plasticity. In this study, we also aimed to recommend a measurement method that would allow us to better understand its physiological basis and the underpinning of visual disorders.

Introduction

In the recent decade, there has been increasing evidence that a new form of temporary binocular plasticity exists in human adults. For instance, patching an eye for a short period strengthens that eye's contribution to binocular vision (Lunghi, Burr et al., 2011; Zhou, Clavagnier et al., 2013). This change has

Citation: Min, S. H., Gong, L., Baldwin, A. S., Reynaud, A., He, Z., Zhou, J., & Hess, R. F. (2021). Some psychophysical tasks measure ocular dominance plasticity more reliably than others. *Journal of Vision*, *21*(8):20, 1–23, https://doi.org/10.1167/jov.21.8.20.

https://doi.org/10.1167/jov.21.8.20 Received October 7, 2020; published August 19, 2021 ISSN 1534-7362

been demonstrated for a patching duration as short as 15 minutes (Kim, Kim et al., 2017; Min, Baldwin et al., 2018). Here, we refer to this neuroplastic change in ocular dominance as a result of short-term monocular deprivation as the patching effect. The patching effect lasts for 30 to 90 minutes (Finn, Baldwin et al., 2019; Lunghi, Burr et al., 2011; Min, Baldwin et al., 2018). It can be induced by both opaque and translucent patches, and by dichoptic video presentation (Bai, Dong et al., 2017; Zhou, Reynaud et al., 2014). The patching effect has been demonstrated with psychophysical, electrophysiological (Lunghi, Berchicci et al., 2015; Zhou, Baker et al., 2015), and neuroimaging (Binda, Kurzawski et al., 2018; Chadnova, Reynaud et al., 2017; Lunghi, Emir et al., 2015) studies. The change in sensory eye dominance as a result of short-term patching seems to be reciprocal between the eyes: the contrast gain of the patched eye is enhanced and that of the non-patched eye weakened (Begum and Tso 2016; Chadnova, Reynaud et al., 2017; Reynaud, Blaize et al., 2020; Reynaud, Roux et al., 2018; Zhou, Clavagnier et al., 2013).

In general, studies agree that short-term patching enhances the contribution of the deprived eye to binocular vision. However, the magnitude of the patching effect has been found to be different. For instance, inconsistent results have been found between different methods and within the same method. There are three possible explanations for the discrepancy. First, the patching effect might be a complex phenomenon rather than a change in a single factor (e.g., an increase in one eye's input gain). In other words, mechanisms underlying the patching effect that are probed by different measurement tasks might exist at different neural sites. For example, the removal of phase information induces the patching effect if it is measured with a binocular rivalry task (Bai, Dong et al., 2017), but not so with a binocular combination task (Bai, Dong et al., 2017; Zhou, Reynaud et al., 2014). Moreover, the patching effect has been shown to be greater and longer lasting in the chromatic visual pathway than in the achromatic visual pathway if it is measured with binocular rivalry (Lunghi, Burr et al., 2013), but not so with binocular combination (Zhou, Reynaud et al., 2017). Furthermore, the site of action is believed to be at an early stage (i.e., striate) in cortical processing by some groups (Begum and Tso 2016; Reynaud, Roux et al., 2018; Tso, Miller et al., 2017; Zhou, Reynaud et al., 2014) and at a later stage (i.e., extrastriate) by others (Bai, Dong et al., 2017; Kim, Kim et al., 2017; Ramamurthy & Blaser 2018). Second, the test–retest variability of one method might yield inconsistent data (Finn, Baldwin et al., 2019; Lunghi & Sale 2015). Third, the patching effect itself in the same subject might fluctuate across separate days or experimental sessions. This possibility has not been explored in the literature.

Some studies have measured the effect of short-term patching for each subject and experimental condition without repeating the entire experiment. This practice assumes that the respective psychophysical methodology is reliable and that the patching effect is consistent across days for each subject. In this study, we question this assumption. We repeat all of our experiments using each task twice on separate days. The test–retest reliability of the three most commonly used tasks (binocular rivalry, binocular combination, and dichoptic masking) and the repeatability of the patching effect for each of the task is evaluated. Two variations for binocular phase combination are used, at one (Zhou, Clavagnier et al., 2013) and many contrasts of the stimuli (Min, Baldwin et al., 2018). Also, two variations of the dichoptic masking task are tested, in which the orientation of the mask grating is either horizontal or vertical (Baldwin & Hess, 2018). Thus, there are five different measurement methods in all. We hope to resolve some of the inconsistencies reported in the literature concerning this form of visual plasticity. We also aim to recommend a measurement method that will allow us to better understand its physiological basis and the underpinning of visual disorders. To do so, we assess five properties of each task:

1. **Baseline reliability**: How well is the baseline performance (i.e., no patching) correlated for each subject between repeated experiments?
2. **Patching effect reliability**: How well is the magnitude of the patching effect correlated for each subject between repeated experiments?
3. **Baseline measurement variability**: What is the expected measurement variability from the task alone, and how does this compare to the overall variability in the baseline conditions?
4. **Patching effect measurement variability**: What is the expected measurement variability from the task alone in the patched conditions, and how does this compare to the overall?
5. **Detectability of the patching effect**: How effective is the task in detecting the patching effect?

Methods

Subjects

Data of 88 adults (age range = 18–33 years) with normal or corrected-to-normal vision were included in this study. The data from 62 subjects have already been reported in publications (Baldwin & Hess, 2018; Finn, Baldwin et al., 2019; Min, Baldwin et al., 2018; Min, Baldwin et al., 2019). For this study alone, we recruited 26 additional subjects. Four subjects participated in multiple experiments. So, there were 92 unique data

points total. This study adhered to the Declaration of Helsinki and was approved by the Institutional Review Boards at McGill University and Wenzhou Medical University. All subjects provided informed written consent.

The issue that this study addresses is to see whether there is minimal difference in data across two repeated sessions. For this reason, a power analysis was not used to determine our sample size because we did not expect to see a statistically significant difference between two repeated experiments. The difference between two experiments could be statistically insignificant, and yet be just large enough to decrease the replicability of the task. Moreover, we did not introduce the effect of treatment on one of the two groups. Since many laboratory groups recruited between 10 and 20 subjects for an experimental condition, we decided that 15 subjects per task would be sufficient. For all methods, subjects were trained extensively before they began the actual experiment and repeated the experiment on a separate day (each session separated by 24 hours) at a similar time.

Monocular deprivation

In all experiments, the dominant eye of each observer was deprived with a translucent patch, which removed all form information and decreased the luminance by 20%. The eye dominance was determined by the Miles test (Miles, 1930). In this test, the subjects were asked to form a peephole with their index finger and the thumb. After placing a visual target within the peephole at arm's length, they alternatively closed each eye. When the dominant eye was closed, the visual target was displaced more within the peephole. For some psychophysical tasks, we tested different patching durations (30, 120, and 150 minutes). Subjects repeated each experiment twice (i.e., two sessions of the same patching duration) on separate days. During patching, subjects browsed the web with either their computer or phone. We were only interested in the immediate patching effect (within 10 minutes), so we did not test the patching effect long after patch removal.

Psychophysical tasks

In this study, we evaluated five psychophysical tasks. Each task is described in detail in this section. Moreover, we extracted a subset of data from four published studies (Baldwin & Hess, 2018; Finn, Baldwin et al., 2019; Min, Baldwin et al., 2018; Min, Baldwin et al., 2019). We additionally recruited 26 unique subjects for three experimental tasks (see Figures 1 and 2; $n = 15$ per task). In this section, we elaborate on the rationale for the data extraction, the process of data analysis, and the experimental procedure for each psychophysical method.

Binocular rivalry

In this method, non-fusible stimuli were shown to the two eyes. The relative strength of each eye was assessed by measuring the length of time for which each eye suppresses the other. Data from 30 subjects were collected from a previous study (Finn, Baldwin et al., 2019), from which we extracted the baseline measurements. An additional 15 subjects were then tested as a part of the current study. Therefore, data from 45 subjects were included in the binocular rivalry analysis.

Stimuli: In the binocular rivalry task used in the study of Finn et al. (2019), two oblique Gabor patches at +45° and −45° were shown separately to the two eyes. The experiment randomly assigned the two orientation of the Gabor patches to remove orientation bias. The Gabor patches had a spatial frequency of 1.5 c/deg, a spatial sigma of 1.3° of visual angle and a contrast of 50%. Shutter glasses were used for the stimulus presentation. Each test block lasted for 180 seconds. Subjects reported continuously using the keyboard whether they perceived a left oblique grating, right oblique grating, or mixed percept throughout the test.

In the new experiment, two orthogonal Gabor patches (0.46 cycle/deg, 4.33° × 4.33°) were dichoptically presented to the two eyes using head-mount goggles (details in the Methods section entitled *Apparatus in the previous studies*). The contrast of the Gabor presented on the non-patched eye was fixed at 80%, and that of the Gabor on the eye to be patched was set so that each subject perceived an equal visibility of the Gabor patches between the two eyes. In short, the ratio of the duration between the eyes was close to 1 after adjusting the contrast for the eye to be patched. This contrast was used throughout the experiment on the same day and was individually established for each subject. The contrast was reset on the second day of the experiment (i.e., second session). Each testing block had two segments of a 90-second trial. Therefore, a single test block lasted for 180 seconds. In the first segment, the orientation of the Gabor was −45° in the non-patched eye and 45° in the patched eye. In the second segment, the orientation of the Gabor was +45° in the non-patched eye and −45° in the patched eye. Subjects were asked to report continuously using the keyboard whether they perceived a left-tilted, right-tilted, or mixed Gabor throughout the test.

Each eye was displayed with one of the two possible orientations of the Gabor patches. Hence, we processed the data by designating the two Gabor gratings at different orientations to each eye's percept (patched or non-patched), and then computed ocular dominance

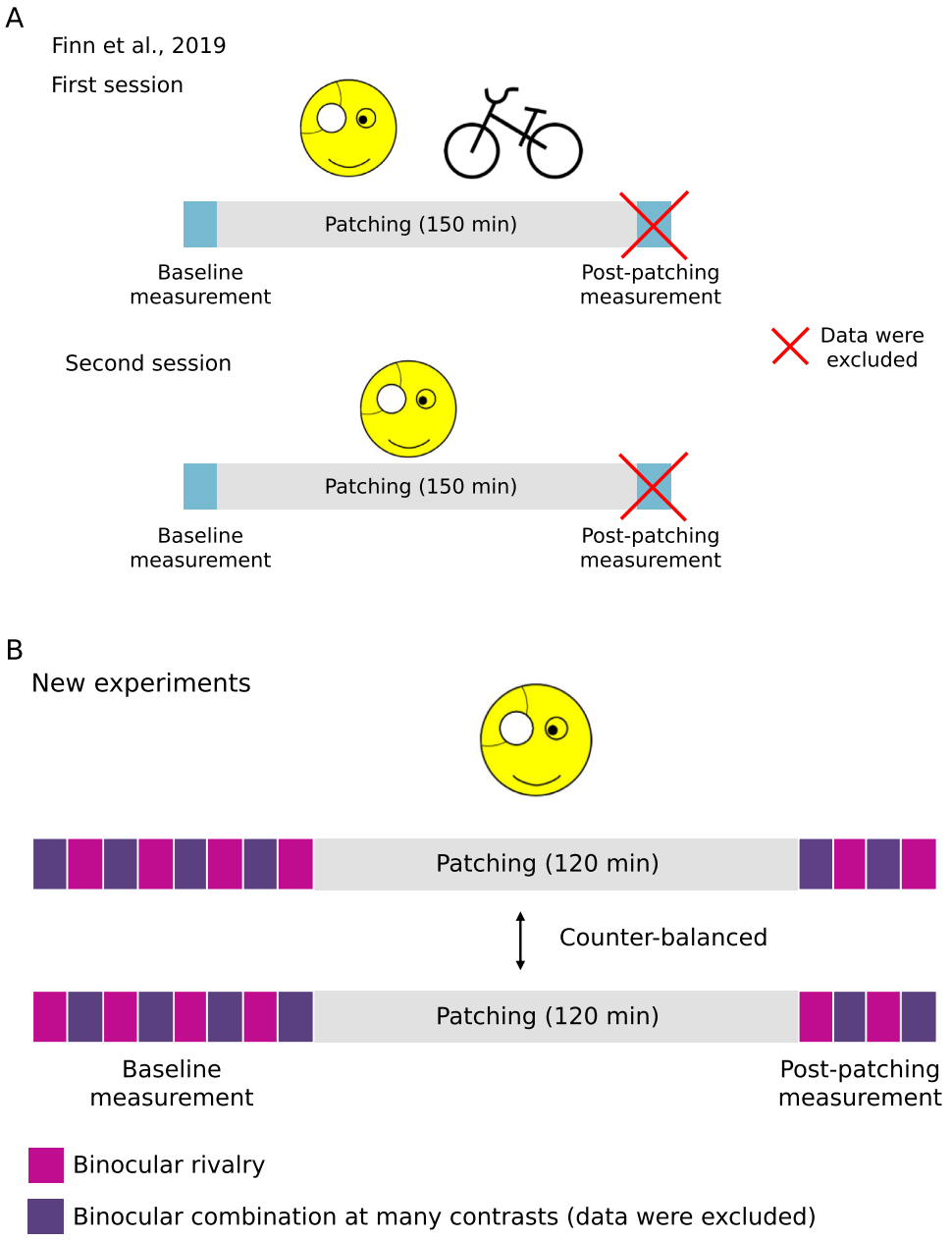

Figure 1. Procedures of experiments using binocular rivalry. (A) Procedure of the experiment in the study of Finn et al. (2019). (B) Procedure of the new experiments in our study.

index (ODI), which represents sensory eye dominance (see below).

Procedure: In the study by Finn et al. (2019), the patching effect was measured in two different experimental conditions. The goal of the study was to examine whether exercise during patching potentiated the patching effect. Since the experimental conditions were not identical, we could not use the postpatching data for our analysis. However, since the baseline measurements made on the two testing days were identical, we included the baseline data in our data analysis ($n = 30$).

However, because we also wanted to evaluate the repeatability of the patching effect, we tested 15 additional subjects. The subjects first performed the baseline measurement during which the binocular rivalry task was performed four times (Figure 1). The binocular rivalry task was interleaved with a binocular combination task (the data from the combination task were not used for analysis owing to a technical mishap). This strategy was used to make the procedure here more comparable with that used to compare the two forms of the combination task (as described in the next section). The baseline tests, therefore,

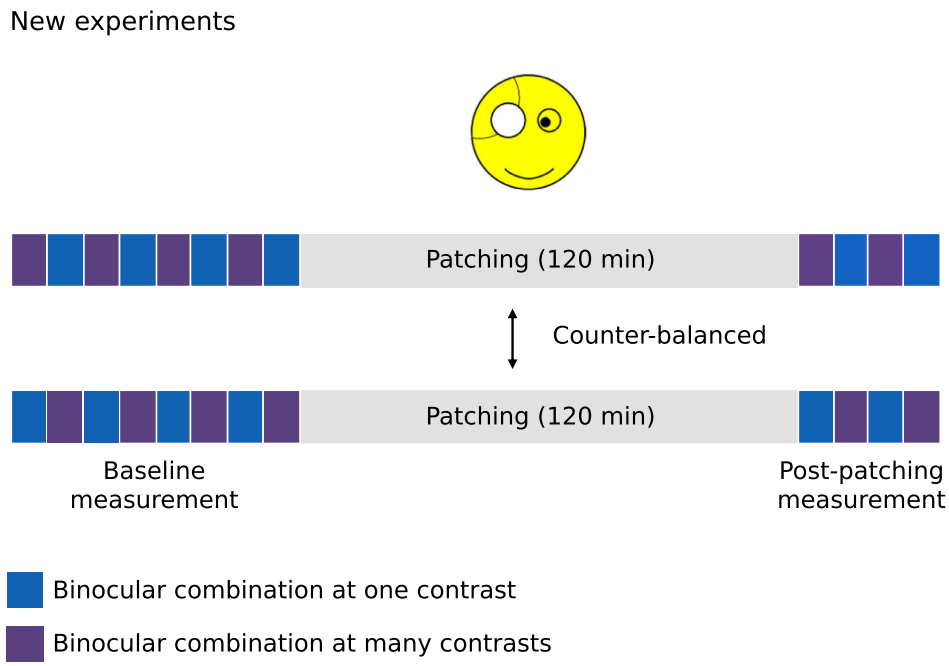

Figure 2. Procedure of the new experiments using, but not limited to, binocular combination at one contrast.

consisted of four experimental blocks of binocular combination and binocular rivalry tasks. After patching for 120 minutes, the subjects were tested again using binocular combination and binocular rivalry for two experimental blocks (two blocks per task).
Data analysis: We computed the ODI as follows:

$$ODI = \frac{d_p - d_n}{d_p + d_n + d_m}, \quad (1)$$

where d_p, d_n, and d_m are the total response durations of the percept perceived by the patched eye, the non-patched eye, and both eyes (i.e., mixed percept), respectively. When the ODI is positive, the total response duration for the percept perceived by the patched eye is longer than that for the non-patched eye's percept. When the ODI is negative, the total response duration for the percept perceived by the non-patched eye is longer than that by the patched eye.

Binocular phase combination task at one contrast

To assess the test–retest reliability of this task, we recruited additional 15 subjects because we had not had any data to extract from previous studies.
Stimuli: Two fusible, separate, and horizontal sine-wave gratings (0.46 cycle/°, 4.33° × 4.33°) with equal and opposite phase shifts (+22.5° and −22.5°) relative to the center of the screen were presented to the two eyes. The perceived phase of fused stimuli was 0° if the two eyes contributed equally to binocular fusion (Figure 5). The subjects were asked to locate their perceived middle portion of the dark patch in the fused grating by positioning a flanking 1-pixel reference line. The stimuli were displayed until subjects completed the tasks. The contrast of the stimuli shown to the non-dominant eye (i.e., non-patched eye) was set at 100% for each subject. Moreover, the contrast of the stimuli shown to the dominant eye (i.e., patched eye) was set so that both eyes contributed equally to binocular vision (i.e., binocularly perceived phase = 0°). The contrast of the stimuli shown to the non-dominant eye was not uniform across subjects. Therefore, there was only one contrast ratio between the stimuli shown separately to the eyes for every subject.
Procedure: As Figure 2 shows, the experimental protocol is identical to the interleaved design described in Figure 1. The subjects performed baseline measurements with psychophysical tasks of binocular combination at one and multiple contrasts (another variation of binocular phase combination, described in the section entitled *Binocular Phase Combination Task at Many Contrasts*). They completed four test blocks of the two different binocular combination tasks (four blocks per task). Each block lasted for approximately 3 to 5 minutes. Then they were patched for 120 minutes. During patching, they performed tasks such as reading and web browsing. After patching, they were tested again using the two methods of binocular combinations

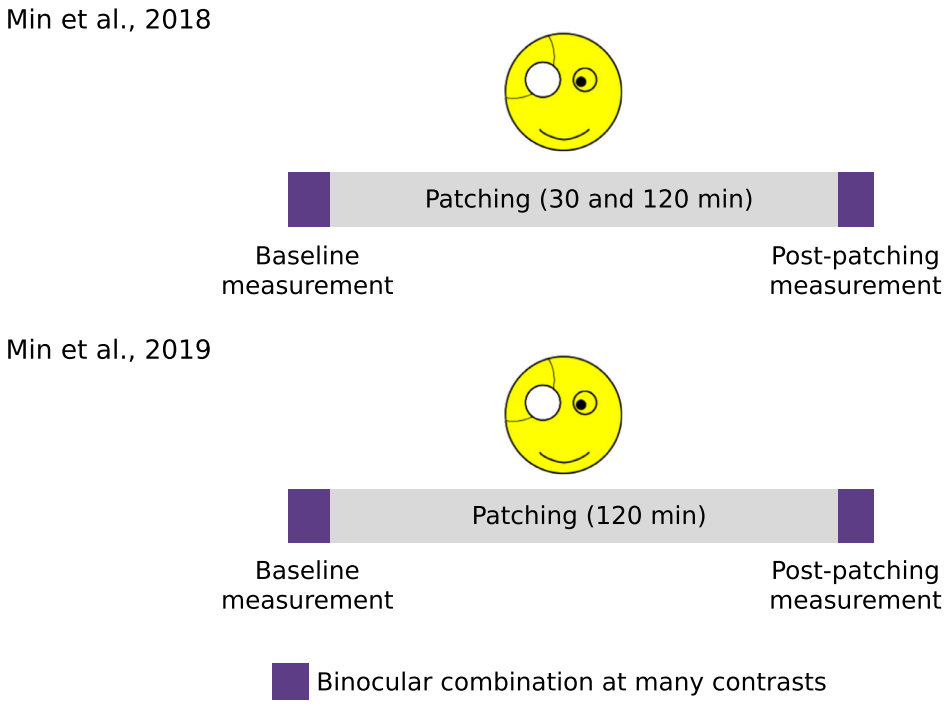

Figure 3. Procedure of experiments using binocular combination at many contrasts.

for two experimental blocks (Figure 2). We randomized the order of the task to be tested and maintained the order across two repeated experiments for each subject.

Binocular phase combination task at many contrasts

Data from 19 subjects were extracted from previous studies (Min, Baldwin et al., 2018; Min, Baldwin et al., 2019). These participants had been patched for 30 or 120 minutes. Fifteen more subjects were additionally recruited (Figure 2) and were patched for 120 minutes. In sum, there are 34 unique data points.

Stimuli: The stimuli were very similar to those in binocular combination at one contrast. Two slightly offset horizontal sinusoidal gratings were presented to the two eyes. The phase difference was 45°: +22.5° for one eye and −22.5° for the other eye. If the two eyes contributed equally to binocular vision, the fused phase percept appeared as exactly the average of the two gratings phases. This was equivalent to the perceived phase of zero (Figure 5).

The interocular contrast ratio between the eyes was changed by increasing the contrast of one eye's stimulus while decreasing the contrast of the other eye's stimulus (Figure 1). Then, the interocular contrast ratio at a perceived phase of 0° was estimated using a contrast gain model (Ding & Sperling, 2006). By comparing the binocular balance before and after patching, we calculated the shift in ocular dominance.

We set five interocular contrast ratios (1/2, $1/\sqrt{2}$, 1, $\sqrt{2}$, 2) for baseline measurement, and three for postpatching measurement ($1/\sqrt{2}$, 1, $\sqrt{2}$). A baseline test block took about 5 minutes to complete, whereas the postpatching test block took 3 minutes. On the other hand, in the binocular phase combination at one contrast task (discussed in the previous section), only a single ratio (i.e., 1) was used.

Procedure: From two previous studies (Min, Baldwin et al., 2018; Min, Baldwin et al., 2019), we extracted data of 19 subjects who had been patched for two patching durations (30 and 120 minutes). We discarded remaining data of the participants who had been patched for other durations (from Min et al., 2018) to not violate the assumption of independence. That is, each data point could only be used once in the data analysis. Before patching, the subjects performed the baseline experiments (Figure 3). After patching for an assigned duration, they completed postpatching experiments at several timepoints between 0 and 48 or 96 minutes after patching. All subjects repeated the experiment twice. Therefore, we were able to include data from baseline and postpatching assessments to evaluate the test–retest repeatability of the task. We only extracted postpatching data at the first three measured postpatching timepoints (0 to 6 minutes) and averaged the values.

As described elsewhere in this article, we tested 15 more subjects to compare the test–retest repeatability directly between the two variations of binocular phase combination. Data had been first collected previously in the procedure described elsewhere in this article (Figure 1). We had first designed the experiment

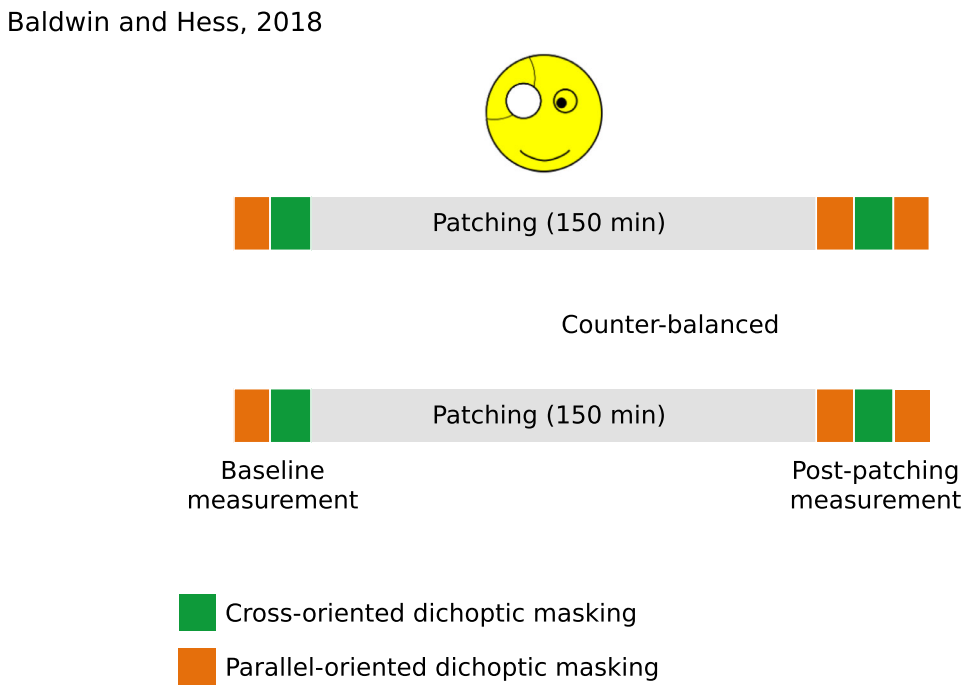

Figure 4. Procedure of experiments using dichoptic masking. The figure has been adapted from the previous study by Baldwin and Hess (2018).

to directly compare between binocular rivalry and combinations at multiple contrasts. However, owing to the improper display of the gratings for binocular combination that we found out after we had finished collecting the data, we decided to discard the data of binocular combination and keep those of binocular rivalry. After resolving the screen issue, we decided to maintain a comparable task design by interleaving two different tasks in the same manner as the procedure described elsewhere in this article (Figure 1). Therefore, we included a binocular phase combination at one contrast, and interleaved it with binocular phase combination at many contrasts (Figure 2).

Data analysis: We averaged the perceived phases across two configurations from each subject. We then fitted these means of perceived phases into a contrast gain control model introduced by Ding and Sperling (2006):

$$\Phi_A = 2tan^{-1}\left[\frac{f(\alpha,\beta,\gamma)-\delta^{1+\gamma}}{f(\alpha,\beta,\gamma)+\delta^{1+\gamma}}tan\left(\frac{\theta}{2}\right)\right], \quad (2)$$

where

$$f(\alpha,\beta,\gamma) = \frac{1+\delta^{\gamma}}{1+\alpha\delta^{\gamma}}, \quad (3)$$

Φ_A = perceived phase from the fused percept of two stimuli, α = gain factor which determines the contrast balance ratio when both eyes contribute equally to binocular vision, γ = slope of the function when both eyes contribute equally to binocular vision, θ = fixed phase displacement between eyes (45°), and δ = interocular contrast balance ratio. After we fitted our data to the contrast gain model function, we estimated the two free parameters, α and γ. We bootstrapped responses trial-to-trial and generated each measurement's sample of α values to generate standard errors for each data point.

α was transformed into log scale as following:

$$\alpha_{dB} = 20 \times log_{10}(\alpha_{ratio}), \quad (4)$$

where

$$\alpha_{ratio} = \frac{\alpha_{DE}}{\alpha_{NDE}}. \quad (5)$$

α_{ratio} = contrast balance ratio when both eyes contribute equally to binocular vision in linear scale and $\alpha_{dB} = \alpha_{ratio}$ in log scale. When the contrast shown to the dominant eye is as twice as strong as the non-dominant to reach the balance point ($\alpha_{DE} = 2\alpha_{NDE}$), then the $\alpha_{ratio} = 2$, thereby resulting in $\alpha_{dB} = 6dB$.

We converted α_{ratio} into α_{dB} to avoid bias for the dominant eye when we quantify binocular balance. We normalized the contrast balance ratios by calculating for the differences in contrast balance ratios between baseline and after patching (dB). Therefore, when Δ

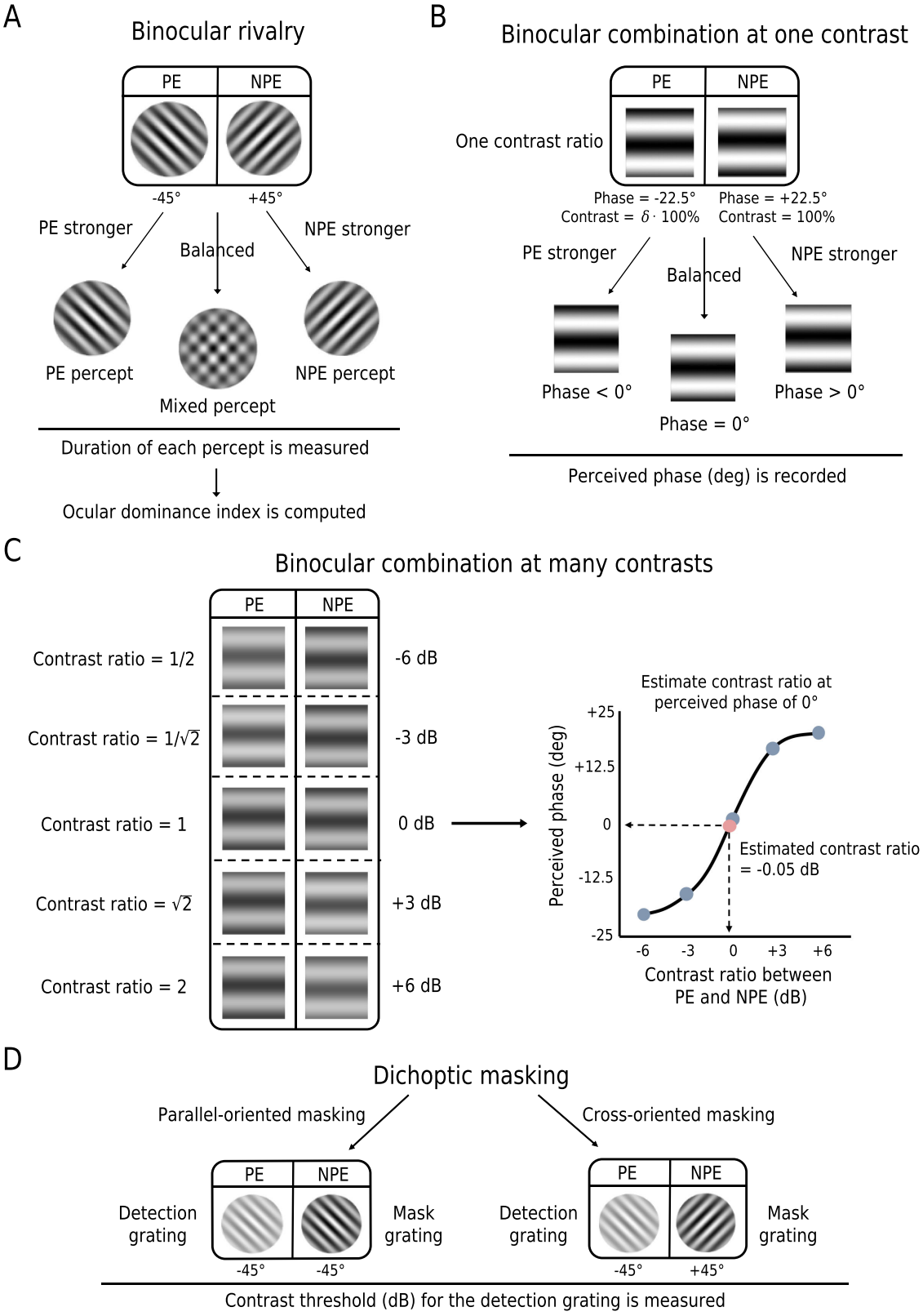

Figure 5. An illustration of stimuli in the five psychophysical task variations. PE = patched eye; NPE = non-patched eye. (A) Binocular rivalry. Two gratings in different orientations were shown separately to both eyes. When the patched eye was dominant, the grating shown to the patched eye would dominate the conscious visual awareness. (B) Binocular phase combination at one contrast. Two fusible gratings were shown dichoptically. Subjects were asked to locate using the keyboard the center of the darkest strip within the middle segment of the fused grating. (C) Binocular combination at many contrasts. Two fusible gratings were shown separately to both eyes. Subjects were asked to locate using the keyboard the center of the darkest strip within the middle segment of the fused

→

←
grating. Five contrast ratios were tested for baseline. Three contrast ratios were used for postpatching measurement. (D) Dichoptic masking. The subjects were asked to detect in which of two intervals the detection grating appeared. Two types of dichoptic mask were used. The parallel mask had the same orientation as the target. The cross-oriented mask had an orthogonal orientation.

contrast balance ratio = 0, it represents no change after patching, whereas a positive Δ contrast balance ratio indicates the shifting of ocular dominance favors the dominance eye (the patched eye).

Parallel- and cross-oriented dichoptic masking task

Data from 14 subjects were extracted from a previous study (Baldwin & Hess, 2018). No additional subjects were tested.

Stimuli: One sinusoidal grating of 0.5 c/deg was presented to each eye. Gratings were presented in a circular raised-cosine envelope. The diameter was 5° of visual angle. The temporal envelope for presenting the gratings was a Gabor (temporal frequency of 2 Hz, duration sigma 500 ms). The contrast in log units (dB) was computed as:

$$c_{dB} = 20 \times log_{10}(c_{\%}).$$

A contrast of 1% translates to 0 dB. A two-fold threshold elevation from masking gives a 6 dB difference between detection thresholds with and without the mask.

The experiment used a two-interval forced choice procedure. Contrast detection thresholds were measured under three conditions: (i) monocularly in the eye to be patched (no mask), (ii) monocularly in the eye to be patched with a dichoptic mask grating shown to the other eye that had the same orientation as the target (parallel), (iii) similar to (ii), but with the mask having an orthogonal orientation (if the left eye's grating were 45°, the right eye's grating would be −45°). The mask contrast was fixed at 4%. When a mask was shown, it would be presented to the non-patched eye in both intervals. In only one of the intervals, the target grating would be shown (to the patched eye). The subject reported the interval (first or second) in which the target grating was presented.

Procedure: During baseline measurement, we measured the detection threshold of the patched eye and that of the patched eye when the mask grating was shown to the non-patched eye (i.e., masked threshold) in two different orientations (parallel and cross). Then the dominant eye was patched for 150 minutes. After patch removal, subjects were asked to immediately perform three blocks of post-patching measurements. The post-patch tests included three test blocks and measured the masked threshold of the patched eye. The sequence of one testing block was either parallel–cross–parallel or cross–parallel–cross for the mask orientation. Each testing block lasted for about 5 minutes. All subjects completed both sequences in a randomized order across the two repeated experiments. The sequence order of the post-test was counterbalanced because the shift in eye dominance after patching would decay over time.

Apparatus for the new experiments

For our new experiments, we measured changes in eye balance after patching using binocular rivalry, binocular combination at one contrast, and binocular combination at many contrasts. We set up the tasks in MATLAB 2012a using PsychToolBox 3.0.9 (Kleiner, Brainard et al., 2007; Pelli, 1997). We presented the stimuli on a Mac computer with gamma-corrected head-mounted goggles (NED Optics Groove pro, OLED). They had a refresh rate of 60 Hz and resolution of 1920 × 1080 to the screen for each eye. The maximum luminance of the goggles was 150 cd/m^2.

Apparatus in the previous studies

Binocular rivalry

During the rivalry task, the gratings were displayed on a projector screen at 2.3 m from the subjects by an Optoma HD26 DLP projector (Finn et al., 2019). The subjects wore a pair of Optoma ZD302 DLP Link Active Shutter 3-dimensional glasses so that the gratings would be displayed dichoptically. For every degree of visual angle, there were 75 pixels in the resolution of the projector. The mean luminance of the screen was set at 95 cd/m^2. The experiment was set up in MATLAB and PsychToolBox (Kleiner, Brainard et al., 2007; Pelli, 1997).

Binocular combination task at many contrasts

The gratings were displayed dichoptically using head-mounted goggles with a refresh rate of 60 Hz, a resolution of 800 × 600 pixels, and a mean luminance of 59 cd/m^2 (Min et al., 2018; Min et al., 2019). For all subjects tested in Min et al. (2018) and for five of the 10 subjects tested in Min et al. (2019), the stimuli were displayed through the eMargin Z800 pro goggles. However, owing to the equipment failure, GOOVIS Cinego G2 goggles were used for the remaining five subjects. These goggles had a resolution of 1920 × 1080 pixels, a refresh rate of 60 Hz, and a mean luminance of 60 cd/m^2.

Dichoptic masking tasks

The detection and mask gratings were displayed on a gamma-corrected Clinton Monoray CRT monitor with a resolution of 800 × 600 pixels and a refresh rate of 150 Hz (Baldwin & Hess, 2018). The subjects completed the task at a viewing distance of 70 cm. There were 27 pixels per degree of visual angle at this viewing distance. To dichoptically display the stimulus, a ViSaGe (Cambridge Research Systems Ltd., Kent, UK) was implemented using FE-1 ferro-electric shutter goggles. The goggles had a refresh rate of 75 Hz.

Standardized data analysis

Data were analyzed using R and Python. Since the five methods have different units, we standardized the raw data into z-scores for each dataset. For instance, z-scores were computed for the dataset of the first session using binocular rivalry for baseline measurement. A z-score of 0 would indicate data that are identical to the mean of the particular dataset (such as our example here). A z-score of 1 would denote that data are 1 standard deviation away from the mean of a particular dataset. The z-score was calculated with this formula:

$$z = \frac{x - \mu}{\sigma}$$

where x is the raw data, μ is the mean of the sample, and σ is the standard deviation of the sample.

The results from each task are analyzed in a similar way. Below we describe each column of our figures in the Results (Figures 6 and 7).

Column (i): Baseline and patching effect reliabilities

To assess test–retest repeatability, Pearson's correlation was calculated using raw data. A strong correlation indicates that a subject's performance from the first experimental session is a good predictor of that in the second session. In this column, figures also show the conversion of raw data into z-scores. Correlation, however, does not guarantee replicability of data. A few extreme points can determine the fate of a correlation. Also, when the means of two samples are significantly different, the data from these samples can still have a strong correlation. Therefore, a strong correlation ($r > 0.7$) does not directly mean that the test–retest replicability is superior. Column (ii), which is a series of Bland–Altman plots, aims to address the inadequacy of correlation.

Column (ii): Baseline and patching effect measurement variabilities

Since correlation is not sufficient to test for replicability, Bland–Altman plots are plotted in column (ii) with the z-score. They illustrate the measurement variability (i.e., test–retest replicability) of either baseline or the patching effect. The y-axis is the difference between the z-scores from the first and second experiments (i.e., sessions). The x-axis is the mean z-score across the two sessions. The mean difference of the z-score between the two days (across subjects) is indicated by the central horizontal dashed line. He 95% limits of agreement are shown by the upper and lower dashed lines; they represent the range within which the difference is most likely to fall for most observers. The wider the limits of agreement, the larger the measurement variability between the tasks. The mean difference (i.e., middle dashed line) is always set to 0 because all the raw data are converted to z-score. Mathematically, the mean of z-scores from one sample has to be 0. Hence, the mean difference of z-scores between two samples also has to be 0.

The two experimental sessions were separated by at least 24 hours. So, we reasoned that the variability indicated by the outer dashed lines can arise from various factors. The first of these could be the measurement error from the task design and testing procedure. The second could be the day-to-day variability in the measured physiological mechanism. In our case, the former was of greater interest. For this reason, we estimated the first of these factors by computing the expected standard error that arose from only the psychophysical task of interest. To obtain the standard error for each task, the median of the standard error from each testing block of the task was obtained either directly from testing (binocular phase combination at one contrast) or estimated by bootstrapping. This was the standard error for a single measure. However, because the Bland–Altman plots analyze the difference between two measurements, the standard errors of both needed to be accounted for. So, we normalized the single standard error by multiplying it by $\sqrt{2}$. To convert this difference standard error to a 95% confidence interval, we multiplied it by 1.96. In short, we calculated the range between the mean of the differences between the two sessions and the expected 95% confidence interval from the measurements. Finally, this result was normalized into the z-score because the Bland–Altman plots were plotted in z-scores. We subsequently shaded this range in grey (Figures 6 and 7). This shaded grey region represents the expected measurement variability from the psychophysical task itself. In short, the narrower the grey region, the better the test–retest replicability of the test. If the range enclosed by the dashed lines indicating the limits of agreement is wider than the shaded region (i.e., measurement variability), then an additional source of variability beyond the measurement alone exists.

Column (iii): Baseline and patching effect correlations

Finally, whether the performance of a single subject across experimental sessions was significantly more correlated than a mismatched pair of subjects was evaluated. To do so, the correlation coefficient was computed from two samples. The first sample was the first session of all subjects (i.e., orderly sample) and the second sample was a randomly sampled data from the second session of all subjects. The resampling of the second sample created a mismatched pair of subjects. If these samples are correlated, then the mismatch will destroy their linear relationship. The second sample was resampled 1000 times, so we were able to compute 1000 correlation coefficients, most of which had a weak linear relationship. The histogram distribution of the correlation coefficients from random sampling is plotted in column (iii) from Figures 6 and 7. Also, the actual correlation coefficient from column (i) is marked in the histogram. If the actual correlation is robust, then the correlation coefficient will be located toward the outer edge of the histogram. However, if the correlation is weak, then the correlation coefficient will reside within the histogram proper.

Results

Baseline measurement

To assess the test–retest variability of the psychophysical tasks, we incorporated data from baseline measurement into our data analysis. Each subject performed two experimental sessions that were separated by at least 24 hours.

Binocular rivalry

For a typical measurement of binocular rivalry, the ODI indicates the relative length of the percepts (patched or non-patched eye) shown separately to both eyes during one test block.

Pearson's correlation was calculated to assess whether the baseline performance of a subject in one day was correlated to that of the same subject from another day. The correlation was not significant ($n = 45$, $r = 0.19$, $p = 0.204$; Figure 6A(i)). Next, the raw data of ODI were converted into the z-score for standardization.

To see if there was a good agreement between the two experimental sessions, we created a Bland–Altman plot. Figure 6A(ii) indicates that the 95% limits of agreement are ±2.49 (z-scores). The limits of agreement (dashed lines) represent the test–retest variability that originate from multiple factors, such as day-to-day variability between the two experimental sessions and the variability from the psychophysical measurement itself. Therefore, we computed the measurement variability of binocular rivalry, which is the median of the bootstrapped standard errors for each test block from baseline measurement. This range, which is shown as a grey shaded area in Figure 6B(ii), is ±1.69 (z-scores). Most of the area within the limits of agreement (i.e., dashed lines) is taken up by the shaded region. This suggests that most of the test–retest variability originates from the binocular rivalry measurement itself rather than the variability from physiological factors.

One might be concerned about the noticeable difference in the spread of the points between the data from Finn et al. (light blue diamonds) and our more recent data (pink diamonds; see Figure 6A(i)) and in the spread of the mean difference from the Bland–Altman plot (Figure 6A(ii)). The correlation of the data from our new data is robust ($n = 15$, $r = 0.52$, $p = 0.043$). However, two samples can have a strong correlation even if their means are significantly different. For this reason, the measurement variability is more representative of replicability than the correlation per se. The measurement variability (grey area) of the data from Finn et al. alone is comparable (±1.72 z-scores) to that of the combined (Finn et al. + new data) baseline dataset (±1.69 z-scores). Even if the new dataset has a robust correlation, the measurement variability for each testing block is still large.

Last, we evaluated whether the performance of a subject from the first experimental session was more correlated with that same subject's performance from the second experimental session rather than that from another, randomly selected subject. The distribution of the 1000 sampled correlation coefficients is plotted in the histogram (see Figure 6A(iii)). As we expected from Figure 6A(i), the correlation between the performance scores in both experimental sessions is weak. So, the correlation coefficient from Figure 6A(i) resides close the middle of the histogram. Our histogram indicates that the test–retest difference is so large that there is little to be gain from using a within-subject protocol to make comparisons.

Binocular combination at one contrast

In this task, a phase of 0° indicates that both eyes are contributing equally to binocular vision. Pearson's correlation revealed a weak correlation ($n = 15$, $r = -0.18$, $p = 0.528$) for the baseline data from the binocular combination task at one contrast.

The Bland–Altman plot (Figure 6B(ii)) shows that the limits of agreement are ±3.01 (z-scores). The measurement variability expected only from the task (grey region) is ±2.22 (z-scores). Since the shaded area makes up most of the area within the limits of agreement (dashed lines), most of the test–retest variability originates from the task measurement variability rather than from other factors.

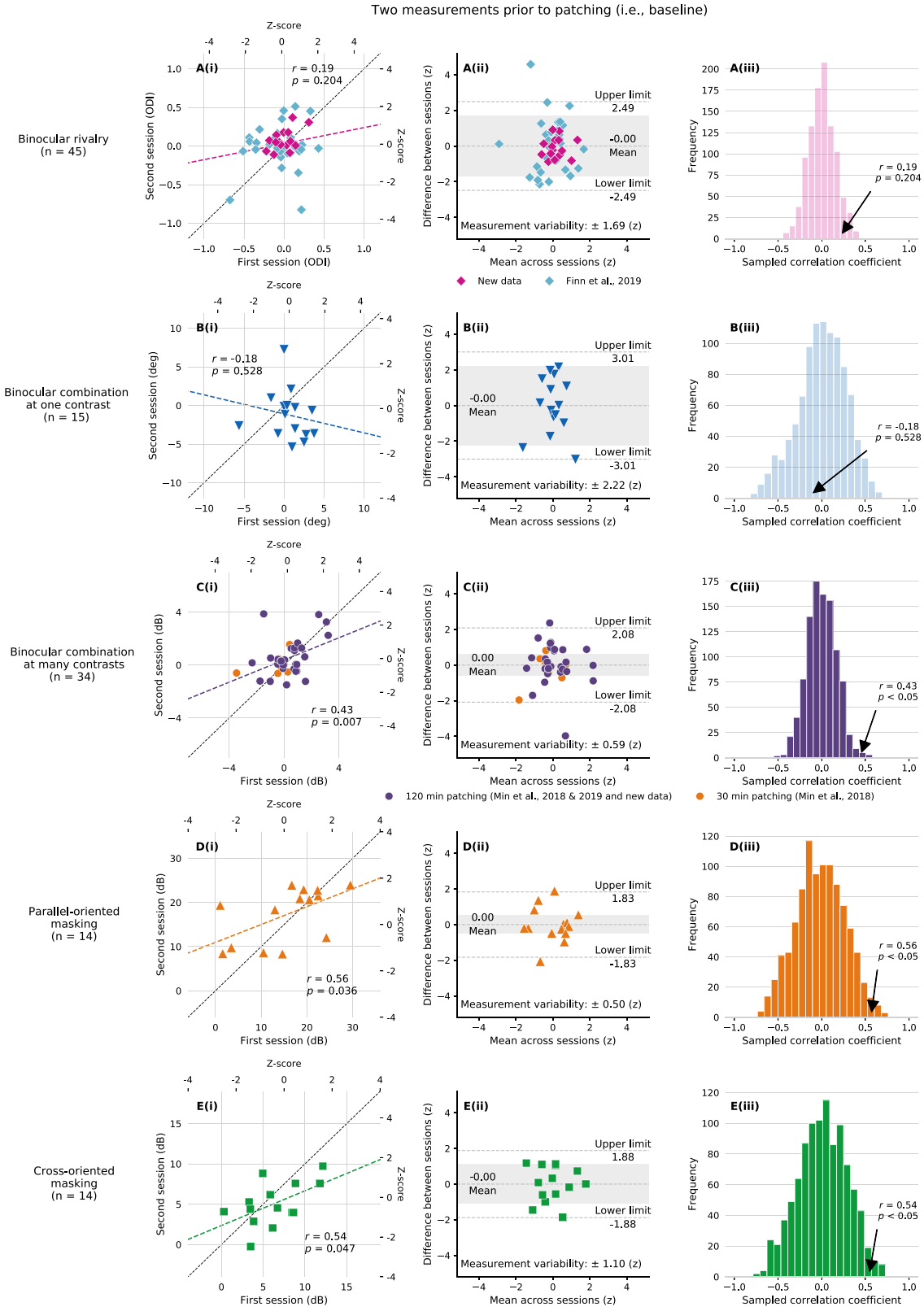

Figure 6. Evaluation of baseline measurement (i.e., no patching) using the five psychophysical tasks. This figure is divided into five rows (task) and three columns (as described in the Standardized Data Analysis section). (A) Binocular rivalry. Pink points represent data from the new experiments, blue points from the study of Finn et al. (2019). (B) Binocular phase combination at one contrast. (C) Binocular phase combination at many contrasts. Different durations of patching are represented in different colors. (D) Parallel-oriented dichoptic masking. (E) Cross-oriented dichoptic masking. Column (i) Baseline reliability. The *x*-axis represents results (e.g., ODI from binocular rivalry) from the first experiment session, and the *y*-axis denotes results from the second session. The secondary

→

←
x- and y-axes represent z-scores from the raw data of ODI. The black dashed line represents the line of equality (first session = second session) and has a slope of 1. The colored dashed line represents the regression line from Pearson's correlation test. Each diamond represents a data point of one subject. Column (ii) Baseline measurement variability in a Bland–Altman plot. Difference in z-scores between the first and second session is plotted as a function of the mean of z-scores across two sessions. The outer horizontal dashed lines indicate 95% limits of agreement. The dashed line in the middle indicates the mean difference of z-scores across the subjects. The gray shaded region within the limits of agreement represents measurement variability of baseline (i.e., the testing variability stemming from only the binocular rivalry task). The unshaded regions within the limits of agreement represent test–retest variability from external factors beside the task itself. Column (iii) Baseline reliability illustrated in a histogram. The sampled reliability coefficients are plotted as a histogram, where the y-axis represents the frequency and the x-axis the sampled correlation coefficient ranging from −1 to 1. The single line value represents the within-subject correlation and this is compared with the distribution of between-subjects correlations.

The sampled correlation coefficients are plotted in histogram (Figure 6B(iii)). The weak correlation coefficient obtained from Figure 6B(i) resides in the middle of the histogram. This finding suggests that the test–retest difference is so considerable that within-subject designs offer little, if any, advantage.

Binocular combination at many contrasts

In this task, 0 dB indicates that both eyes contribute equally to binocular vision. This task is different from the binocular phase combination task at one contrast because it makes measurements at multiple contrast ratios and calculates the shift in ocular dominance using a model.

Pearson's correlation (see Figure 6C(i)) revealed a significant correlation ($n = 34$, r = 0.435, $p = 0.0072$). The Bland–Altman plot (Figure 6C(ii)) indicates that the limits of agreement are ±2.08 (z-scores). The measurement variability (grey shaded area) from the task itself is ±0.59 (z-scores). The shaded area only represents a small fraction of the area within the limits, suggesting that most of the test–retest variability originates from external factors such as day-to-day variability in physiological mechanisms.

Last, the sampled correlation coefficients are plotted in a histogram (Figure 6C(iii)). As observed in Figure 6C(i), the correlation between the performance scores in both experimental sessions is robust. This finding is confirmed in Figure 6C(iii), where the correlation coefficient obtained from Figure 6C(i) resides in the outer edge of the histogram. This finding suggests there is much to be gained from using within-subject testing protocols.

Parallel-oriented dichoptic masking

Pearson's correlation test revealed a significant correlation ($n = 14$, r = 0.56, $p < 0.05$; Figure 6D(i)). A Bland–Altman plot (Figure 6D(ii)) shows that the limits of agreement are ±1.83 (z-scores). The measurement variability (grey shaded area in Figure 6D(ii)) is ±0.50 (z-scores). The shaded area only represents a small fraction of the area within the limits of agreement. This finding suggests that most of the test–retest variability originates from external factors such as day-to-day variability.

Last, the distribution of the sampled correlation coefficients is plotted (see Figure 6D(iii)). As we observed in Figure 6D(i), the correlation between the performance scores in both experimental sessions is robust. This finding is confirmed in Figure 6D(iii), where the correlation coefficient obtained from Figure 6D(i) seems to reside in the outer edge of the histogram. Therefore, there is an advantage from within-subject testing protocols.

Cross-oriented dichoptic masking

Pearson's correlation test found a significant correlation ($n = 14$, r = 0.54, $p < 0.05$; Figure 6E(i)). The Bland–Altman plot (Figure 6E(ii)) shows that the limits of agreement are ±1.88 (z-scores). The measurement variability (grey shaded area in Figure 6E(ii)) is ±1.10 (z-scores). It seems that the larger portion of the areas within the limits of agreement are attributable to the measurement variability from the dichoptic masking task itself rather that from external factors such as day-to-day variability. However, it is notable that the additional area within the limits of agreement that is attributable to external factors is of a similar size.

Last, the distribution of the sampled correlation coefficients is plotted (Figure 6E(iii)). As we observed in Figure 6E(i), the correlation between the performance scores in both experimental sessions is strong. This finding is confirmed in Figure 6E(iii) where the correlation coefficient obtained from Figure 6E(i) resides in the outer edge of the histogram, suggesting that within-subject testing protocols are advantageous.

Magnitude of changes in sensory eye balance after short-term patching

In this section, we analyze data that represent the magnitude of change in eye dominance as a result of

short-term patching (i.e., patching effect). We follow the convention, where the differences between postpatching data and baseline data are used to quantify this effect.

Binocular rivalry

The patching effect is represented by the difference in ODI between baseline and postpatching measurements. The more positive the Δ ODI, the stronger the patching effect. A Pearson's correlation test revealed a nonsignificant correlation ($n = 15$, $r = 0.15$, $p = 0.597$) between the patching effects of the two repeated sessions.

The Bland–Altman plot in Figure 7A(ii) indicates that the limits of agreement are ±2.56 (z-scores). The measurement variability from the binocular rivalry task itself (grey shaded area in Figure 7A(ii)) is ±1.48 (z-scores). This corresponds with the median of the bootstrapped standard error for each testing block from both baseline and postpatching experiments. Unlike in the baseline measurements, the shaded area covers only one-half of the area within the limits of agreement. This finding suggests that one-half of the test–retest variability of the patching effect originates from the measurement error of the binocular rivalry task itself, rather than cognitive factors such as attention.

The weak correlation from Figure 7A(i) is confirmed in Figure 7A(iii), where the correlation coefficient obtained from Figure 7A(i) resides in the middle of the histogram, suggesting that it is not beneficial to use a within-subjects design.

Binocular combination at one contrast

The change in sensory eye dominance from patching is represented by the difference in perceived phase (degrees) between baseline and postpatching measurements. The more negative the difference in perceived phase, the stronger the patching effect).

A Pearson's correlation test found a significant correlation ($n = 15$, $r = 0.83$, $p < 0.001$) between the patching effects in both experimental sessions. The Bland–Altman plot in Figure 7B(ii) indicates that the limits of agreement are ±1.13 (z-scores). The expected measurement variability from the binocular combination task itself (grey shaded area in Figure 7B(ii)) is ±0.81 (z-scores).

The robust correlation from Figure 7B(iii) is corroborated in Figure 7B(iii), where the correlation coefficient obtained from Figure 7B(i) is located at the outer edge the histogram, suggesting that within-subjects designs are beneficial.

Binocular combination at many contrasts

The change in sensory eye dominance from short-term patching is represented by the difference in contrast ratio (dB) between baseline and postpatching measurements. The more positive the difference in contrast ratio (Δ dB), the stronger the patching effect. The correlation was not significant ($n = 34$, $r = 0.298$, $p = 0.073$; Figure 7C(i)), probably owing to some extreme points. However, these points are within three standard deviations and, therefore, were not categorized as outliers.

The Bland–Altman plot in Figure 7C(ii) indicates that the limits of agreement are ±2.32 (z-scores). The expected measurement variability from the binocular combination task itself (grey shaded area in Figure 7C(ii)) is ±0.46 (z-scores). Most of the area within the limits of agreement is not shaded in grey. That means most of the test–retest variability from the patching effect originates from factors other than the measurement variability associated with binocular combination task itself.

The insignificant correlation from Figure 7C(i) surprisingly resides at the outer edge of the histogram, suggesting that within-subjects designs are more sensitive than between-subject designs.

Parallel-oriented dichoptic masking

The change in sensory eye dominance from patching is represented by the difference in contrast ratio (dB) between baseline and postpatching measurements. The more negative the difference in the contrast threshold for the test grating (Δ dB), the stronger the patching effect. This applies to both parallel- and cross-oriented dichoptic masking. A Pearson's correlation test revealed a significant correlation ($n = 14$, $r = 0.57$, $p < 0.05$; Figure 7D(i)).

The Bland–Altman plot in Figure 7D(ii) indicates that the limits of agreement are ±1.82 (z-scores). The expected measurement variability from the task (grey area in Figure 7D(ii)) is ±0.56 (z-scores). Most of the area within the limits of agreement is not shaded in grey. This finding indicates that most of the test–retest variability of the patching effect originates from factors other than the task measurement error.

The strong correlation from Figure 7D(i) is confirmed in Figure 7D(iii), where the correlation coefficient resides at the outer edge of the histogram, suggesting that within-subject designs are superior to between-subject designs.

Cross-oriented dichoptic masking

A Pearson's correlation test indicated a significant correlation ($n = 14$, $r = 0.60$, $p < 0.05$; Figure 7E(i)). The Bland–Altman plot in Figure 7E(ii) indicates that the limits of agreement are ±1.75 (z-scores). The expected measurement variability from the task itself (grey shaded area in Figure 7E(ii)) is ±1.42 (z-scores). Most of the area within the limits of agreement is

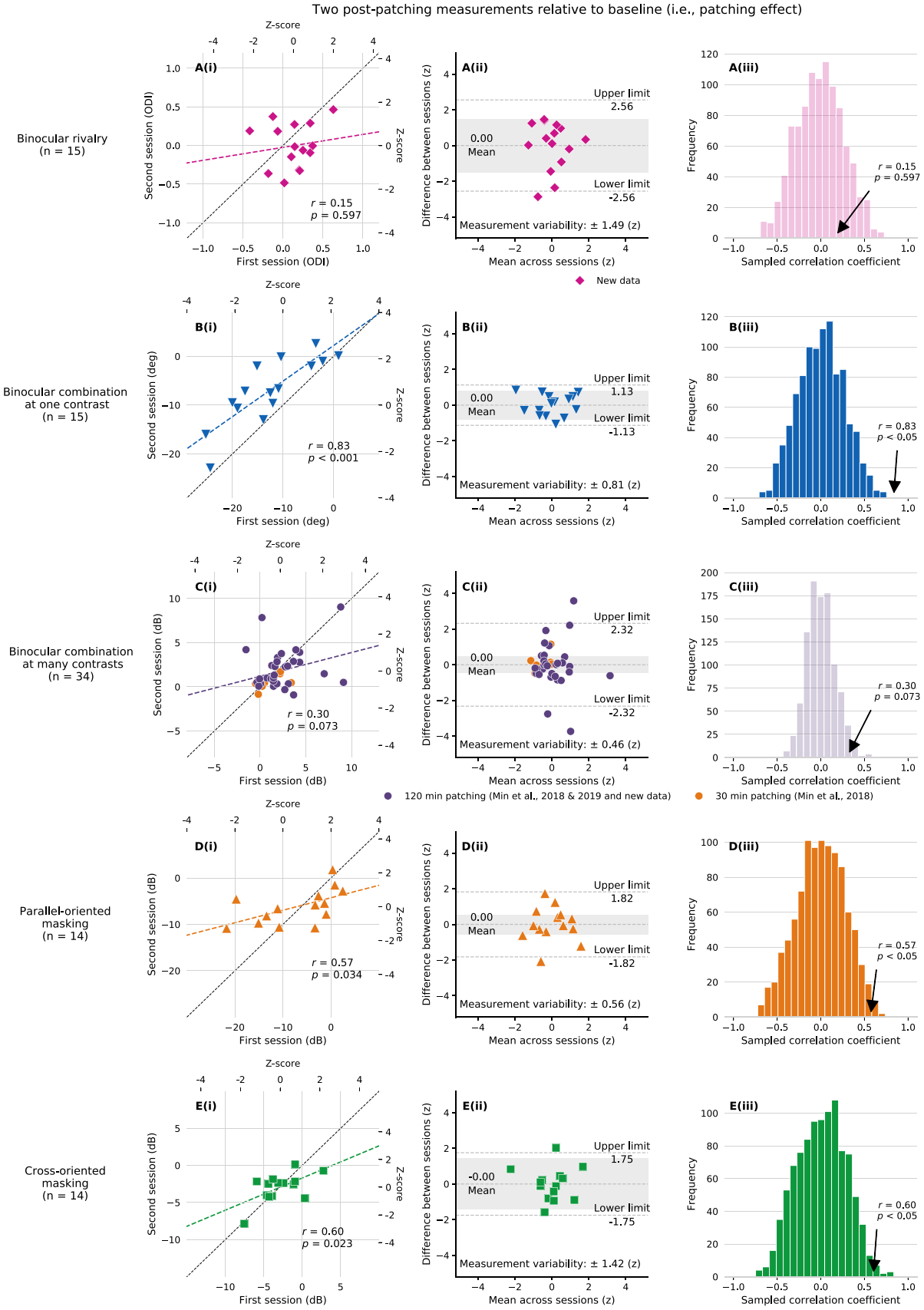

Figure 7. Repeatability of the patching effect as measured in the five psychophysical tasks. This figure is divided into five rows (task) and three columns (data analyses). (A) Binocular rivalry. Fifteen subjects were patched for 120 minutes. (B) Binocular phase combination at one contrast. Fifteen subjects were patched for 120 minutes. (C) Binocular phase combination at many contrasts. Seven subjects were patched for 30 minutes. Twenty-seven subjects were patched for 120 minutes. (D) Parallel-oriented dichoptic

→

←
masking. Fourteen subjects were patched for 150 minutes. (E) Cross-oriented dichoptic masking. 14 subjects were patched for 150 minutes. Column (i) Baseline reliability. Column (ii) Baseline measurement variability in a Bland–Altman plot. Column (iii) Baseline reliability illustrated in a histogram. The columns present data in the same manner as in Figure 6.

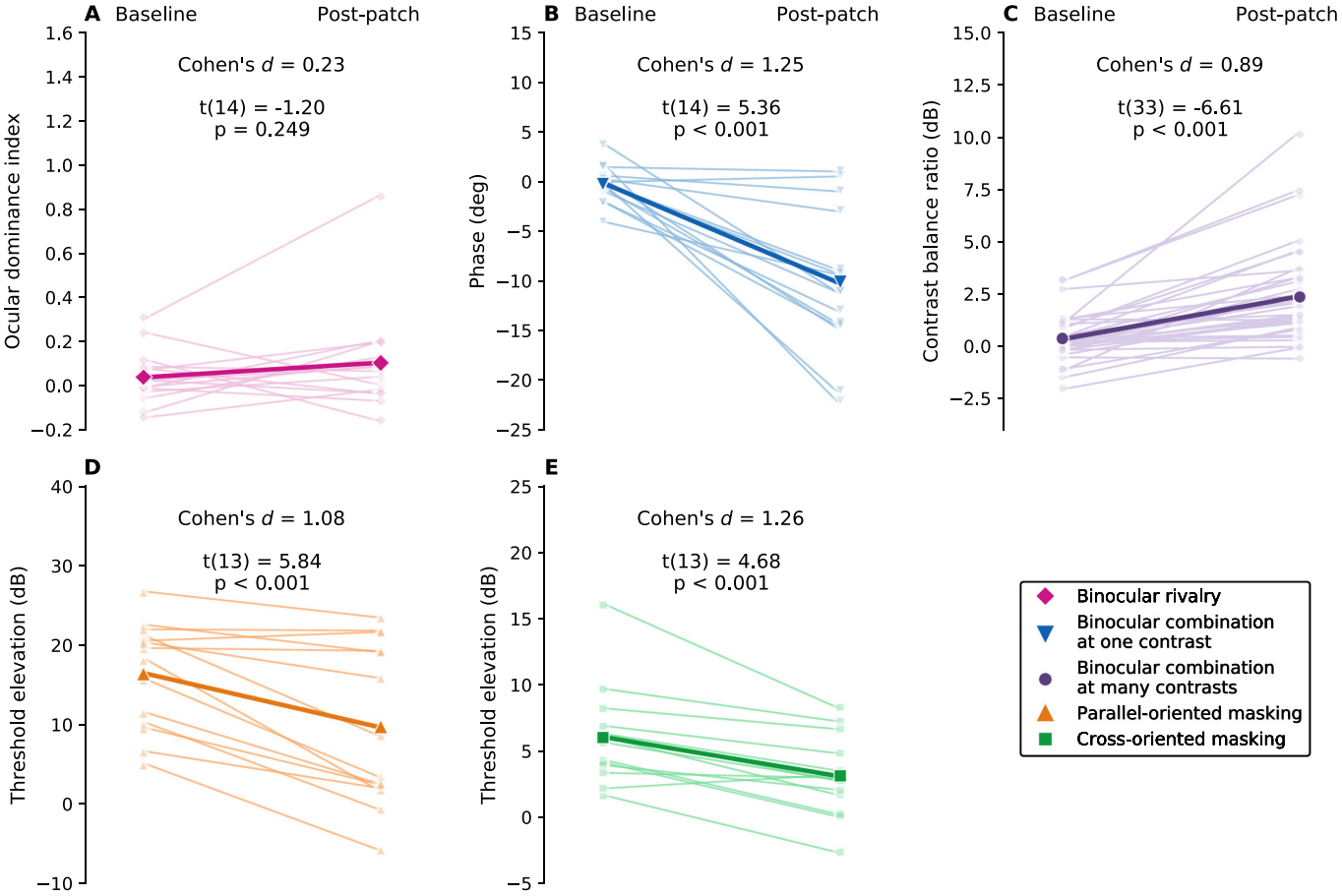

Figure 8. A slope chart that compares the baseline and postpatching measurements for each of the five psychophysical tasks. Solid plot and points represent the average of all subjects across two sessions. Transparent points and plots represent the averaged individual data across two sessions. Cohen's *d* (effect size) and results from a two-tailed sampled *t*-test are included for each panel. (A) Binocular rivalry. (B) Binocular combination at one contrast. (C) Binocular combination at many contrasts. (D) Parallel-oriented masking. (E) Cross-oriented masking.

shaded in grey. This finding suggests that most of the test–retest variability of the patching effect originates from the task measurement itself.

The robust correlation is confirmed in Figure 7E(iii), where the correlation coefficient resides at the outer edge of the histogram, indicating that there is an advantage of using a within-subject design for this task.

Detectability of the patching effect

Besides the test–retest reliability and the measurement variability, the detectability of the patching effect has to be robust for a task to be considered effective. To quantify the detectability of the patching effect, we computed the effect size (Cohen's *d*) of the changes in eye balance after short-term patching for all tasks and experimental sessions (Figure 8). The effect size does not depend on the sample size. Therefore, the differences in the sample sizes across tasks are irrelevant here. The greater the effect size, the larger the detectability of the patching effect in a given task. The effect size was used to perform a power analysis and determine the necessary sample size to detect the patching effect with a statistical significance (alpha = 0.05, power = 0.80) from a one-tailed paired *t*-test. A one-tailed paired *t*-test was selected for the power

analysis because short-term patching shifts the balance of the eyes in favor of the deprived eye only (i.e., one direction of expected change).

For binocular rivalry, the effect size in changes of ODIs (i.e., ODI) between postpatching and baseline results was 0.23 (Figure 8A). A power analysis revealed that 230 subjects would be necessary to detect the patching effect with a statistical significance. As for the phase combination task at one contrast, the effect size was 1.25 (Figure 8B) and the necessary sample size was found to be nine subjects. For the phase combination task at many contrasts, the effect size was 0.89 (Figure 8C) and the necessary sample size was 16 subjects. For the parallel-oriented masking task, the effect size was 1.08 (Figure 8D) and the necessary sample size was 12. As for the cross-oriented masking task, the effect size was 1.26 (Figure 8E) and the required sample size was 9. In summary, it seems that cross-oriented masking and binocular combination at one contrast are most sensitive in detecting the patching effect, closely followed by parallel-oriented masking and binocular combination at many contrasts. However, binocular rivalry seems to be least sensitive, requiring an extraordinarily large sample size for a reliable detection of the patching effect.

Moreover, we conducted a two-tailed paired *t*-test to compare the raw baseline and postpatching data for each task. The baseline and postpatching data were averaged across the two sessions for each subject. Whether the difference in the visual measure for each task was statistically significant after patching relative to baseline was assessed. In binocular rivalry, the change of eye dominance from patching was not significant, $t(14) = -1.20$, $p = 0.249$. However, for the other four tasks, the change from patching was highly significant (p's < 0.001). Along with the effect size, the t-test also demonstrates that binocular rivalry task is least effective in capturing the patching effect.

Summary of results

In this section, five properties of the five psychophysical tasks are ranked (see Figure 9). These properties are baseline reliability, patching effect reliability, baseline measurement variability, patching effect measurement variability, and detectability of the patching effect (as defined in the Introduction).

The correlations (i.e., measurement reliabilities) for baseline measurements and the magnitude of the patching effect are summarized in the form of the correlation coefficient from Pearson's correlation tests between the raw data from the first and second experimental sessions. The *p* values were not used as a summary index for measurement reliability because they heavily depend on the sample size. The larger the sample size, the smaller the *p*-value. This point is exemplified by the large sample size for the binocular phase combination task at many contrasts and its low p-value from correlation in Figure 6.

The baseline and the patching effect measurement variabilities (z-scores) correspond with the width of the shaded gray regions in the Bland–Altman plots. They are the measurement error from the psychophysical task itself rather than extraneous errors such as day-to-day variability and attention levels. What do the measurement variability indicate? The measurement variability of the patching effect includes variability from the baseline measurement and the changes in the strength of the patching effect across days. However, the measurement variability of the baseline data includes variance associated with the task performance.

As for the detectability of the patching effect, the effect size (i.e., Cohen's *d*) in changes of eye balance between baseline and postpatching was computed. The greater the effect size, the higher the detectability of the patching effect, and the lower the required sample size to detect the patching effect with a statistical significance (as shown by the power analysis above).

To rank the psychophysical tasks from best to worst, we normalized the statistical values (correlation coefficient, measurement variability in z-scores, and Cohen's *d* for the effect size). If the normalized value is 1, it indicates that it is the best of all tasks; if the normalized value is 0, it indicates that it is the worst of all tasks. As for measurement variability (i.e., correlation coefficient) and measurement detectability of the patching effect (i.e., Cohen's *d*), the highest values among all tasks were converted to 1 and the lowest to 0 using this formula:

$$normalized\ score = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

where x_i indicates a raw statistical score that is to be normalized (e.g., binocular rivalry: Cohen's $d = 0.30$), x_{min} the minimum value, and x_{max} the maximum value. Therefore, this equation normalizes the largest value to 1 and the smallest to 0. However, in the case of measurement variability (z-score), a smaller value is superior. Therefore, we used this equation below to normalize the smallest value to 1 and the largest to 0.

$$normalized\ score = 1 - \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Discussion

Are different psychophysical tasks associated with distinct neural sites and mechanisms?

Studies using binocular rivalry and binocular combination at one contrast have revealed different

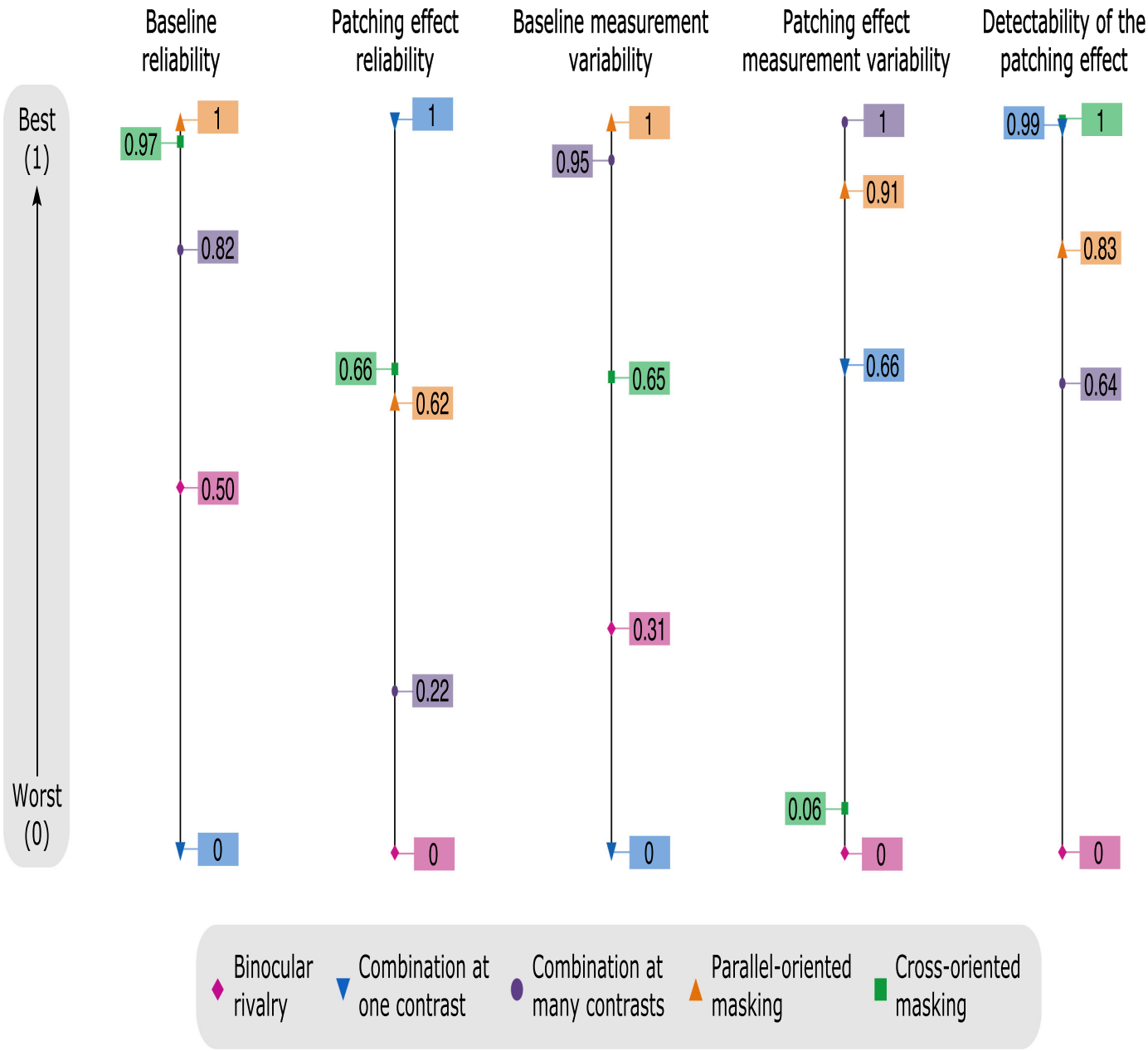

Figure 9. Summary of results. Baseline reliability refers to the correlation coefficient (column (i) in Figure 6). Patching effect reliability indicates the correlation coefficient from correlation analysis of the difference between postpatching and prepatching baseline data (column (ii) in Figure 7). Baseline measurement variability refers to the median of the standard error for each testing block from baseline measurement; this is represented by the gray areas in column (ii) of Figure 6. Patching effect measurement variability denotes the median of the standard error for each testing block from both postpatching and prepatching baseline data; this is represented by the gray areas in column (ii) of Figure 7. Detectability of the patching effect is the effect size (Cohen's *d*) between baseline and postpatching measurement (see Figure 8). Each value was normalized in a scale where 1 represents best and 0 worst of all tasks.

magnitudes of the patching effect (Bai, Dong et al., 2017; Lunghi, Burr et al., 2013; Zhou, Reynaud et al., 2017). This finding supports that the patching effect takes place at multiple neural sites. For example, Baldwin and Hess (2018) interleaved two dichoptic masking tasks with different orientations of the mask grating within one experiment and repeated the experiment twice (Baldwin & Hess, 2018). Their experimental design minimized measurement variability between the two tasks. They reported that the orientation of the mask determines the magnitude of the patching effect. This finding reinforces the notion

that the patching effect is multifaceted and that one psychophysical task might capture only one aspect of the change in neural plasticity. If this interpretation is true, different psychophysical tasks can be associated with different aspects or sites.

However, we show that this difference in results can also be attributed to a wide measurement variability of the patching effect owing to the method itself, such as binocular rivalry (see Figure 9). Moreover, the measurement variability of the patching effect between the parallel- and cross-oriented masking tasks is different, although the patching effect reliabilities are both robust. The measurement variability is much wider in the cross-oriented masking task (Figure 9). This outcome suggests that, when gratings of orthogonal orientations are presented dichoptically, the measurement variability of the patching effect can increase. This reasoning also applies to the obvious difference in the measurement variability of the patching effect between binocular rivalry and binocular combination tasks. Since the variability directly confounds the outcome of interest (i.e., magnitude of the patching effect), we cannot yet conclude that the results from different psychophysical tasks reflect separate neural sites. If our logic is correct, it will be more beneficial to use a task, such as binocular combination and parallel-oriented masking, that presents gratings at a parallel orientation to both eyes to measure the patching effect.

How reliable is the baseline measurement for each task?

As Figure 9 shows, binocular rivalry and binocular combination at one contrast have poor reliability and measurement variability in baseline measurements. Conversely, a binocular combination at many contrasts and parallel-oriented dichoptic masking seem to measure baseline in a consistent fashion. What can be the contributing factors for the poor reliability of binocular rivalry and binocular phase combination at one contrast?

To begin with, binocular rivalry measures competition, rather than the combination, between the eyes by presenting two rivalrous images separately to both eyes. The interocular competition during rivalry causes a rapid and irregular fluctuation of sensory eye dominance over visual space and time (Blake, Fox et al., 1971; Blake & Logothetis, 2002). The random nature of binocular rivalry might widen the measurement variability. Moreover, attention can affect the temporal dynamics of the rivalry (Paffen & Alais, 2011), suggesting that this task is influenced significantly by cognitive factors (Bai, Dong et al., 2017; Ramamurthy & Blaser, 2018). The poor reliability of the baseline measurement between the two separate days of testing might indicate that the level of attention throughout the task between the sessions differed. More important, binocular rivalry task is the only method that captures continuous time-series data of the subject, thereby adding one more dimension to the data (i.e., time). All other four tasks yielded discrete, rather than continuous, data. The discrete structure of the data might decrease the source of measurement error. Therefore, the random dynamic nature of binocular rivalry and the influence of top-down attentional factors might have increased the measurement variability of baseline measurement.

Our results are surprising given the fact that binocular rivalry has been used to study a wide range of visual phenomena (Blake & Logothetis, 2002), such as sensory eye dominance at a population level (Dieter, Sy et al., 2017) and within the visual field (Dieter, Sy et al., 2017) and its changes after short-term patching (Ooi & He, 2020). It has also been used as a gold standard when a novel test for measuring sensory eye dominance is developed (Rice, Leske et al., 2008). A study has investigated the test–retest reliability of binocular rivalry measurement (Dieter, Sy et al., 2017), reporting a robust correlation between the two experimental sessions. The authors highlight the correlation as evidence to claim that binocular rivalry is a reliable test. However, the correlation coefficient is not indicative of test–retest replicability, because two samples with significantly different means can still have a strong linear relationship. In our study, we also found a good correlation for binocular rivalry for the baseline measurement of our new data ($n = 15$, $r = 0.52$, $p = 0.043$; pink points in Figure 6A(i)). However, a large measurement variability was observed (±1.72 z-scores) in the dataset.

In the case of binocular combination task at one contrast, as its name implies, only one contrast ratio between the eyes is used. We believe that using only one contrast ratio of the stimuli might have widened the measurement variability at baseline. Conversely, in binocular combination at many contrasts, the various contrast ratios were used to display the stimulus. Then the contrast ratio, where the perceived phase is, 0 was estimated by fitting a contrast gain model (Ding & Sperling, 2006) to the data across all contrast levels. Therefore, the version of the task in which data was collected across multiple contrast values, not surprisingly, had a much smaller measurement variability than the binocular combination task at one contrast.

Interestingly, we found a stark difference in the measurement variability (gray areas in the Bland–Altman plots from Figure 6) between parallel- and cross-oriented dichoptic masking tasks. The 14 subjects in this experiment were identical as the two tasks were interleaved alternately (Figure 4). The only difference in

these methods was the orientation of the mask grating. Cross-oriented masking induces binocular competition between the eyes because the orientations of the mask and detection gratings are orthogonal. On the other hand, parallel-oriented masking does not induce any competition since both the mask and detection gratings are identically oriented. The orthogonal (i.e., non-fusible) orientations of the gratings might account for the large difference in the measurement variability between the two masking tasks. This explanation might also help to explain the large measurement variability in the binocular rivalry task.

Is the patching effect stable across days?

Studies using various tests have demonstrated the replicability of the patching effect. However, the magnitude of the plasticity change has been reported to be not uniform across tasks. One can observe the variability from the patching effect by comparing the measurement variabilities (z-scores) between the patching effect and baseline for each task. If the range of ± z-scores from the measurement variability of the patching effect is larger than that of the baseline, then the data of the patching effect are more variable than the baseline data. In other words, in this scenario, baseline measurement data have one less source of variability, which is patching. On the other hand, if the measurement variability is similar between baseline and the patching effect measurement, then one can infer that patching itself does not introduce additional variance to the data. According to our results, all tasks except for cross-oriented masking have a similar or narrower measurement variability (Figure 7(ii)) compared with that from baseline. These tasks suggest that the patching effect is stable across days within the same subject. Also, our results indicate that patching does not necessarily increase the variance of the data.

Do all five psychophysical tasks reliably detect changes in sensory eye balance?

For a given task, the detectability (i.e., sensitivity) of changes in eye balance must be reliably consistent to be useful. If the detectability is poor, it can lead to conflicting results from different laboratory groups even if the methodology is identical. To quantify the detectability, we computed the effect size (Cohen's d) between data from baseline and those from postpatching measurements (see Figure 8). Our analysis indicates that binocular combination at one and many contrasts, parallel- and cross-oriented masking tasks show a reliable detectability to the patching effect (Cohen's $d > 0.8$), whereas binocular rivalry does not (Cohen's $d = 0.23$). In fact, for this reason, it might be that different results have been reported regarding the role of physical exercise in potentiating the patching effect (Finn, Baldwin et al., 2019; Lunghi & Sale, 2015). Although Lunghi et al. found that exercise magnifies the changes in eye balance after short-term patching, Finn et al. did not find such an effect (Finn, Baldwin et al., 2019; Lunghi & Sale, 2015). Nevertheless, it might come as a surprise that binocular rivalry has such a poor detectability when the first seminal study that reported the phenomenon of short-term monocular deprivation used binocular rivalry to measure eye dominance (Lunghi, Burr et al., 2011).

There are several possible explanations as to why our data do not show a large effect size, whereas some previous studies have shown otherwise. First, we used a modified version of ODI to compute the effect of short-term monocular deprivation as measured in binocular rivalry after previous studies (Dieter, Sy et al., 2017; Finn, Baldwin et al., 2019). However, the indices for eye dominance as measured in binocular rivalry do not seem to be identical across studies; these include, but are not limited to, normalized dominance duration (Kim, Kim et al., 2017), phase duration ratios (Lunghi, Emir et al., 2015), and deprivation index (Binda & Lunghi, 2017). In addition, many studies normalize the change in eye balance after short-term patching relative to baseline (i.e., 0), and then conduct data analyses by solely using the difference in data between postpatching and baseline; hence, the relative difference to baseline might seem to be large. As shown in Figure 8, however, we used our raw data from baseline and postpatching to directly compute the effect size to be more precise. Also, many studies have a smaller sample size than ours ($n = 15$), so the reported effect size might have been more prone to a few outliers who displayed an exceptionally large change in sensory eye balance (Bai, Dong et al., 2017; Binda, Kurzawski et al., 2018; Binda & Lunghi, 2017; Kim, Kim et al., 2017; Lo Verde, Morrone et al., 2017; Lunghi, Burr et al., 2011; Lunghi, Burr et al., 2013; Lunghi, Galli-Resta et al., 2019; Ramamurthy & Blaser, 2018; Wang, McGraw et al., 2020). In addition, although applying a diffuser to one eye is a common method to monocularly deprive its visual input, the methods of monocular deprivation can differ across studies. For example, Bai et al. (Bai, Dong, He, & Bao, 2017) used pink noise and mean color to deprive the visual input of one eye, whereas Ramamurthy et al. (Ramamurthy & Blaser, 2018) used kaleidoscopic monocular deprivation.

Last, the duration of the test for the binocular rivalry task might differ between our protocol and those from previous studies. Most other studies show data that were collected during longer testing periods (e.g., eight blocks of 180 sec) of a binocular rivalry task (Lunghi,

Burr et al., 2011; Pettigrew & Miller, 1998), thereby decreasing the margin of error and possibly increasing the effect size. However, because we wanted to test whether the five methods would be appropriate in a clinical setting, where time is often constrained, we asked our subjects to complete only four blocks of 180 seconds each to measure the baseline data and two for postpatching data.

Which psychophysical tasks should be used in the clinical setting to measure sensory eye dominance and the patching effect?

Recent clinical studies on amblyopes have incorporated training protocols that involve patching the dysfunctional eye (Chen, He et al., 2020; Lunghi, Sframeli et al., 2019; Zhou, He et al., 2019), a design that is identical to the one used in short-term patching studies in normal observers. To ensure that the findings from preliminary studies are replicable in a wider population, the choice of test in clinical studies is important.

To begin, our findings show that binocular rivalry and binocular combination at only one contrast have poor test–retest replicability in baseline measurement. In addition, binocular rivalry exhibits a large test–retest variability and low detectability of the patching effect. This finding may limit its usefulness for clinical studies. Instead, psychophysical tasks that capture stable baseline performance and a repeatable patching effect and detect the patching effect easily will be most useful. According to our results, these tasks are parallel-oriented dichoptic masking and binocular phase combination at many contrasts.

Limitations of the study

In hindsight, a study design that compares two variations of a binocular rivalry task in an interleaved fashion as we did so with binocular phase combination tasks (one contrast vs. many contrasts) would have been preferable. As we mentioned elsewhere in this article, we first attempted to compare binocular rivalry and combination at many contrasts by interleaving them in a single design (Figure 1). However, owing to the failure of the stimuli display of the phase combination task, we had to discard the data of binocular phase combination but keep those of binocular rivalry. To maintain a comparable design, we decided to recollect data of binocular phase combination at many contrasts by interleaving with another variation of binocular phase combination (at one contrast).

Moreover, the number of testing blocks and the duration of time for each block were not identical across the five psychophysical tasks. Furthermore, the subjects were not always paired across the five tasks. This discrepancy might have affected the difference in the effect size of the patching effect. Performing such a controlled comparison would require a large study designed from the outset for that purpose. In our case, the study we present is a meta-analysis across several published studies. We therefore do expect extraneous differences between those studies to account for a part of the differences we see between tasks.

Conclusion

There have been conflicting reports on the patching effect from short-term deprivation in adults and children. The magnitude of the patching effect has been found to be variable across different tests (binocular rivalry and combination) and within the identical test (binocular rivalry) across conditions. In the Introduction, three explanations for these discrepancies are introduced. First, the mechanism of the patching effect might be multifaceted and different tasks might reflect different processing sites. If this notion holds true, each psychophysical task might capture only one aspect of the entire plasticity change. Previous psychophysical studies have advocated this reasoning (Bai et al., 2017; Baldwin & Hess, 2018). Second, the measurement error associated with the tasks might be poorer with certain tasks. In light of our findings, this claim is reasonable for some tasks. For instance, the presentation of orthogonal gratings (e.g., binocular rivalry and cross-oriented dichoptic masking tasks) seems to increase the measurement variability of the task directly, thereby making the baseline or the patching effect more variable. Third, the patching effect might be itself an unstable phenomenon. Our findings show that this is not the case, because we do not find evidence for any additional source of variability for the patching effect. Finally, our results indicate that binocular phase combination at many contrasts and parallel-oriented dichoptic masking are most reliable for measuring the patching effect.

Keywords: ocular dominance plasticity, monocular deprivation, methodology, measurement error

Acknowledgments

Supported by the National Natural Science Foundation of China (31970975), the Qianjiang Talent Project (QJD1702021), the Wenzhou Medical University grant QTJ16005 and the Project of State Key Laboratory of Ophthalmology, Optometry and Visual Science, Wenzhou Medical University (K171206) to JZ, the Zhejiang Basic Public Welfare Research Project

(LGJ20H120001) to ZH, the Canadian Institutes of Health Research Grants CCI-125686, NSERC grant 228103, and an ERA-NET Neuron grant (JTC2015) to RH, and Canadian Institutes of Health Research graduate award to SM. The sponsor or funding organization had no role in the design or conduct of this research.

Commercial relationships: none.
Corresponding authors: Zhifen He, Robert F. Hess.
Emails: zhifen0821@163.com, robert.hess@mcgill.ca.
Address: McGill University, 1650 Cedar Ave, Montreal General Hospital, Montreal, Quebec H3G1A4, Canada.

References

Bai, J., Dong, X., He, S., & Bao, M. (2017). Monocular deprivation of Fourier phase information boosts the deprived eye's dominance during interocular competition but not interocular phase combination. *Neuroscience, 352*, 122–130.

Baldwin, A. S., & Hess, R. F. (2018). The mechanism of short-term monocular deprivation is not simple: Separate effects on parallel and cross-oriented dichoptic masking. *Scientific Reports, 8*(1), 6191.

Begum, M., & Tso, D. (2016). Shifts in interocular balance resulting from short-term monocular deprivation in adult macaque visual cortex are not magno-dominated. *Journal of Vision, 16*(12), 1328–1328.

Binda, P., Kurzawski, J. W., Lunghi, C., Biagi, L., Tosetti, M., & Morrone, M. C. (2018). Response to short-term deprivation of the human adult visual cortex measured with 7T BOLD. *Elife, 7*, e40014.

Binda, P., & Lunghi, C. (2017). Short-term monocular deprivation enhances physiological pupillary oscillations. *Neural Plasticity, 2017*, 6724631.

Blake, R., & Logothetis, N. K. (2002). Visual competition. *Nature Reviews Neuroscience, 3*(1), 13–21.

Blake, R. R., Fox, R., & McIntyre, C. (1971). Stochastic properties of stabilized-image binocular rivalry alternations. *Journal of Experimental Psychology, 88*(3), 327–332.

Chadnova, E., Reynaud, A., Clavagnier, S., & Hess, R. F. (2017). Short-term monocular occlusion produces changes in ocular dominance by a reciprocal modulation of interocular inhibition. *Scientific Reports, 7*, 41747.

Chen, Y., He, Z., Mao, Y., Chen, H., Zhou, J., & Hess, R. F. (2020). Patching and suppression in amblyopia: One mechanism or two? *Frontiers in Neuroscience, 13*, 1364.

Dieter, K. C., Sy, J. L., & Blake, R. (2017). Individual differences in sensory eye dominance reflected in the dynamics of binocular rivalry. *Vision Research, 141*, 40–50.

Dieter, K. C., Sy, J. L., & Blake, R. (2017). Persistent biases in binocular rivalry dynamics within the visual field. *Vision (Basel), 1*(3), 10.

Ding, J., & Sperling, G. (2006). A gain-control theory of binocular combination. *Proceedings of the National Academy of Sciences of the United States of American, 103*(4), 1141–1146.

Finn, A. E., Baldwin, A. S., Reynaud, A., & Hess, R. F. (2019). Visual plasticity and exercise revisited: No evidence for a "cycling lane" Finn, Baldwin, Reynaud, & Hess. *Journal of Vision, 19*(6), 21–21.

Kim, H. W., Kim, C. Y., & Blake, R. (2017). Monocular perceptual deprivation from interocular suppression temporarily imbalances ocular dominance. *Current Biology, 27*(6), 884–889.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in psychtoolbox-3. *Perception, 36*(14), 1–16.

Lo Verde, L., Morrone, M. C., & Lunghi, C. (2017). Early cross-modal plasticity in adults. *Journal of Cognitive Neuroscience, 29*(3), 520–529.

Lunghi, C., Berchicci, M., Morrone, M. C., & Di Russo, F. (2015). Short-term monocular deprivation alters early components of visual evoked potentials. *Journal of Physiology, 593*(19), 4361–4372.

Lunghi, C., Burr, D. C., & Morrone, C. (2011). Brief periods of monocular deprivation disrupt ocular balance in human adult visual cortex. *Current Biology, 21*(14), R538–539.

Lunghi, C., Burr, D. C., & Morrone, M. C. (2013). Long-term effects of monocular deprivation revealed with binocular rivalry gratings modulated in luminance and in color. *Journal of Vision, 13*(6), 1–1.

Lunghi, C., Emir, U. E., Morrone, M. C., & Bridge, H. (2015). Short-term monocular deprivation alters GABA in the adult human visual cortex. *Current Biology, 25*(11), 1496–1501.

Lunghi, C., Galli-Resta, L., Binda, P., Cicchini, G. M., Placidi, G., Falsini, B., . . . Morrone, M. C. (2019). Visual cortical plasticity in retinitis pigmentosa. *Investigative Ophthalmology & Visual Science, 60*(7), 2753–2763.

Lunghi, C., & Sale, A. (2015). A cycling lane for brain rewiring. *Current Biology, 25*(23), R1122–1123.

Lunghi, C., Sframeli, A. T., Lepri, A., Lepri, M., Lisi, D., Sale, A., . . . Morrone, M. C. (2019). A new counterintuitive training for adult amblyopia. *Annals of Clinical and Translational Neurology, 6*(2), 274–284.

Miles, W. R. (1930). Ocular dominance in human adults. *Journal of General Psychology, 3*, 412–430.

Min, S. H., Baldwin, A. S., & Hess, R. F. (2019). Ocular dominance plasticity: A binocular combination task finds no cumulative effect with repeated patching. *Vision Research, 161*, 36–42.

Min, S. H., Baldwin, A. S., Reynaud, A., & Hess, R. F. (2018). The shift in ocular dominance from short-term monocular deprivation exhibits no dependence on duration of deprivation. *Scientific Reports, 8*(1), 17083.

Ooi, T. L., & He, Z. J. (2020). Sensory eye dominance: Relationship between eye and brain. *Eye Brain, 12*, 25–31.

Paffen, C. L. E., & Alais, D. (2011). Attentional modulation of binocular rivalry. *Frontiers in Human Neuroscience, 5*, 105.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision, 10*(4), 437–442.

Pettigrew, J. D., & Miller, S. M. (1998). A "sticky" interhemispheric switch in bipolar disorder? *Proceedings of the Royal Society of London. Series B: Biological Sciences, 265*(1411), 2141–2148.

Ramamurthy, M., & Blaser, E. (2018). Assessing the kaleidoscope of monocular deprivation effects. *Journal of Vision, 18*(13), 14.

Reynaud, A., Blaize, K., Chavane, F., & Hess, R. F. Monocular vision is intrinsically unstable: a side-effect of binocular homeostasis. *BioRXiv*, https://doi.org/10.1101/2020.03.17.987362.

Reynaud, A., Roux, S., Chemla, S., Chavane, F., & Hess, R. (2018). Interocular normalization in monkey primary visual cortex. *Journal of Vision, 18*(10), 534–534.

Rice, M. L., Leske, D. A., Smestad, C. E., & Holmes, J. M. (2008). Results of ocular dominance testing depend on assessment method. *Journal of American Association for Pediatric Ophthalmology and Strabismus, 12*(4), 365–369.

Tso, D., Miller, R., & Begum, M. (2017). Neuronal responses underlying shifts in interocular balance induced by short-term deprivation in adult macaque visual cortex. *Journal of Vision, 17*(10), 576–576.

Wang, M., McGraw, P., & Ledgeway, T. (2020). Short-term monocular deprivation reduces inter-ocular suppression of the deprived eye. *Vision Research, 173*, 29–40.

Zhou, J., Baker, D. H., Simard, M., Saint-Amour, D., & Hess, R. F. (2015). Short-term monocular patching boosts the patched eye's response in visual cortex. *Restorative Neurology and Neuroscience, 33*(3), 381–387.

Zhou, J., Clavagnier, S., & Hess, R. F. (2013). Short-term monocular deprivation strengthens the patched eye's contribution to binocular combination. *Journal of Vision, 13*(5), 12–12.

Zhou, J., He, Z., Wu, Y., Chen, Y., Chen, X., Liang, Y., & Hess, R. F. (2019). Inverse occlusion: A binocularly motivated treatment for amblyopia. *Neural Plasticity, 2019*, 12.

Zhou, J., Reynaud, A., & Hess, R. F. (2014). Real-time modulation of perceptual eye dominance in humans. *Proceedings. Biological Sciences, 281*(1795), 20141717.

Zhou, J., Reynaud, A., Kim, Y. J., Mullen, K. T., & Hess, R. F. (2017). Chromatic and achromatic monocular deprivation produce separable changes of eye dominance in adults. *Proceedings. Biological Sciences, 284*(1867), 20171669.