Perceptual integration without conscious access

Johannes J. Fahrenfort[a,1], Jonathan van Leeuwen[a], Christian N. L. Olivers[a], and Hinze Hogendoorn[b]

[a]Department of Experimental and Applied Psychology, Vrije Universiteit Amsterdam, 1081BT Amsterdam, The Netherlands; and [b]Helmholtz Institute, Experimental Psychology, Utrecht University, 3584CS Utrecht, The Netherlands

Edited by Randolph Blake, Vanderbilt University, Nashville, TN, and approved February 22, 2017 (received for review October 18, 2016)

The visual system has the remarkable ability to integrate fragmentary visual input into a perceptually organized collection of surfaces and objects, a process we refer to as perceptual integration. Despite a long tradition of perception research, it is not known whether access to consciousness is required to complete perceptual integration. To investigate this question, we manipulated access to consciousness using the attentional blink. We show that, behaviorally, the attentional blink impairs conscious decisions about the presence of integrated surface structure from fragmented input. However, despite conscious access being impaired, the ability to decode the presence of integrated percepts remains intact, as shown through multivariate classification analyses of electroencephalogram (EEG) data. In contrast, when disrupting perception through masking, decisions about integrated percepts and decoding of integrated percepts are impaired in tandem, while leaving feedforward representations intact. Together, these data show that access consciousness and perceptual integration can be dissociated.

phenomenal consciousness | access consciousness | perceptual integration | masking | attentional blink

There has been a long-standing debate about the relationship between consciousness and the integration of information, dating back at least to Helmholtz who proposed that conscious perception is the result of unconscious integration of spatially scattered features, allowing the brain to make perceptual inferences about visual input (1). However, information integration can occur locally, within sensory modules, or globally, when information is communicated to widespread areas across the brain, including response modules. Whereas the first type—to which we refer here as perceptual integration—has been related to phenomenal consciousness (subjective experience; refs. 2, 3), the latter type of integration has been linked to access (or in some views "true") consciousness (availability for report; refs. 4, 5). In the current study, we investigate whether perceptual integration is ontologically independent from conscious access, by determining to what extent it maintains its neural signature when access to consciousness is disturbed.

We use the Kanizsa illusion (Fig. 1*A*), together with two well-known manipulations of consciousness, to assess whether neural representations can reach a state of integration in which features are combined to form perceptual entities, despite not being consciously reported. Kanizsa figures are similar to control figures in terms of physical input, but they have very different perceptual outcomes, notably an illusory surface region with accompanying contours (6) and increased brightness (7). These emergent properties are a primary demonstration of perceptual integration, as the constituent parts in isolation (the inducers) do not carry any of the effects that are brought about by their configuration.

Earlier work has shown that Kanizsa configurations can facilitate detection of target stimuli, with and without competing objects (8–10). However, in these studies, conscious access has been implemented in various ways, whereas the dependent measure was always a behavioral response. The only study that has measured the neural substrate of perceptual integration in the absence of conscious report, postponed the behavioral response until after data collection (11), leaving open the possibility that subjects were consciously accessing the stimulus during scanning but had forgotten it at test time. The level at which conscious access and perceptual integration interact thus remains unclear. The current study used several electroencephalographic (EEG) measures to investigate the neural substrate of perceptual integration under two different types of manipulations known to affect consciousness: masking and the attentional blink (AB) (see Fig. 1*B* for the factorial design).

Masking is known to leave feedforward processing largely intact while selectively interfering with local processes of perceptual integration, as well as behavioral detection (12–14). See refs. 15 and 16 for more in-depth reviews about the distinction between feedforward integration and local recurrence-based perceptual integration. We therefore expected masking to interfere with both behavior and perceptual integration. The AB, on the other hand, is thought to impair behavioral detection and access to consciousness (17) by disrupting long-range integration (18), but to what degree perceptual integration occurs without conscious access has not been established. Therefore, the crucial question in the current study was whether neural markers of perceptual integration would be impaired when access was disturbed by the AB.

NEUROSCIENCE

PSYCHOLOGICAL AND COGNITIVE SCIENCES

Results

The main experiment consisted of two phases. In the first phase, 16 subjects performed a behavioral training session to become familiar with the task and the stimulus set. Subjects who performed adequately were enrolled in the EEG phase (11 out of 16; see *SI Methods* for details). In the second phase, we recorded 64-channel EEG data in two EEG sessions. Two black target figures (T1 and T2) were shown in a rapid serial visual presentation (RSVP) containing red distractors. Each target could either be a Kanizsa or a control figure (Fig. 1*A* and Fig. S1).

Significance

Our brain constantly selects salient and/or goal-relevant objects from the visual environment, so that it can operate on neural representations of these objects, but what is the fate of objects that are not selected? Are these discarded so that the brain only has an impoverished nonperceptual representation of them, or does the brain construct perceptually rich representations, even when objects are not consciously accessed by our cognitive system? Here, we answer that question by manipulating the information that enters into awareness, while simultaneously measuring cortical activity using EEG. We show that objects that do not enter consciousness can nevertheless have a neural signature that is indistinguishable from perceptually rich representations that occur for objects that do enter into conscious awareness.

Author contributions: J.J.F. and J.v.L. designed research; J.v.L. performed research; J.J.F., J.v.L., and H.H. analyzed data; and J.J.F., J.v.L., C.N.L.O., and H.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: All data and analysis scripts are stored in a repository at the Department of Experimental and Applied Psychology, Vrije Universiteit Amsterdam, and can be obtained by contacting the corresponding author.

[1]To whom correspondence should be addressed. Email: fahrenfort.work@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1617268114/-/DCSupplemental.

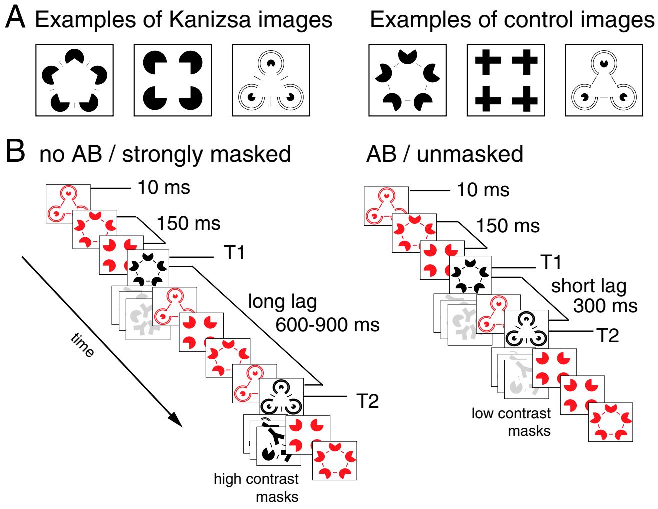


Fig. 1. Experimental design. (*A*) Examples of different Kanizsa images and their controls as used in the experiment (see Fig. S1 for the complete stimulus set). (*B*) Examples of two of the four trial types in the factorial design: without an AB (long lag) and strong masking (*Left*) and with an AB (short lag) and no masking (*Right*).

T1 and T2 lag was varied, inducing an AB at short lags (300 ms) with recovery at long lags (≥600 ms). In one-half of the trials, T2 was strongly masked using high-contrast masks. In the other half, low-contrast masks were used, so that there was no effect of masking (see examples of masks in Fig. S2). Examples of two of the four trial types are shown in Fig. 1*B*. At the end of each trial, subjects indicated whether T1 and/or T2 contained a surface region (see *SI Methods* for details). The ability to distinguish surface from control figures was computed as the hit rate (HR) minus the false-alarm rate (FAR), serving as a behavioral index of perceptual integration. T1 accuracy was high, at 0.90 (SEM, 0.02).

To establish a neural index for perceptual integration, we trained a linear discriminant classifier to categorize trials as either Kanizsa or control, using the amplitude of the EEG signal across electrodes as features for classification (see *SI Methods* for details). To prevent task and response-related processes from contaminating this neural index of perceptual integration, the training set was obtained from an independent RSVP task containing Kanizsas and controls. In this task, subjects pressed a button on black figure repeats (1-back task on black targets while ignoring red distractors; Fig. S3). This prevented task, response, or decision mechanisms from confounding classification performance, as target identity (repeat or not) was independent from stimulus class (Kanizsa or control), and all trials on which a response was given were excluded.

Next, we used this classifier on the experimental runs, computing classification accuracy (HR − FAR, just as in the behavioral measure) for every time sample, yielding classification accuracy over time. As in behavior, Kanizsa vs. control classification accuracy for T1 was well above chance, peaking at ∼264 ms (Fig. 2*A*), and was strongly occipital in nature [see correlation/class-separability map (19) in Fig. 2*B*]. For this reason, classifications in this analysis were restricted to occipital electrodes (see *SI Methods* and Fig. S4 for details). The fact that the classifier was able to discriminate Kanizsas from controls was reassuring, but we also wanted to establish a direct link between peak classification accuracy and perceptual integration.

To achieve this, we first computed behavioral accuracy separately for the 12 Kanizsa–control pairs that were used in the experiment. Different Kanizsa–control pairs yielded different behavioral accuracies due to inherent differences regarding the ease with which Kanizsa figures are perceptually integrated to result in surface perception (see Fig. S1 for the full Kanizsa–control stimulus set). Next, we applied a robust linear regression analysis (20) to determine whether T1 peak classification performance for these pairs (the neural index for perceptual integration) would be able to predict behavioral accuracy at T1. Peak classification accuracy was able to predict behavioral performance with remarkably high accuracy ($R^2 = 0.61$, $P < 0.005$; see Fig. 2*C*, using colored dots to refer to the specific Kanizsa–control pairs in Fig. S1), providing independent evidence that peak classification accuracy captures the signal underlying perceptual integration. Similar results were obtained for the individual experimental conditions (*SI Results*, *Prediction of Behavioral Accuracy Based on Neural Classification Accuracy*, and Fig. S5). Finally, we also checked whether there was a contribution of frontal electrodes to perceptual integration. Although we observed a small but significant effect of classification accuracy at T1 and at long lags, this signal was not predictive of perceptual integration across the 12 Kanizsa–control pairs, suggesting that it reflects a generic presence–absence signal (*SI Results*, *The Contribution of Frontal Cortex to Perceptual Integration*, and Fig. S6).

Next, we wanted to establish how the AB and masking affect the neural marker of perceptual integration. In terms of behavior, we observed the classic detrimental effects of both masking (mask vs. no mask, $F_{1,10} = 426.54$, $P < 10^{-8}$) and the AB (short vs. long lag, $F_{1,10} = 51.89$, $P < 10^{-4}$) on accuracy (Fig. 2*F*). There was also an interaction ($F_{1,10} = 52.17$, $P < 10^{-4}$), which was entirely driven by the difference between unmasked long- and short-lag trials (post hoc *t* test, $P < 10^{-4}$).

We hypothesized that, if both masking and the AB impact perceptual integration, they should both affect neural markers of perceptual integration in similar ways. To enable a direct comparison with behavior, we extracted classification accuracy in the four experimental T2 conditions. Fig. 2*D* shows the entire time course, and Fig. 2*E* shows peak classification accuracy at 264 ms (latency taken from T1). A 2 × 2 analysis of variance (ANOVA) showed a highly significant main effect of masking ($F_{1,10} = 37.68$, $P < 0.001$), but no AB effect (short vs. long lag) ($F_{1,10} = 2.16$, $P = 0.172$), and no significant interaction between masking and the AB ($F_{1,10} = 0.02$, $P = 0.963$). Post hoc *t* tests confirmed significant costs for masked vs. unmasked stimuli for both long and short lag (both $P < 0.001$), but no significant differences between long lag and short lag (both $P > 0.25$). Moreover, we were able to show this by training the classifier on an independent RSVP training set that was not confounded by task-related decision or response mechanisms. Note that we confirm these findings in an analysis in which we train the classifier on T1 using all electrodes, intended to investigate the contribution of such mechanisms to the observed data pattern (see Fig. 5).

Thus, although we observe a strong effect of masking in both brain and behavior, the classic AB effect only occurs in behavior. To further statistically underpin the differential effect of conscious access on behavioral and neural measures of perceptual integration, we entered both measurements into a 2 × 2 × 2 ANOVA with factors measure (normalized behavioral/normalized neural), AB (yes/no), and masking (yes/no). The validity of treating neural and behavioral HR − FAR data as repeated measures of the same thing (i.e., classification of a perceptual object) is discussed in *SI Methods*. In line with the other results, this analysis showed a three-way interaction effect driven by differences in behavioral and neural classification accuracies ($F_{1,10} = 9.30$, $P = 0.012$), as well as a two-way interaction between measure and AB ($F_{1,10} = 10.92$, $P = 0.008$) but no interaction between measure and masking ($F_{1,10} = 1.51$, $P = 0.247$). We also provide a confirmatory analysis in which we analyze the neural data contingent on behavioral selection, but would like to stress that such an approach has severe pitfalls and limitations (*SI Results*, *Seen–Unseen Analysis*, and Fig. S7).

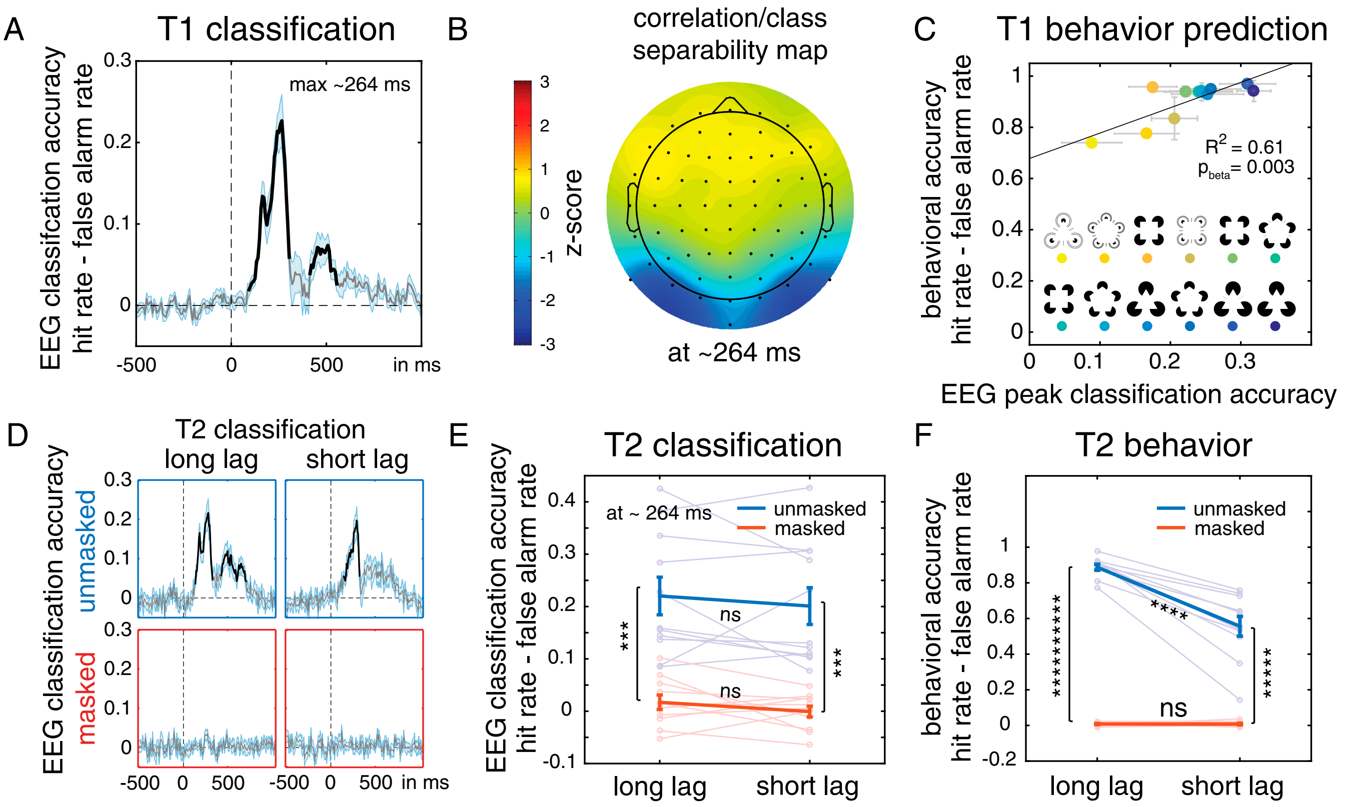


Fig. 2. Peak classification accuracy reflects perceptual integration. (*A*) T1 EEG mean decoding accuracy of perceptual integration over time. Line graphs are average ± SEM in light blue; thick black lines reflect $P < 0.05$, cluster-based permutation test. (*B*) The correlation/class separability map reflecting the underlying neural sources for maximum decoding at ∼264 ms (*SI Methods*). (*C*) The degree to which classification accuracy at ∼264 ms predicts behavioral sensitivity to perceptual integration at T1 for the 12 Kanizsa–control pairs when performing robust linear regression. Each colored data point is a Kanizsa–control pair (only the Kanizsa is shown in this figure; see Fig. S1 for the full figure legend including the control counterparts). (*D*) T2 EEG decoding accuracy over time for the four experimental conditions and (*E*) maximum decoding accuracy at ∼264 ms for these conditions. (*F*) Behavioral sensitivity to perceptual integration for the four conditions (compare with *E*). Error bars are mean ± SEM; individual data points are plotted using low contrast in the background. ns, not significant ($P > 0.05$). ***$P < 0.001$, ****$P < 10^{-4}$, *****$P < 10^{-5}$, **********$P < 10^{-12}$.

These data show that masking disrupts perceptual integration, whereas the AB does not. However, a concern might be that masking wiped out all processing of the stimulus, rather than specifically affecting perceptual integration, resulting in a floor effect. To test this, we reanalyzed the data from the main experiment by selecting a subset of the stimulus set that could be divided orthogonally according to its impact on input energy (contrast) or its impact on perceptual integration (surface perception). Fig. 3*A* illustrates this: the horizontal axis captures differences in perceptual integration (surface perception on the right but not on the left), whereas the vertical axis captures difference in bottom-up energy (high contrast between the inducers and the background vs. low contrast between inducers and background; see Fig. S8 and *SI Methods* for a specification of the entire stimulus set). If masking wipes out all stimulus processing, we should not be able to classify high- vs. low-contrast stimuli. We computed classification accuracy for feature contrast on the one hand and perceptual integration on the other, using a within-condition eightfold cross-validation scheme (collapsing over short and long lag; see *SI Methods* for details).

The results are shown in Fig. 3 *B* and *C*. In an early time window of ∼80–90 ms, both masked and unmasked stimuli showed highly significant classification accuracies for feature contrast [left panels, masked: $t_{(10)} = 7.45$, $P < 10^{-4}$; unmasked: $t_{(10)} = 8.82$, $P < 10^{-5}$, statistics at ∼92 ms, T1 peak latency]. Thus, despite strong masking, the bottom-up signal is processed up to the point of contrast detection. Conversely, masking does wipe out classification accuracy on the perceptual integration dimension [right panels, masked: $t_{(10)} = -0.19$, $P = 0.852$; unmasked: $t_{(10)} = 6.82$, $P < 10^{-4}$]. Note that the same type of masks would follow all stimulus classes (regardless of whether these were Kanizsa, control, high or low contrast), such that the masks themselves could not bias classification accuracy. These results confirm that masking selectively abolishes perceptual integration, leaving feedforward processing largely intact (12–14). In addition, this shows that the reduced classification accuracy for perceptual integration cannot be explained by a generic effect of reduced classification sensitivity under masking.

Another concern might be that EEG classification accuracy is an all-or-none phenomenon, whereas behavior relies on graded evidence. In such a scenario, the behavioral effects on perceptual integration (Fig. 2*F*) might not be reflected in classification accuracy (Fig. 2*E*) due to a lack of sensitivity of the classifier to smaller effects such as those observed during the AB. To test this hypothesis, we conducted a control experiment in which we used a staircase to titrate mask contrast to get a weaker behavioral effect of masking, similar in magnitude to the effect of the AB in Fig. 2*F* (see *SI Methods* for details). Fig. 4*A* shows the resulting behavioral effect of weak masking in this experiment. When computing classification accuracy on these data, we see that it nicely follows behavior (Fig. 4 *B* and *C*) [$t_{(5)} = 3.82$, $P = 0.012$]. Together, these results show that the drops in behavioral accuracy caused by masking and the AB have different root causes: masking impacts perceptual integration directly, whereas the AB leaves it intact. A more detailed treatment can be found in *SI Discussion*, *Mechanisms of Masking and the AB*.

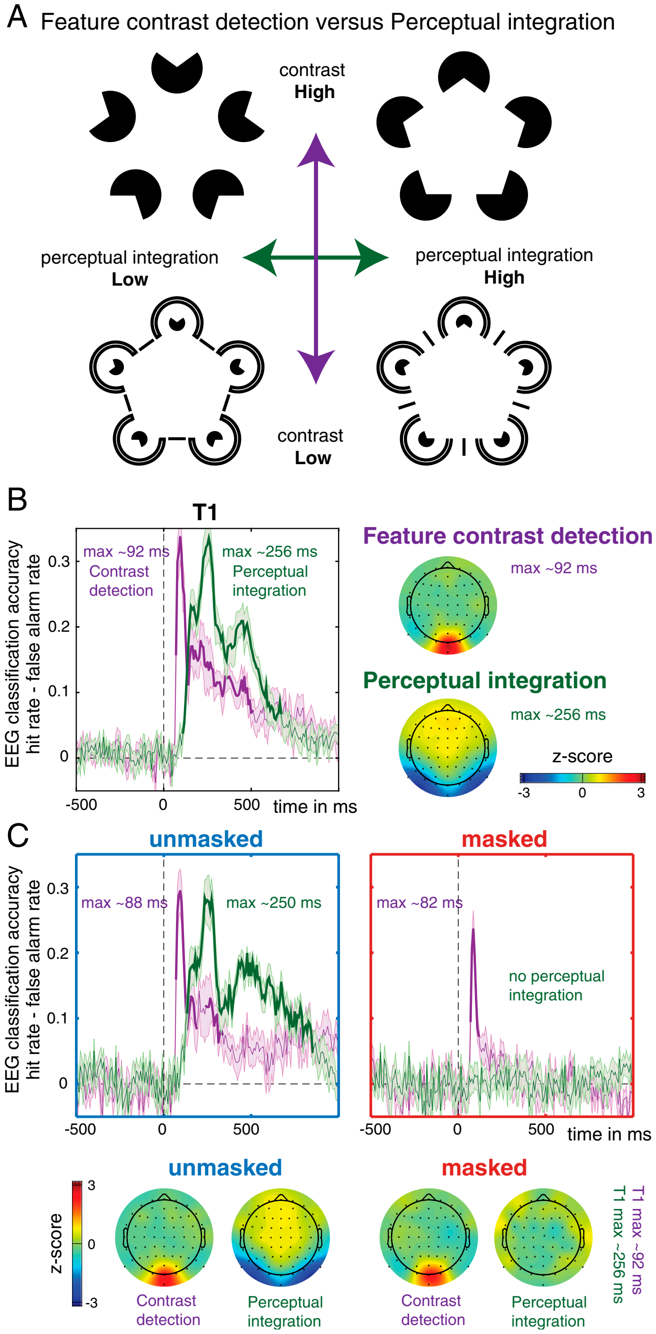


Fig. 3. Separating out perceptual integration and feature contrast detection. (*A*) Example stimuli that were used to orthogonally classify feature contrast and perceptual integration on the same data. (*B*) Classification accuracies across time for contrast detection and perceptual integration (*Left*) as well as correlation/class separability maps (*Right*) for T1, (*C*) and for unmasked (*Left*) and strongly masked trials (*Right*). Line graphs contain mean ± SEM. Thick lines are $P < 0.05$, cluster-based permutation test.

So what neural process causes the dip in behavioral accuracy during the AB? A natural hypothesis would be that the AB interferes with conscious access after perceptual integration has already taken place. If true, we should be able to observe evidence of a selection process that results in conscious access at a later point in time. Investigating this issue requires a classifier that is sensitive to selection. Because the independent training runs that we used for training the classifier in the first analysis were designed to control for the direct influence of decisions and responses, these might not capture such a mechanism. The neural response to T1, however, does involve a conscious decision about the presence of a Kanizsa. We therefore trained a classifier on T1 data and tested it on T2 data, this time using all electrodes as this resulted in better classification accuracy than occipital electrodes only (*SI Methods* and the top row of Fig. S9). Fig. 5 *A* and *B* show classification accuracies for the four experimental conditions when using this T1 classifier.

We again find the initial peak at 264 ms that was described before. Despite the potential contribution of frontal electrodes and decision mechanisms to classification accuracy when training on T1, this peak follows a pattern that is similar to the pattern that we observed when training on the independent training runs using only occipital electrodes (Figs. 2*E* and 5*B*, *Top*), and that does not follow behavioral accuracy (Fig. 2*F*; see *SI Results*, *T1-Based Classification at 264 ms*, for statistical tests). So at what point in time is the behavioral effect of the AB reflected in the neural data? The most notable difference when training on T1 is a second peak in classification accuracy occurring around 406 ms, which is heavily modulated by the AB (Fig. 5*A*). At this time point, the pattern of results is identical to that obtained in behavior (Figs. 2*F* and 5*B*, *Bottom*, as well as Fig. S9, *Bottom*). All manipulations had highly significant effects on classification accuracy: a main effect of AB ($F_{1,10} = 7.96$, $P = 0.018$), a main effect of masking ($F_{1,10} = 130.19$, $P < 10^{-6}$), as well as a strong interaction effect ($F_{1,10} = 14.92$, $P = 0.003$).

To directly compare behavioral to neural data at 406 ms, we again entered the normalized measurements into a 2 × 2 × 2 ANOVA with factors measure (behavioral/neural), AB (yes/no), and masking (yes/no). The results show highly significant main effects of the AB ($F_{1,10} = 23.65$, $P < 0.001$), masking ($F_{1,10} = 528.18$, $P < 10^{-9}$), and a strong interaction effect between AB and masking $F_{1,10} = 51.55$, $P < 10^{-4}$), but importantly no two- or three-way interaction effects with measurement (neural/behavioral, all $F_{1,10} < 3.08$, all $P > 0.110$). This underpins the similarity

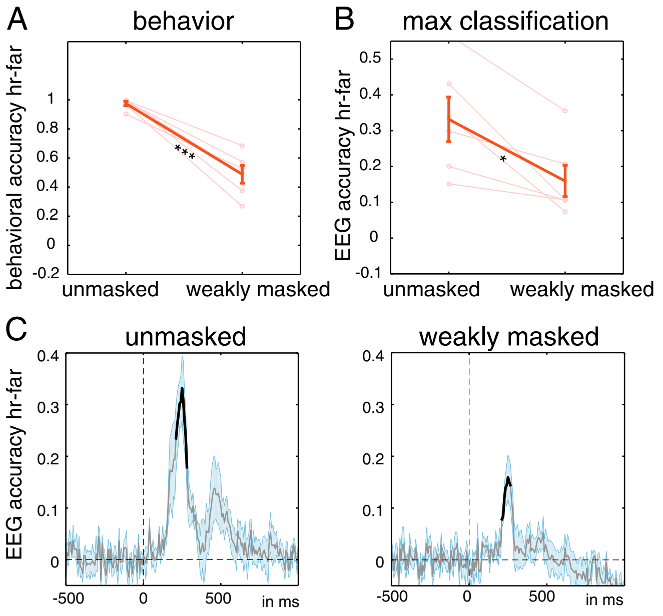


Fig. 4. Masking control experiment. (*A*) Behavioral results. (*B*) Maximum classification accuracy. Error bars are mean ± SEM; individual data points are plotted in light in the background. *$P < 0.05$, ***$P < 0.001$. (*C*) Raw decoding accuracies over time for unmasked and weakly masked conditions. Line graphs contain mean ± SEM; black line reflects $P < 0.05$, cluster-based permutation test.

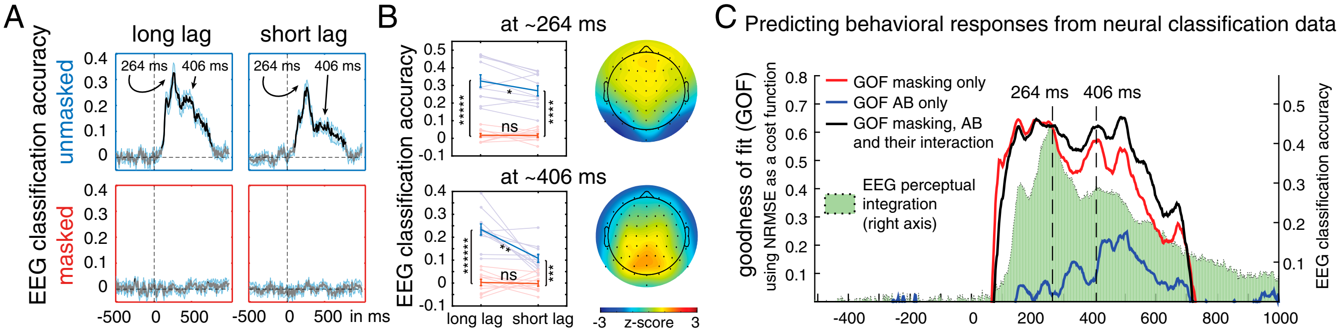


Fig. 5. The impact of masking and AB on perceptual integration over time. (*A*) EEG classification accuracy for the four experimental T2 conditions when training on T1. (*B*) EEG classification accuracies and correlation/class separability maps plotted at peak classification performance at 264 ms (*Top*) and at the second peak at 406 ms (*Bottom*). Blue lines represent the unmasked condition; red lines represent the masked condition. The 406-ms time point follows the same pattern as behavioral accuracy (see main text for statistics) and has a spatial distribution that is homologous to that of a classical P300. ns, not significant ($P > 0.05$). $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 10^{-4}$, $^{*****}P < 10^{-5}$, $^{******}P < 10^{-6}$. (*C*) An estimation of the GOF when using the normalized EEG classification accuracy data as a model for the normalized behavioral detection data (left axis). Datasets are either collapsed over the AB dimension (GOF masking), over the masking dimension (GOF AB), or without collapsing over either dimension (GOF masking, AB, and their interaction). T1 classification accuracy is plotted as a green shade in the background for reference (right axis). Not until after the perceptual integration signal has peaked at 264 ms does the black line overtake the red line, showing a postperceptual contribution of the AB to behavioral accuracy.

between the behavioral and neural data pattern at this time point. The correlation/class separability map at 406 ms (Fig. 5*B*, *Bottom*) has the same topology as that of a classical P300 (or P3b), which has been frequently associated with conscious access and perceptual decision-making (e.g., ref. 18). These data provide converging evidence that neural signals around the time frame of the P300 reflect a postperceptual signal that reflects conscious access, rather than perceptual integration itself. What we unambiguously show is that perceptual integration precedes such conscious access.

In a statistical sense, we have so far regarded behavioral and classification accuracy data as repeated measures of the same underlying perceptual object. Another approach would be to assess the degree to which the neural data are able to serve as a model for behavior across time. To do this, we used normalized classification accuracies from the T1 classifier as reference points to determine the goodness of fit (GOF) with normalized behavioral accuracies as test data. As a measure of GOF, we used the normalized root mean square error (NRMSE) cost function given by the following:

$$fit(t) = 1 - \frac{||xref(:,t) - x(:)||}{||xref(:,t) - mean(xref(:,t))||},$$

where x denotes the test data (behavioral accuracy), $xref$ denotes the neural data (classification accuracy), || indicates the 2-norm (Euclidean length) of a vector, fit is a row vector of length Nt and $t = 1, ..., Nt$, where Nt is the number of time points. NRMSE costs vary between –Infinity (bad fit) to 1 (perfect fit). If the GOF cost function is equal to zero, then x is no better than a straight line at matching $xref$. We obtained this fitness measure separately for the different factors by collapsing the neural and behavioral data either across the masking factor, across the AB factor, or without regard to either factor (*SI Methods*). The results are shown in Fig. 5*C*, where we also plot T1 classification accuracy as a reference for the time course of perceptual integration. Fig. 5*C* confirms that, up to 264 ms, the masking manipulation uniquely models (predicts) behavior, indeed better than when the AB is also allowed to contribute to the fit. Only after 264 ms does the AB start to contribute to behavioral outcomes, trailing the perceptual integration signal itself and in line with prior analyses.

Discussion

We show that EEG can be used to decode the presence of integrated percepts in visual cortex. Furthermore, we show that masking obliterates behavioral accuracy and classifier performance. Because the ability to decode feature contrast is retained under masking, the effects of masking on perceptual integration cannot be attributed to generic effects of masking on the sensitivity of the classifier. Rather, masking selectively disrupts perceptual integration while leaving feedforward signals intact (12–14). Interestingly, however, peak classification performance on integration remains unchanged during the AB, despite causing a marked dip in behavioral accuracy. This shows that the brain is able to integrate features into perceptual objects when conscious access is impaired. These results seem to fit nicely with early findings showing semantic effects without conscious access (17) and more recent findings that the meaning of multiple words can even be integrated unconsciously to reflect semantic valence (21), but see *SI Discussion*, *How Does Perceptual Integration Relate to Semantic Integration of Words*, for a more nuanced treatment.

The idea that perceptual integration can occur without conscious access is seemingly at odds with experiments on object-based attention. For example, in an experiment by Roelfsema et al. (22), monkeys were trained to perform a curve-tracing task. Attention to the task-relevant curve resulted in a spreading activation across V1 neurons that coded the features belonging to the curve, thus binding the constituent elements of the curve together. This suggests that serial access is the glue that unites an object, in line with the classical framework put forward by Treisman and Gelade (23), and inconsistent with the position that perceptual integration can occur without conscious selection. With some exceptions, however (e.g., refs. 24 and 25), attention and conscious access are usually conflated, precluding their disentanglement. Hence, the perceived relationship between conscious access and perceptual integration may be mediated by attention, such that perceptual integration may occur without conscious access as long as stimuli are attended and/or intrinsically task relevant. Importantly, we do not claim that perceptual integration is not modulated by attention or task relevance.

Rather, we show that Kanizsa figures can be integrated in visual cortex despite not being promoted to a consciously accessible state. In contrast, masking destroys perceptual integration regardless of task demands. Naturally, this difference must be reflected in neural mechanisms. Dynamic feature grouping that underlies perceptual integration is thought to rely on corticocortical feedback. Although much remains to be learned about the origin of these feedback signals, evidence suggests that they originate from within visual cortex and thus are local in nature

(14–16, 26–28). Although conscious access also involves feedback, such long-range feedback originates from frontoparietal cortex (18). In the consciousness literature, long-range integration is often referred to as "global ignition" occurring in the time frame of the P300 at 400–500 ms (4), whereas local integration within visual cortex is associated with activity in the 200- to 300-ms time frame (29). The current data suggest that perceptual integration occurs in this early time frame and does not require global ignition.

These results also speak to a current debate about whether consciousness overflows cognitive access (30, 31). In this debate, the question is whether access causes representational content to be extracted, or whether it acts to select from a rich representational set that cannot be accessed in its entirety. In support of the latter position, retrocueing studies show that the representational capacity in early visual cortex is much larger than what can be accessed at any given moment (e.g., ref. 32). A recent study has questioned these results, suggesting that a retrocue might serve to postdictively impact perception (33). The current study resolves this issue by using a direct neural measure of perceptual integration to show that perceptual integration precedes conscious access. We provide a more detailed treatment of the implications of these results in *SI Discussion*, *Implications for Global Neuronal Workspace Theory and the Debate About the Existence of Phenomenal vs. Access Consciousness*.

This study was performed in accordance with the Declaration of Helsinki under approval of the ethics committee of the Faculty of Social and Behavioral Sciences, Utrecht University. All participants signed an informed consent.

ACKNOWLEDGMENTS. We thank Anouk van Loon for helpful comments during this project.

1. Helmholtz HV (1867) *Handbuch der Physiologischen Optik* (Leopold Voss, Leipzig, Germany).
2. Block N (2005) Two neural correlates of consciousness. *Trends Cogn Sci* 9(2):46–52.
3. Lamme VAF (2010) How neuroscience will change our view on consciousness. *Cogn Neurosci* 1(3):204–220.
4. Dehaene S, Changeux JP, Naccache L, Sackur J, Sergent C (2006) Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends Cogn Sci* 10(5): 204–211.
5. Baars BJ (2005) Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience. *Prog Brain Res* 150:45–53.
6. Ginsburg AP (1975) Is the illusory triangle physical or imaginary? *Nature* 257(5523): 219–220.
7. Laeng B, Endestad T (2012) Bright illusions reduce the eye's pupil. *Proc Natl Acad Sci USA* 109(6):2162–2167.
8. Davis G, Driver J (1994) Parallel detection of Kanizsa subjective figures in the human visual system. *Nature* 371(6500):791–793.
9. Vuilleumier P, Landis T (1998) Illusory contours and spatial neglect. *Neuroreport* 9(11): 2481–2484.
10. Mattingley JB, Davis G, Driver J (1997) Preattentive filling-in of visual surfaces in parietal extinction. *Science* 275(5300):671–674.
11. Vandenbroucke ARE, Fahrenfort JJ, Sligte IG, Lamme VAF (2014) Seeing without knowing: Neural signatures of perceptual inference in the absence of report. *J Cogn Neurosci* 26(5):955–969.
12. Harris JJ, Schwarzkopf DS, Song C, Bahrami B, Rees G (2011) Contextual illusions reveal the limit of unconscious visual processing. *Psychol Sci* 22(3):399–405.
13. Kovács G, Vogels R, Orban GA (1995) Cortical correlate of pattern backward masking. *Proc Natl Acad Sci USA* 92(12):5587–5591.
14. Fahrenfort JJ, Scholte HS, Lamme VAF (2007) Masking disrupts reentrant processing in human visual cortex. *J Cogn Neurosci* 19(9):1488–1497.
15. Roelfsema PR (2006) Cortical algorithms for perceptual grouping. *Annu Rev Neurosci* 29:203–227.
16. Wyatte D, Jilk DJ, O'Reilly RC (2014) Early recurrent feedback facilitates visual object recognition under challenging conditions. *Front Psychol* 5:674.
17. Luck SJ, Vogel EK, Shapiro KL (1996) Word meanings can be accessed but not reported during the attentional blink. *Nature* 383(6601):616–618.
18. Sergent C, Baillet S, Dehaene S (2005) Timing of the brain events underlying access to consciousness during the attentional blink. *Nat Neurosci* 8(10):1391–1400.
19. Haufe S, et al. (2014) On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage* 87:96–110.
20. Rousseeuw PJ, Leroy AM (2005) *Robust Regression and Outlier Detection* (John Wiley & Sons, Inc., New York).
21. van Gaal S, et al. (2014) Can the meaning of multiple words be integrated unconsciously? *Philos Trans R Soc Lond B Biol Sci* 369(1641):20130212.
22. Roelfsema PR, Lamme VAF, Spekreijse H (1998) Object-based attention in the primary visual cortex of the macaque monkey. *Nature* 395(6700):376–381.
23. Treisman AM, Gelade G (1980) A feature-integration theory of attention. *Cognit Psychol* 12(1):97–136.
24. Kentridge RW, Heywood CA, Weiskrantz L (1999) Attention without awareness in blindsight. *Proc Biol Sci* 266(1430):1805–1811.
25. Naccache L, Blandin E, Dehaene S (2002) Unconscious masked priming depends on temporal attention. *Psychol Sci* 13(5):416–424.
26. Fahrenfort JJ, et al. (2012) Neuronal integration in visual cortex elevates face category tuning to conscious face perception. *Proc Natl Acad Sci USA* 109(52):21504–21509.
27. Kok P, Bains LJ, van Mourik T, Norris DG, de Lange FP (2016) Selective activation of the deep layers of the human primary visual cortex by top-down feedback. *Curr Biol* 26(3):371–376.
28. Wokke ME, Vandenbroucke AR, Scholte HS, Lamme VA (2013) Confuse your illusion: Feedback to early visual cortex contributes to perceptual completion. *Psychol Sci* 24(1):63–71.
29. Koivisto M, Salminen-Vaparanta N, Grassini S, Revonsuo A (2016) Subjective visual awareness emerges prior to P3. *Eur J Neurosci* 43(12):1601–1611.
30. Block N (2011) Perceptual consciousness overflows cognitive access. *Trends Cogn Sci* 15(12):567–575.
31. Phillips I (2016) No watershed for overflow: Recent work on the richness of consciousness. *Philos Psychol* 29(2):236–249.
32. Sligte IG, Vandenbroucke ARE, Scholte HS, Lamme VAF (2010) Detailed sensory memory, sloppy working memory. *Front Psychol* 1:175.
33. Sergent C, et al. (2013) Cueing attention after the stimulus is gone can retrospectively trigger conscious perception. *Curr Biol* 23(2):150–155.
34. Kaernbach C (1991) Simple adaptive testing with the weighted up-down method. *Percept Psychophys* 49(3):227–229.
35. Delorme A, Makeig S (2004) EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J Neurosci Methods* 134(1):9–21.
36. Oostenveld R, Fries P, Maris E, Schoffelen JM (2011) FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput Intell Neurosci* 2011:156869.
37. Maris E, Oostenveld R (2007) Nonparametric statistical testing of EEG- and MEG-data. *J Neurosci Methods* 164(1):177–190.
38. Huber PJ (1973) Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann Stat* 1(5):799–821.
39. Aru J, Bachmann T, Singer W, Melloni L (2012) Distilling the neural correlates of consciousness. *Neurosci Biobehav Rev* 36(2):737–746.
40. Rahnev D, et al. (2011) Attention induces conservative subjective biases in visual perception. *Nat Neurosci* 14(12):1513–1515.
41. King JR, Dehaene S (2014) A model of subjective report and objective discrimination as categorical decisions in a vast representational space. *Philos Trans R Soc Lond B Biol Sci* 369(1641):20130204.
42. Enns JT, Di Lollo V (2000) What's new in visual masking? *Trends Cogn Sci* 4(9):345–352.
43. Kelly AJ, Dux PE (2011) Different attentional blink tasks reflect distinct information processing limitations: An individual differences approach. *J Exp Psychol Hum Percept Perform* 37(6):1867–1873.
44. Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. *Cereb Cortex* 1(1):1–47.
45. Kouider S, Dehaene S (2007) Levels of processing during non-conscious perception: A critical review of visual masking. *Philos Trans R Soc Lond B Biol Sci* 362(1481):857–875.
46. Sergent C, Dehaene S (2004) Is consciousness a gradual phenomenon? Evidence for an all-or-none bifurcation during the attentional blink. *Psychol Sci* 15(11):720–728.
47. van Gaal S, Ridderinkhof KR, Fahrenfort JJ, Scholte HS, Lamme VAF (2008) Frontal cortex mediates unconsciously triggered inhibitory control. *J Neurosci* 28(32):8053–8062.
48. Gazzaniga MS, Bogen JE, Sperry RW (1962) Some functional effects of sectioning the cerebral commissures in man. *Proc Natl Acad Sci USA* 48(10):1765–1769.
49. Marshall JC, Halligan PW (1988) Blindsight and insight in visuo-spatial neglect. *Nature* 336(6201):766–767.
50. Overgaard M, Rote J, Mouridsen K, Ramsøy TZ (2006) Is conscious perception gradual or dichotomous? A comparison of report methodologies during a visual task. *Conscious Cogn* 15(4):700–708.
51. Pitts MA, Metzler S, Hillyard SA (2014) Isolating neural correlates of conscious perception from neural correlates of reporting one's perception. *Front Psychol* 5:1078.
52. Tononi G (2008) Consciousness as integrated information: A provisional manifesto. *Biol Bull* 215(3):216–242.
53. Greenwald AG, Klinger MR, Liu TJ (1989) Unconscious processing of dichoptically masked words. *Mem Cognit* 17(1):35–47.
54. Holender D (1986) Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: A survey and appraisal. *Behav Brain Sci* 9:1–23.
55. Dehaene S, et al. (1998) Imaging unconscious semantic priming. *Nature* 395(6702): 597–600.
56. McClelland JL, Rumelhart DE (1981) An interactive activation model of context effects in letter perception. 1. An account of basic findings. *Psychol Rev* 88(5):375–407.
57. Fiez JA, Petersen SE (1998) Neuroimaging studies of word reading. *Proc Natl Acad Sci USA* 95(3):914–921.

Supporting Information

Fahrenfort et al. 10.1073/pnas.1617268114

SI Methods

Participants. Sixteen right-handed participants (nine females, mean age of 24 y; and seven males, mean age of 25 y) participated in the study for monetary reward. All participants had normal, or corrected-to-normal, vision (mean left eye, 2020; mean right eye, 2030).

Stimuli. We constructed a set of 12 Kanizsa–control pairs of different shapes: triangles, squares, and pentagons (Fig. S1). Traditionally, controls are created by rotating the inducers outwardly. Although such controls retain the overall configuration of the inducers, they allow for Kanizsa recognition using low-pass spatial frequency filters (6). Moreover, the support ratio of the stimulus (the ratio between the physically specified side length and illusory side length) is obliterated in such controls. We therefore constructed controls in which one or more of these characteristics were optimally matched with their Kanizsa counterpart; keeping the shapes of the inducers intact (see numbers 1, 2, 3, 4, 6, 7, 9, 10, and 11 in Fig. S1), retaining the low spatial frequency characteristics of the global stimulus (see numbers 1, 4, 5, 8, and 12 in Fig. S1), as well as maximizing the support ratio of controls (see numbers 1, 2, 3, 4, 5, 6, 8, and 11 in Fig. S1) compared with their Kanizsa counterparts. In cases where large inducers were rotated outwardly (see numbers 1, 6, 7, 9, 10, and 11 in Fig. S1), we rotated the inducers around their center of gravity rather than around their "veridical" center, so as to further minimize differences between Kanizsas and controls in terms of their global spatial-frequency characteristics.

As a result, the stimulus set contained large variations in terms of physical properties across stimulus instances but had similar physical properties within any given Kanizsa–control pair. Because the classification analyses involve single-trial extraction of class membership that needed to carry over from one stimulus instance to the next to be able to work (i.e., the task was to classify stimuli based on the existence of surface information, irrespective of the physical features of the inducers or the shape of the configuration of the inducers), differences we observed in the Kanizsa–control dimension could not be explained by any single physical property but were particular to differences resulting from perceptual integration. Put differently: neither the subjects nor the classifier could solve the task by using particular features of any of the inducers; the only way of solving the classification task was to perceptually integrate the features and establish the existence of surface information to determine class membership. Finally, the total region covered by Kanizsa figures (including inducers) was 9.4° × 8.5° (degrees visual angle) for triangles, 7.4° × 7.4° for squares, and 7.7° × 7.7° for pentagons, keeping only the size of the illusory surface region approximately the same between shape types.

Masks were created by randomly rotating inducer elements from the Kanizsa and control images (Fig. S2). There were 10 masks for each stimulus shape. Masks were picked randomly from these sets, but always matching masks to shapes, so that triangular Kanizsas and controls would be followed by triangularly organized masks, square Kanizsas and controls by square masks, and pentagonal Kanizsas and controls by pentagonal masks. All stimuli and masks were generated using Adobe Illustrator CS6 (Adobe Systems).

Procedure and Tasks of Main Experiment. All tasks were programmed in Presentation (Neurobehavioral Systems) and displayed on a 19-inch CRT-monitor running at 100 Hz. Subjects participated in a total of three sessions. The first session was a training session to make subjects familiar with the task and the stimulus set. In this task, Kanizsa and control images were presented for 10 ms, and participants were prompted to identify whether the image contained a surface. When they were able to perform this task with an accuracy of more than 90%, they continued with the next task. In the second part of the practice session, participants performed a no-blink (long lag) version of the experimental task to determine whether they were able to correctly identify black T1 and T2 targets amid an RSVP of red distractors. Subject performance was computed as hit rate (HR) (the fraction of Kanizsa figures categorized as containing a surface) minus false-alarm rate (FAR) (the fraction of control figures categorized as containing a surface). If their HR minus FAR exceeded 0.8 in both T1 and T2, they went on to the third and final part of the practice session. In this part, they performed two versions of the experimental task to determine at what latency the T2 induced the largest AB for that subject.

The task was the same as the experimental task but did not include the masked conditions. The difference between these two versions was the interstimulus interval (ISI). In the first version, the ISI was 150 ms (resulting in a short AB lag of 300 ms), whereas the second version had an ISI of 100 ms (resulting in a short AB lag of 200 ms). If the participants were not able to perform adequately in one of the tasks or did not show a sufficiently strong AB, they were excluded from the rest of the experiment. Eight participants performed the EEG sessions at an ISI of 150 ms (short AB lag: 300 ms), three participants did the task at an ISI of 100 ms (short AB lag: 200 ms), and five participants were excluded after the first training session for not meeting one or more of the above criteria for inclusion.

Subsequently, subjects took part in two separate sessions on separate days, in which they performed the experimental task while their EEG was recorded. The experimental task consisted of an RSVP in which they had to detect two targets (T1 and T2, see Fig. 1*B*). The first 8–15 RSVP elements before T1 were red distractors, either followed by one red distractor (short lag) or by four to seven red distractors (long lag). Another five to seven distractors would appear between T2 and the response screen, asking subjects to indicate whether the first and/or second target contained a surface. One-half of the T2 targets was followed by a strong high-contrast mask, and the other half was followed by a weak low-contrast mask. The weak mask was only used to make the conditions as comparable as possible, but not intended to impede visibility. Subjects gave two responses, the first for T1, and the second for T2. Responses consisted of button presses using a two-button box attached to the right arm of the chair, with the left button indicating "I perceived a surface" and the right button indicating "I did not perceive a surface." Each of the two sessions consisted of nine blocks of this experimental task. Each block consisted of 24 Kanizsa and 24 control images for each of the four conditions resulting in 192 trials per block. Across both experimental sessions, the participants performed a total of 3,456 trials of the experimental task.

In addition to the experimental blocks, they also performed a 1-back RSVP task, which was used to train the multivariate discriminant classifier. In this 1-back task, black images (Kanizsa and control) and red distractor images were displayed in an RSVP, interleaved with one another, at an ISI of 1,000 ms (±50-ms jitter). Each image was displayed for 10 ms (Fig. S3). Image type (Kanizsa or control) was randomized, with three randomly occurring repetitions in every 10 black images. Participants were required to press a button every time a black image repeated itself while ignoring the red images. There was no relationship between stimulus type (Kanizsa or control) and image repetition, the task was purely intended to keep attention focused on the screen, and the behavioral data were not analyzed further. Over both experimental sessions, participants performed 1,152 trials of this task, split across eight blocks.

Procedure and Tasks of the Masking Control Experiment. Six subjects from the main experiment took part in the masking control experiment (see main text for task rationale), which consisted of one EEG session. The difference with the main experiment was that the AB manipulation was not included and that the strong mask condition was replaced by a variable weak-mask condition. The experiment was identical with respect to timing and response method. Before testing, each subject performed a staircase to titrate the contrast of the weak masks so that subject performance was the same as their performance in the unmasked short-lag AB condition of the main experiment. After the staircase, subjects performed nine blocks of the control experiment (1,728 trials) while their EEG was recorded. In addition, they performed four blocks of the same RSVP 1-back task as in the main experiment (576 trials).

We used a double-staircase procedure using the weighted up-down method (34). Masks were presented on a white background. Contrast was adjusted by changing the intensity of the masks. One staircase started out at the minimum contrast, and the other started at the maximum contrast. The staircase was updated only on trials with a Kanizsa figure: detection of the Kanizsa (hit) increased the difficulty (S_{down}), and indicating absence of a Kanizsa (miss) decreased the difficulty (S_{up}). The step size with which mask contrast was changed, was determined using the weighted rule $S_{\mathrm{up}}{}^{*}p = S_{\mathrm{down}}{}^{*}(1 - p)$, in which S_{up} is the upward step size corresponding to a decrease of mask contrast, whereas S_{down} is the downward step size corresponding to an increase of mask contrast, and p is the percentage correct onto which the staircase should converge. For example, if a subject had a HR of 0.7 in the short-lag condition of the main experiment, the relationship between the two step sizes would be $S_{\mathrm{up}}{}^{*}0.7 = S_{\mathrm{down}}{}^{*}0.3$, rounding off to the nearest available values that fit given the available contrast levels (there were 20 available contrast steps between minimal and maximal).

The staircase ended after 12 reversals. The median reversal contrast for both staircases was used as starting point for mask contrast. During the experimental blocks, mask contrast was updated after each block, based on the behavioral performance in the previous block. The updating was done to keep the behavioral performance as close as possible to the unmasked short-lag condition in the main experiment. Updating was rare; for four of the subjects, mask contrast was adjusted only twice (on the first two blocks). One subject had mask contrast adjusted once (after the first four blocks), and one subject did not have the mask contrast adjusted at all.

Behavioral Analysis. Where applicable, all reported statistical tests are double sided. Responses were scored as hits (Kanizsa correct) misses (Kanizsa incorrect) correct rejections (control correct) and false alarms (control incorrect). The HR was computed as the fraction of Kanizsa figures categorized as containing a surface, whereas the FAR was computed as the fraction of control figures categorized as containing a surface. Behavioral performance was computed as HR minus FAR for each of the conditions to determine how well they performed on the task. Repeated-measures ANOVAs were used to detect main and interaction effects of the conditions.

EEG Data Collection and Preprocessing. EEG data were collected at 2,048 Hz using a 64-channel ActiveTwo system (BioSemi). EEG data analysis was performed using Matlab (MathWorks), the EEGLAB toolbox (35), and custom-written Matlab scripts to perform multivariate classification.

All data were referenced to the average of the mastoids, downsampled to 512 Hz, and epoched between −500 and 1,000 ms. Trials containing muscle artifacts were removed using an adapted version of the ft_artifact_zvalue muscle artifact detection function taken from the FieldTrip toolbox (36). This function applies a frequency filter between 110 and 140 Hz and assigns a Z value to each time point to ascertain the degree to which power values in that frequency range deviate from normality. Trials that contained Z-score outliers more than 3 SDs away from the absolute value of the minimum negative Z value were discarded. Next, the data were high-pass filtered at 0.1 Hz. No low-pass filtering was applied.

We did not apply baseline correction to the T2 data obtained from the main AB/mask experiment, as baseline correction introduces unwanted confounding effects on short-lag vs. long-lag trials. There are two potential ways of performing baseline correction in this experimental design: (*i*) either one chooses a fixed baseline time window before a T2 target or (*ii*) one applies a baseline that comes from a fixed time window before T1. Both approaches are problematic. The first approach only works when picking a clean baseline period before trial onset (so before T1), keeping the distance between baseline and T2 fixed (which would result in a different baseline time window, depending on whether T2 was a short- or long-lag trial). However, this would have required an extremely long clean intertrial interval, which given the long trial sequence we already had was not feasible. When picking a baseline window that is closer to T2, the baseline period would overlap with T1 or with the T1–T2 lag period depending on whether it is a short- or long-lag T2. In that case, task-related activity during the baseline period would get introduced into the T2 period. The second approach is also problematic because the period between the baseline period and T2 onset would be different for short- and long-lag trials, allowing long-lag trials to drift off more than short-lag trials. We investigated this and confirmed that such a procedure indeed artificially boosts the short-lag T2 signal compared with the long-lag T2 signal, counteracting a potential impact of the AB. Instead, we therefore performed a 0.1 high-pass filter, which takes slow drifts out of the signal, similar to performing a baseline correction, but does not have any of the aforementioned problems. However, we did perform baseline correction on the RSVP and T1 training sets, and on the masking testing set from the control experiment, because none of these carry the T2-specific baseline problems outlined above. When baseline correction was carried out, it was always applied on the period of −250 to 0 ms before stimulus onset.

Finally, we ran a number of control analyses to ascertain the influence of eyeblinks on the classification analysis, using both an independent component analysis to remove eyeblink components as well as using a procedure to remove all trials containing eyeblinks altogether. Neither procedure had quantitative or qualitative effects on any of the classification results compared with leaving the eyeblinks in, so we opted to retain the signal in its original form and not remove eyeblinks.

EEG Multivariate Pattern Analyses. For each participant, we applied a backward decoding classification algorithm either using the independent RSVP data for training, using the T1 data for training, or using an eightfold cross-validation scheme (explained further down below). In all analyses, we trained a linear discriminant classifier to discriminate Kanizsa and control images using the raw EEG activity across electrodes as the features used for classification. Next, we computed classification accuracy of the classifier as the HR (the fraction of Kanizsa figures that were classified as Kanizsa) minus the FAR (the fraction of control figures that were classified as Kanizsa) for each subject, and for each of the conditions: T1, masked AB, unmasked AB, masked without AB, and unmasked without AB. The procedure was executed for every time sample in a trial, yielding the evolution of classification accuracy over time for each of the conditions. All statistical tests were double-sided t tests across subjects of classification accuracy (HR − FAR) against zero. When plotting significant intervals over time, t tests were corrected for multiple comparisons using cluster-based permutation testing (1,000 iterations, at a threshold of 0.05). In this procedure, the sum of the t values in the observed cluster of contiguously significant data points is compared with the sum of the cluster of contiguously significant data points under random permutation. The P value is the number of times that the cluster sum under permutation exceeds that of the observed cluster sum, divided by the number of iterations (37).

Because the classifier weights that result from the training procedure result from a backward model, they do not unambiguously reflect neural sources. They may have small amplitudes for electrodes containing the signal of interest, but also large amplitudes at electrodes not containing this signal, and may therefore result in both type I and type II errors. To mitigate this problem, we obtained topographic maps by using a method recently described by Haufe et al. (19), in which the classifier weights are multiplied by the data correlation matrix (Fig. S4*A*, *Left*, for classifier weights). This operation creates a correlation/class separability map (Fig. S4*A*, *Right*) that generates interpretable neural sources for which nonzero activity is only observed at channels for which the task-related signal is both strong and highly correlated with the task, while at the same time minimizing the influence of potential artifacts.

We normalized both the weight and class/correlation separability maps across electrodes for each subject, to be able to compute topographic plots of condition averages across subjects. Fig. S4*A* provides a direct comparison between classifier weights and the correlation/class separability map. Perhaps unsurprisingly, the effect of perceptual integration was strongly occipital in nature. Because the occipital electrodes yielded nonzero classifier weights (Fig. S4*A*) and the highest classification accuracies (Fig. S4*B*), we restricted the initial analyses of the experimental conditions by using only occipital electrodes as features for classification (PO7, PO3, O1, Iz, Oz, POz, PO8, PO4, and O2) to ensure that any effects we observed were not due to poor signal-to-noise ratio. Control analyses revealed that using all electrodes did not change any of the effects that we observed.

Next, we used robust linear regression to characterize the relationship between peak accuracy of the classifier and behavioral accuracy at T1, using the 12 Kanizsa–control pairs as data points (Fig. S1). Robust linear regression guards against violations of assumptions that are required for standard regression, as well as the unwanted influence of outliers (38). This analysis underpins the validity of viewing peak classification accuracy as a neural measure for perceptual integration, evidenced by its strong predictive power of the behavioral response regarding surface perception. For more details regarding this analysis, see *SI Results*, *Prediction of Behavioral Accuracy Based on Neural Classification Accuracy*.

To be able to compare the differential effect of the four T2 conditions under behavioral and neural measures of perceptual integration (HR − FAR), we entered the measurements into a 2 × 2 × 2 ANOVA of measure (normalized behavioral/normalized neural), AB (yes/no), and masking (yes/no). The normalization step Z-scores the data, separately within the behavioral and within the neural matrix, subtracting their respective means and dividing by their respective SDs. It is important to realize that this normalization step does not change any of the statistics that result from the initial 2 × 2 (masking yes/no ×AB yes/no) ANOVA analyses. Whether entering the normalized or the nonnormalized data into such an analysis, all F statistics, P values, and all other aspects of the analysis remain the same. The only thing that changes when entering both of these normalized matrices into a large 2 × 2 × 2 ANOVA, is that any main effects of measure fall out because the measure means have been subtracted out.

The rationale for doing this is that we are not interested in main effects of measure, which is differentially affected by the signal-to-noise ratio for behavioral and EEG data. Rather, we want to know whether the pattern that we observe under behavioral and neural measures is the same or not, which can be obtained by looking at the interaction between measure (normalized behavioral/normalized neural) and the other factors. Whether one can regard normalized behavioral and neural measures as repeated measures of the same perceptual object can best be understood by drawing an analogy. Let us say we want to know whether there is a differential effect of X on Y at night and during the day, but there is an overshadowing main effect on these measurements during daytime and nighttime that we are not interested in (simply because there is more light during the day, our measurement is affected by this). In such a case, it would be valid to separately normalize the measurements during day and during night, removing the main day–night effect on the measurements to see whether there is an interaction between factor X and moment of measurement (day/night). This is essentially what we do here by regarding the behavioral and neural measures of perceptual integration as repeated measures of the same thing, albeit with different overall averages. An interaction between that factor and the other factors shows that the underlying data pattern is not the same for the two measurements, which suggests that the experimental manipulations impact behavioral markers differently from neural markers.

In the next analysis, we looked at the degree to which a classifier would be able to determine class membership regarding high or low contrast on the one hand and high or low perceptual integration on the other (see Fig. S8 for the stimulus set), under masking and no-masking conditions (collapsing across AB and no AB trials). Because any potential contribution of decision mechanisms was irrelevant in this analysis (subjects did not have to respond to feature contrast), we used an eightfold training–testing algorithm. In this scheme, we first removed information about the order in which trials were acquired during the experiment by randomizing the order in which trials were stored on disk. Next, we split up the dataset into eight equally sized subsets. Subsequently, a linear discriminant classifier was trained to discriminate between stimulus classes using seven-eighths of the data, and was tested on the remaining one-eighth of the data, thereby ensuring independence of training and testing sets, repeating that scheme until all data were used for testing once, but never using the same data for training and testing in one train–test cycle. To obtain final accuracy scores, we averaged across the eight iterations. As before, the EEG activity at individual electrodes was used as features for classification and the cross-validation procedure was executed for every time sample in a trial, yielding the evolution of classification accuracy over time.

Finally, we wanted to determine the point in time at which neural signals could best explain the behavioral results. For this analysis, rather than controlling for the influence of decision mechanisms as we did initially, we now wanted to include this influence on classification accuracy. Therefore, we used the T1 data as training set for the linear discriminant. Because decision mechanisms and conscious access are known to involve frontal cortex (4, 18), we went back to including all electrodes in this analysis. A control analysis confirmed that, when training on T1, classification accuracy was indeed better for all electrodes compared with restricting to occipital electrodes (Fig. S9, *Top*; compare Fig. S4*A*, where the reverse is the case). All final analyses are therefore executed on T1 trained data, using all electrodes. Again, we performed 2 × 2 ANOVAs on the behavioral and neural data as before, and again we performed large 2 × 2 × 2 ANOVAs, which include the normalized behavioral and normalized neural data as a repeated measure (see Fig. S9, *Bottom*, for normalized responses).

To further fully characterize the moment in time at which the neural data are able to explain the behavioral data, we quantified the degree to which the neural data can serve as a model for the behavioral data using a goodness of fit on the behavioral data, taking the neural data as a reference (see main text for details). We computed this measure on the normalized neural and behavioral data, using the same rationale for normalization as before. Goodness of fit was calculated for every time point of the neural data, using a 40-ms moving average (we used a forward-looking moving average to maintain liberal estimates of fit onsets). This was done separately for the masking factor, the AB factor, and for all data. The masking factor was computed by averaging accuracy scores across the AB conditions, the AB factor was computed by averaging across the masking conditions, and the total data (masking plus attention plus their interaction) was computed by averaging across pairs of values within each condition. Using this averaging procedure, the total number of points was kept constant for each estimation, while still being able to generate separate estimates for masking, AB, and all data. However, given the uneven number of subjects ($n = 11$), we could not create a balanced set when averaging within conditions for the total data. Therefore, the procedure was repeated 11 times for all fit types (masking, AB, and total), leaving out a subject at each iteration to acquire an even number, and then averaging over the 11 resulting fits to obtain the final values.

SI Results

Prediction of Behavioral Accuracy Based on Neural Classification Accuracy. The 12 Kanizsa–control pairs (Fig. S1) differ in the degree to which they result in perceptual integration. This is reflected in variations in behavioral accuracy at distinguishing between a Kanizsa and its control across these pairs. To establish a direct link between EEG classification accuracy and perceptual integration, we used variations in peak classification accuracy to predict variations in behavioral accuracy. To obtain classification accuracies for the 12 pairs, we used the same classifier as was used in the other analyses (see Fig. S3 for the training task). Importantly, we did not train separate classifiers for separate Kanizsa–control pairs; only the testing was performed separately for the 12 pairs. Because we used a single classifier that was trained indiscriminately on the entire stimulus set, it is only sensitive to differences in perceptual integration that generalize across the entire set. This is important because it prevented classification accuracy for any Kanizsa–control pair from being confounded by idiosyncratic features in that pair (such as luminance or the makeup of the inducers).

Next, we averaged across subjects and used robust linear regression to predict behavioral accuracy using classifier performance. Fig. 2*C* and Fig. S5 show regression slopes and corresponding R^2 values when predicting behavioral accuracy using peak EEG classification accuracy at 264 ms across the 12 Kanizsa–control pairs. Robust linear regression guards against violations of assumptions underlying ordinary least squares, and guards against the influence of outliers (38). Fig. S5*A* shows regressions for each of the four experimental conditions (as was done in the main text for T1 in Fig. 2*C*). This analysis shows that the T1 effect of Fig. 2*C* is replicated: peak EEG classification performance is predictive for behavior in both unmasked conditions, but unsurprisingly not in the masked conditions (where both behavior and classification accuracy was at chance).

However, the unmasked short-lag condition seems to have slightly less predictive power compared with the unmasked long-lag condition (lower R^2 and higher P value for the top-right panel compared with the top-left panel). This is to be expected if the AB manipulation affects behavioral performance without affecting perceptual integration. If the lack of conscious access in the AB (short lag) indeed selectively affects behavioral performance but not perceptual integration itself, one would expect better predictive power when using these data to predict behavioral performance that was not affected the AB, such as behavior at T1. The ability of peak classification accuracy in the four experimental conditions to predict T1 behavior across the 12 pairs is shown in Fig. S5*B*.

Indeed, short-lag T2 EEG classification is better at predicting T1 behavior than it is at predicting short-lag T2 behavior (compare the top right panel of Fig. S5*B* to the top right panel of Fig. S5*A*). There, predictive performance is very similar for the unmasked long- and short-lag conditions, as would be expected if the neural processes involved in perceptual integration are not impacted by the AB. Together, these data show independent conformation that peak classification accuracy at 264 ms is able to predict behavioral accuracy across the Kanizsa–control pairs, confirming its validity as a neural index of perceptual integration.

The Contribution of Frontal Cortex to Perceptual Integration. To investigate the contribution of frontal regions to perceptual integration, we also applied the classification analysis and the brain–behavior regression from Fig. 2*C* to the frontal electrodes: Fp1, AF7, AF3, Fpz, Fp2, AF8, AF4, AFz, and Fz (bottom right panel of Fig. S6*A*; black dots show the electrode selection in the topographic map). This analysis shows some modulation of classification accuracy in frontal cortex, both in the 264- and in the 406-ms time frame—albeit much lower than what is observed in occipital cortex (Fig. S6*A*). Importantly, however, although classification accuracy seems to show a difference between the AB and no-AB condition, none of these modulations predicts behavioral accuracy across the Kanizsa–control pairs, neither in the T1 condition nor in any of the other conditions (Fig. S6*B*; compare with Fig. 2*C* in the main manuscript and Fig. S5*A*).

This shows that the frontal signal is not causally involved in the strength of perceptual integration, consistent with the distribution of the perceptual integration signal that we observed in Fig. 2 *B* and *C*. This is in line with our finding that a selection of occipital electrodes is advantageous when using independent training runs to obtain a "pure" measure of perceptual integration (Fig. S4). Because the frontal signal is nonselective with respect to the strength of perceptual integration, it is likely to reflect a generic presence/absence signal as a precursor to global ignition and conscious access later on (which is known to occur in the range of the P300).

Note that we also performed an analysis on all electrodes, while training on T1 (Fig. 5 in the main manuscript). This analysis was intended to look at the contribution of other signals and mechanisms to behavior than perceptual integration alone. There, we further showed that no notable classification advantage was gained at 264 ms over what was observed in occipital cortex when adding frontal electrodes. The pattern of results was largely the same as what was observed when restricting the analysis to electrodes in occipital cortex (compare Fig. 2 to Fig. 5 in the main manuscript). This shows that the information contained in frontal cortex in the 264-ms time frame does not meaningfully contribute to classification accuracy over and above what is already present in occipital cortex. Later in time, however, we do see a contribution of centroparietal and frontal electrodes to classification accuracy at 406 ms, on par with the outcome of the behavioral decision, the distribution of which can be observed in Fig. 5*B* (*Bottom*).

Together, these analyses show that the occipital cortex contains a signal that uniquely reflects perceptual integration, and that this signal is not modulated by the presence or absence of conscious access. Frontal cortex does contain a weak signal in the 264-ms range that seems sensitive to whether a perceptually integrated signal will be reported, but this signal is not diagnostic or selective for the strength of perceptual integration, and does not provide a classification advantage over the signals that are already present in occipital cortex.

Seen–Unseen Analysis. A common analysis approach in consciousness research has been to perform a post hoc selection of neural data based on whether trials are behaviorally seen or unseen. Although such an analysis can in principle be useful in addition to a main analysis, it also has intrinsic pitfalls. Importantly, one cannot dissociate between the possibility that any observed effect of seen or unseen trials is a cause, a consequence, or a correlate of consciousness (also see ref. 39). Even when equating objective performance between seen and unseen conditions (40), such an approach can never determine with certainty whether the equated objective performance between seen and unseen might not be caused by uneven mixes of low-level stimulus or other bottom-up–related effects on the one hand and cognitive factors (i.e., attention) on the other (e.g., see discussion in ref. 41). Therefore, before presenting this seen–unseen analysis, we make the disclaimer that the only way of establishing cause and effect is by manipulating an independent variable (e.g., through masking or the AB) and determining the effect of that manipulation on behavior and neural processing across all trials, as is done in the main text. Again, any analysis in which a post hoc selection of neural data are made based on subject responses cannot establish with certainty whether the observed effects are caused by the manipulation in question, or are merely a consequence of coincidental differences in initial stimulus strength, noise levels in the neural machinery (e.g., waxing and waning of attention), criterion setting, incidental response errors, or any combination thereof.

This becomes apparent when inspecting the T1 plots (top row) of Fig. S7*A*. Here, we selected T1 trials based on whether a Kanizsa was seen or not, and looked at classification accuracy over time for these trials (classification accuracy was computed using the same classifier as was used in the core analysis of Fig. 2 in the main text; *SI Methods*). Although classification accuracy is clearly modulated by visibility of T1, it is impossible to know what caused this modulation. The “unseen” stimuli may have escaped report because the subject had his eyes closed on some of these trials, was momentarily not attending, because the stimulus had less bottom-up strength than its “seen” counterparts, because subjects had a conservative response criterion, because the wrong button was accidentally pressed, or any combination of these. It is evidently questionable what one can conclude about the effect of access consciousness on perceptual integration based on such a seen–unseen analysis of T1 trials, because access consciousness was not manipulated here. Importantly, however, this is not only because we did not explicitly manipulate consciousness for T1 (although this makes the flaw in the approach more apparent), but rather because one cannot attribute cause and effect using an approach in which a dependent measure (the seen–unseen response) is used to generate experimental conditions. As a general reminder: an experimental condition should always be one that is under the control of the experimenter, not under control of the subject.

The same shortcoming applies in any post hoc seen–unseen analysis approach of neural data, even when the experiment does contain an explicit manipulation of consciousness such as masking or the AB, and even when controlling for objective performance. Indeed, as we can see in the four experimental conditions (row 2 and 3) of Fig. S7*A*, there is a clear effect of seen–unseen on short-lag (AB) trials. Unseen short-lag trials (fourth column, second row) have a lower classification accuracy than seen short-lag trials (second column, second row). However, as for T1, one cannot attribute this seen–unseen modulation to differences in conscious access, as differences in bottom-up stimulus strength, attention, as well as response errors have a big influence on whether a trial is classified as seen or unseen. The seen and unseen conditions will not be balanced with respect to these coincidental properties and can thus not be sensibly compared. Therefore, if any, the only somewhat legitimate comparison in terms of the effect of conscious access on perceptual integration would be between short and long lag within the seen category (so between the first and second column) on the one hand, or between short and long lag within the unseen category (so between the third and fourth column) on the other.

These within category comparisons clearly show that (in)visibility is not modulated by lag (i.e., classification accuracy is equally strong for short- and long-lag seen trials, as well as for short and long-lag unseen trials). If anything, perceptual integration is stronger for the short-lag trials than for the long-lag trials, both within the seen and within the unseen category, although this is hard to ascertain due to the fact that different stimulus counts go into these categories as a result of post hoc selection. Also consistent with the conclusion of the main text, we see that unseen short-lag trials show a clear signature of perceptual integration, further supporting the main conclusion of this study that perceptual integration can occur in the absence of conscious access. In short, despite disclaimers about a seen–unseen analysis approach, the seen–unseen data are consistent with the results in the main text: perceptual integration is not modulated by conscious access.

Interestingly—and again in line with the main text—the same does not hold for the control experiment in which weak masking was applied. If we do the same seen–unseen analysis for this control experiment, as shown in Fig. S7*B*, a different picture emerges. Here, we see a clear effect of masking on perceptual integration within the seen category, in contrast with the AB effect of Fig. S7*A*. Weakly masked seen trials result in evidently lower peak classification accuracy than unmasked seen trials, supporting the notion that masking impacts perceptual integration directly. A similar comparison could not be made in the unseen category, because not enough trials went undetected in the unmasked condition. However, it is noteworthy that weakly masked unseen trials could not be classified above chance, again in line with the conclusion from the main text that masking impacts visibility by disrupting perceptual integration directly, although, once again, it is important to realize that many other factors could have contributed to classification performance in this “weakly masked” unseen category (erroneous button presses, lapses of attention, etc.).

T1-Based Classification at 264 ms. When training the classifier on T1 data using all electrodes, and testing this classifier on the T2 data, the 264-ms time point showed a strong main effect of masking ($F_{1,10} = 91.63$, $P < 10^{-5}$), a main effect of AB ($F_{1,10} = 8.22$, $P = 0.017$), and a trending interaction between masking and AB ($F_{1,10} = 4.06$, $P = 0.071$) (Fig. 5*B*, *Top*). To test directly whether the measurement source (neural or behavioral) at 264 ms results in a differential effect on classification accuracy, we again entered the normalized measurements into a large 2 × 2 × 2 ANOVA with factors measure (behavioral/neural), AB (yes/no), and masking (yes/no) (*SI Methods*). There was no interaction between measure and masking ($F_{1,10} = 0.274$, $P = 0.61$), but importantly there was an interaction between measure and AB ($F_{1,10} = 6.75$, $P = 0.027$), as well as a trending three-way interaction ($F_{1,10} = 4.50$, $P = 0.060$). The impact of measure confirms that, even when decision, selection, and response mechanisms are allowed contribute to classifier performance, the neural data at 264 ms cannot explain the pattern of results that is observed in behavior.

SI Discussion

Mechanisms of Masking and the AB. In line with many previous studies, the current manuscript shows that different neural mechanisms are involved in masking and the AB. In this section, we provide a short description of what these mechanisms may be. Before we begin, it is important to note that there are different types of masking, such as forward masking, backward masking, metacontrast masking, pattern masking, and object substitution masking (42), and that these are likely to have different neural substrates. Here, we follow the experimental manipulation of this manuscript and focus only on backward pattern masking, in which a randomly structured pattern follows the target stimulus, but this may not hold for other types of masking. In addition, we make the disclaimer that not all AB tasks are necessarily the same, and different AB tasks may be supported by different mechanisms (43).

Consistent with the data shown in the current manuscript, however, many studies suggest that backward pattern masking leaves feedforward processing largely intact, while selectively interfering with recurrent processing in visual cortex (13, 14). In this theoretical framework, a trailing mask is not able to catch up with the initial feedforward volley that is initiated by the target stimulus, but disrupts recurrent signals coming back from regions further up in the visual hierarchy. The existence of abundant reciprocal pathways from regions higher in the cortex to lower-tier regions is well documented (e.g., ref. 44). Local feedback pathways within visual cortex have been have been ascribed a variety of (plausibly related) functional roles, such as predictive coding and perceptual hypothesis testing, figure–ground segregation, perception of visual detail, perceptual integration, object-based attention, binding, perceptual grouping, as well as (phenomenal) consciousness. Although the degree to which these functions overlap has not been exhaustively determined, overwhelming evidence now suggests that backward masking has a detrimental effect on functions that rely on local recurrent processing within visual cortex (12–14).

The AB, on the other hand, is thought to leave both feedforward processing and local recurrent processing within visual cortex largely unaffected. Indeed, early studies showed that word targets that go undetected in an AB paradigm are affected during late stages of visual processing (17). The late time frame of the AB was later confirmed in a study by Sergent et al. (18), in which the effects of the AB are attributed to long-range interactions between frontal cortex and a distributed network of cortical association areas (4, 45). These global long-range recurrent interactions have been shown to carry an all-or-none character that is insensitive to gradual perceptual changes such as those that arise from recurrent processing within visual cortex (46).

When adding our own results to these, a picture emerges in which masking disrupts local recurrent interactions that reflect perceptual integration in visual cortex, whereas the AB impacts later conscious access by disrupting long-range integration (or "global ignition") across frontal cortex and the sensory areas, while potentially leaving perceptual integration intact.

Implications for Global Neuronal Workspace Theory and the Debate About the Existence of Phenomenal vs. Access Consciousness. The Global Neuronal Workspace (GNW) model, as advocated by Dehaene and others (4, 45), does not exclude the possibility of perceptual integration in occipital cortex in the absence of consciousness, as can also be seen the lower left quadrant of Dehaene's taxonomy of conscious and unconscious processing (see figure 1 in ref. 4). In the nomenclature of that taxonomy, this quadrant is "preconscious." The same taxonomy also contains a quadrant that is called "subliminal" (or one might say "truly unconscious") and that does not undergo neuronal/perceptual integration. "True" conscious perception in GNW theory is then reserved only for the lower right quadrant as a result of global ignition across frontal cortex and the association areas. Thus, although the GNW model concedes the possibility of preconscious perceptual integration (in addition to unconscious, this distinction is important), it further attributes true perception (consciousness) only to global ignition. However, there is a tension in GNW theory that is central to the argument that consciousness is uniquely associated with access (or global ignition). This tension has spawned a debate about the existence of phenomenal consciousness, a form of consciousness that is hypothesized to embody the contents of conscious experience, and that is thought to exist also for representations that are not cognitively accessed (30). The debate is partly caused by the fact that it is unclear in GNW theory what counts as true perception (consciousness), if one is to avoid the circularity of defining consciousness as global ignition without reference to some independent benchmark.

GNWs initial way out of the conundrum of consciousness has been to use behavioral report as such a benchmark (4). However, there are many examples in which "unconscious" or preconscious representations have been shown to exert clear influences on behavior, even when they are not accessed. For example, unconscious representations have been shown to influence cognitive control (47) and cause above-chance behavioral performance in blind-sight studies (24). Even more notable are results from split-brain and neglect patients. Split-brain patients report seeing nothing when words are presented to their right visual field, but when asked to draw them they are still able translate these words into complex drawings (48). Similarly, a patient with left-sided neglect could not overtly discriminate between two houses, despite the left side of only one of the two houses being on fire. However, when asked to choose one of the two "identical" houses, the patient would consistently prefer the house that was not on fire (49). Although some of these results are acknowledged by GNW theory, they are incompletely represented in the traditional GNW taxonomy put forward by Dehaene and colleagues, as this would require both the subliminal and the preconscious quadrant to penetrate into frontal cortex. Complex behaviors as those observed in split brains are preconscious by GNW logic, but would require an extended workspace involving frontal cortex that is not accessible for verbal report or internal narrative. Such representations are typically excluded from the true correlate of consciousness by definition, without providing proper argumentation to do so (e.g., also see ref. 5).

Given the well-documented ability of unconscious and preconscious information to penetrate frontal cortex and influence behavior, the pressing question becomes what separates "report" from "conscious report"? Barring the circularity of defining consciousness as global ignition, GNW's only available point of exit has been to relate conscious access uniquely to representations that are "perceptual" in nature (representations underlying the contents of experience, referred to as perceptual integration in this manuscript). Consequently, global ignition is only conscious if it is subserved by perceptual representations; hence the lower right quadrant in figure 1 of ref. 4, where perceptual integration is combined with long-range integration (global ignition). However, taking perceptual representations as the basis for evaluating consciousness in a GNW framework seems to run directly counter to the claim that perceptual representations can exist without consciousness (as GNW puts forward). Put differently, GNW theory asserts that something is conscious when it is perceptual, but only when it is accessed (or ignited), without seeming to provide further argument why or when ignition is required before perceptual representations become "truly" conscious.

As a result, GNW theory claims that perceptual integration in occipital cortex can occur outside conscious experience, while at the same time requiring perceptual integration as the core prerequisite for catapulting global ignition into the realm of conscious experience. An analogy would be to say that a plane cannot fly without a pilot, and that if it does fly without a pilot, is not really flying (but rather "preflying") because there is no pilot. Or to push it even further: it is like saying that the pilot is more crucial for flying than the wings of the plane (admittedly, a pilot does make the flight last longer and gives it purpose, but the act of flying does not require one, as we show in the main manuscript). This is noteworthy in the context of the ever-increasing literature on consciousness and the debate on phenomenal vs. access consciousness. On the one hand, there is a large literature on access consciousness that identifies the neural correlate of consciousness as global ignition in the P3 time frame (4, 18, 45). On the other hand, there is a large literature that identifies a neural correlate of consciousness earlier in time, coinciding with neural markers of perceptual integration, and which scales with subjective experience (14, 26, 29). Although the latter literature typically does not deny the existence of a late (or global) correlate related to conscious access, the former literature uniquely ascribes consciousness to conscious access only.

In the current manuscript, we show an occipital signal that reflects the contents of experience (perceptual integration, Fig. 2*C*). Importantly, this signal is able to escape conscious access yet retain a clear perceptual signature (Fig. S5*B*). Rephrasing the question, how can "real" conscious perception depend on global ignition (access) when the data uncover a signal reflecting the contents of conscious perception that is impervious to the same access signal and that is not represented frontally (Figs. S4 and S6)? Logically, the argument that the contents of experience are represented in one region, yet real experience requires global ignition to other regions (without clarifying why that ignition is required to invoke the basis for experience), seems problematic. No argument is given why experiencing stuff only happens when perceptual representations are "broadcast" to a global neuronal workspace (e.g., also see refs. 50 and 51), or how this step increases the explanatory power of the framework in relation to conscious experience.

In our view, a promising alternative to GNW theory is integrated information theory (IIT) by Giulio Tononi (52). IIT is the only theory that specifies in mathematical terms which representations give rise to consciousness and which do not. IIT provides a precise mathematical description of how conscious experience relates to the physical world, without having to rely on a particular biological implementation with elusive defining features. Interestingly, IIT would have no problem attributing consciousness to integrated information within visual cortex (= phenomenal consciousness). In the framework of IIT, phenomenal consciousness would be a label for "phi" in a subnetwork (or complex, in IIT terminology), and access consciousness would be the name for a larger encompassing network with larger phi (presumably larger, although this is an empirical question). What the current manuscript shows is that the subnetwork can potentially operate without being modulated by the larger network, giving it an ontologically independent status.

How Does Perceptual Integration Relate to Semantic Integration of Words. Early studies asked whether the meaning of words can be processed without conscious awareness of those words (e.g., see ref. 53). This was initially disputed (e.g., ref. 54), but later EEG and imaging studies firmly established that words could be processed up to a semantic level despite not being consciously perceived (17, 55). This is in line with the notion that extraction of complex information can occur unconsciously during the initial feedforward sweep of activation (13, 26). However, there have also been recent reports that the effects of unconscious words extend to more complex integration, in which the brain combines masked multiword utterances to reflect a single semantic valence (21). For ease of reference, we refer to these multiword integration processes as semantic integration.

In this section, we discuss the differences and similarities between perceptual integration and semantic integration. First and foremost, we would like to stress that semantic integration and perceptual integration have substantially different psychological outcomes. Perceptual integration manifests when small changes in the input or configuration of the stimulus can have dramatic perceptual effects, such as when the inducers of a Kanizsa are rotated or modified to induce or remove surface perception. In contrast, during semantic integration, changes in words or numbers have small perceptual effects, but big semantic effects. For example, the word combinations "bad rape" and "bad grape" do not differ much in terms of visual appearance, but differ vastly with respect to their perceived meaning. This makes perceptual and semantic integration hard to compare directly on a psychological level.

Accordingly, these integration processes are most likely subserved by different pathways and levels in the cortical hierarchy. This becomes most apparent when comparing the effects of masking on perceptual integration to the effects of masking on semantic integration (for an extensive review of effects of semantic subliminal processing, see ref. 45). Masking is known to abolish recurrent mechanisms involved in perceptual integration within visual cortex (13, 14), whereas late effects of semantic integration seem to survive masking. For example, semantic integration results in N400 modulations (21), a time frame at which perceptual integration is abolished under masking (also see the current manuscript). So what differences might exist between perceptual integration and semantic integration in neural terms?

One tentative explanation might be that the orthography of words can be extracted more or less automatically in the feedforward sweep, even resulting in activation of semantic representations, but that more extensive multiword semantic integration (unlike perceptual integration) occurs higher up in the hierarchy, where the incoming mask is not able to effectively interrupt integration. Indeed, evidence suggests that the visual system is able to unconsciously extract word meanings in the feedforward sweep (45), just like shapes and object categories can be extracted unconsciously in the feedforward sweep (13, 26). Early models of word reading have a feedforward architecture that is conceptually similar to feedforward models of the visual hierarchy, for example, compare ref. 56 to ref. 15. However, an important difference may be that perceptual integration occurs when features are dynamically bound in perception across visual cortex (15), whereas multiword semantic integration might take place in word-reading networks higher up in the hierarchy beyond the visual word form area (57), in accordance with the known anatomy of reading networks and early models of word reading (56, 57). The latter networks may be less sensitive to masking, such that perceptual integration (and hence conscious visual experience) of a masked word is abolished due to interruption within visual cortex, while still allowing for some (albeit minimal) degree of semantic integration of word pairs further up in the hierarchy.

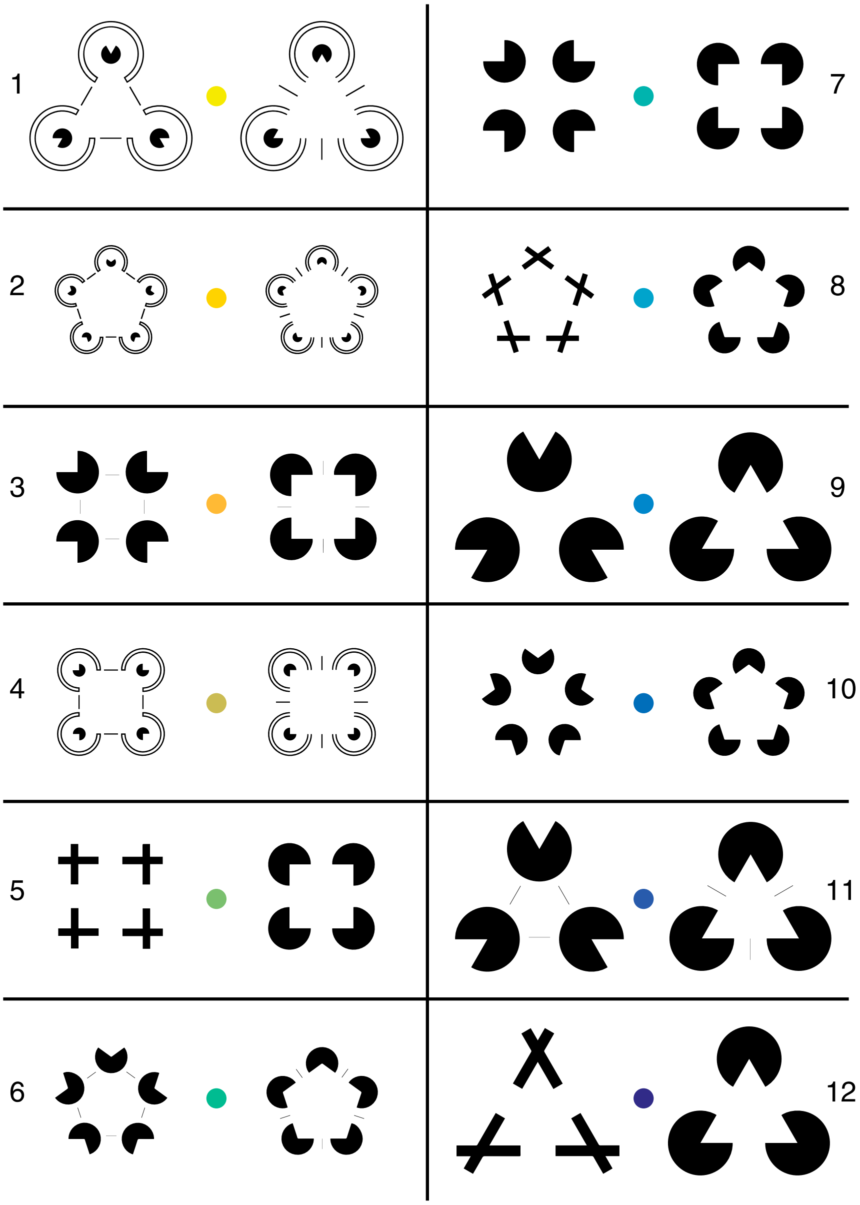


Fig. S1. The 12 Kanizsa–control pairs; see *SI Methods* for rationale behind stimulus design.

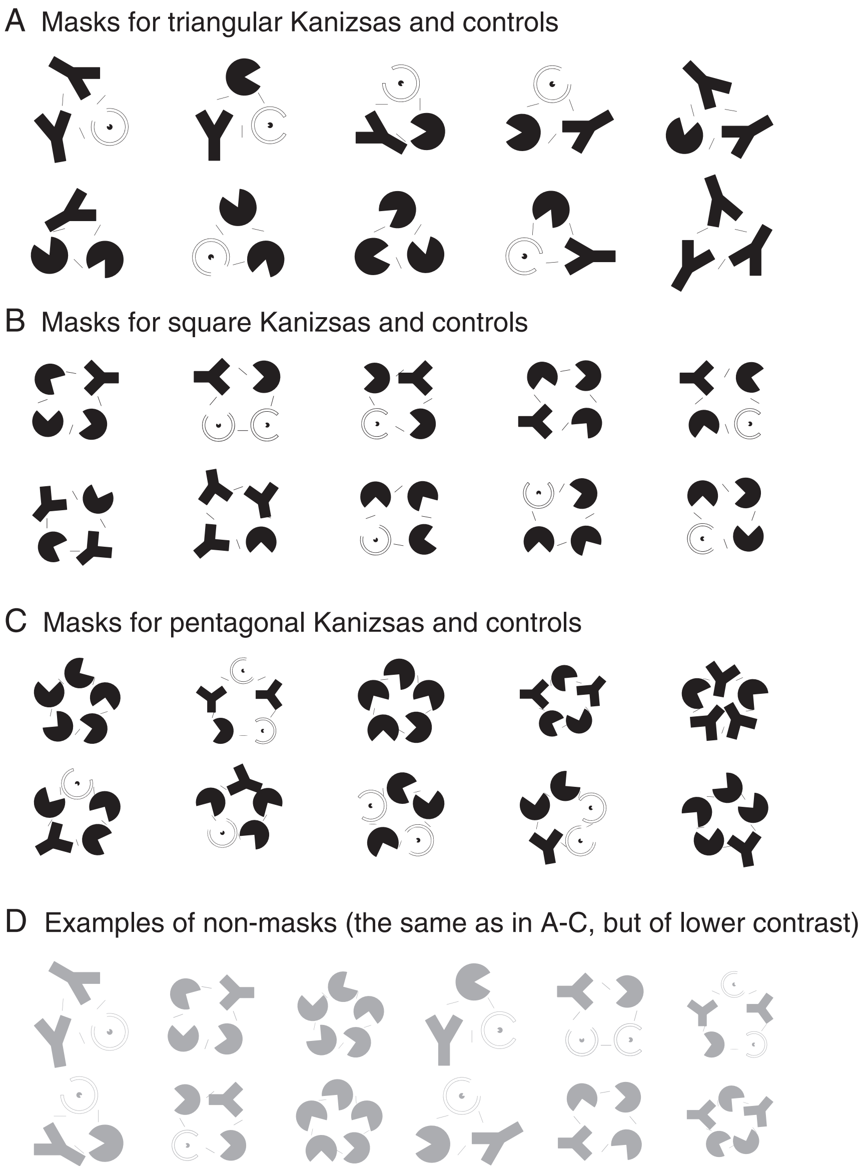

A Masks for triangular Kanizsas and controls
B Masks for square Kanizsas and controls
C Masks for pentagonal Kanizsas and controls
D Examples of non-masks (the same as in A-C, but of lower contrast)

Fig. S2. Masks used during the experimental tasks. Triangular (*A*), square (*B*), pentagonal (*C*), and examples of lower-contrast nonmasks (*D*).

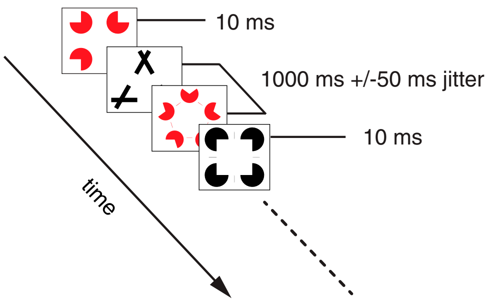


Fig. S3. Independent RSVP task that was used to train the EEG classifier. Subjects were required to press a button whenever a black target would repeat (regardless of whether this target contained a Kanizsa or not), while ignoring the red distractors. Note that this task allowed us to train the classifier using a signal that was not contaminated by response mechanisms, decision mechanisms, or task relevance. We also performed an analysis in which these mechanisms were able to contribute, by training on T1 (Fig. 5).

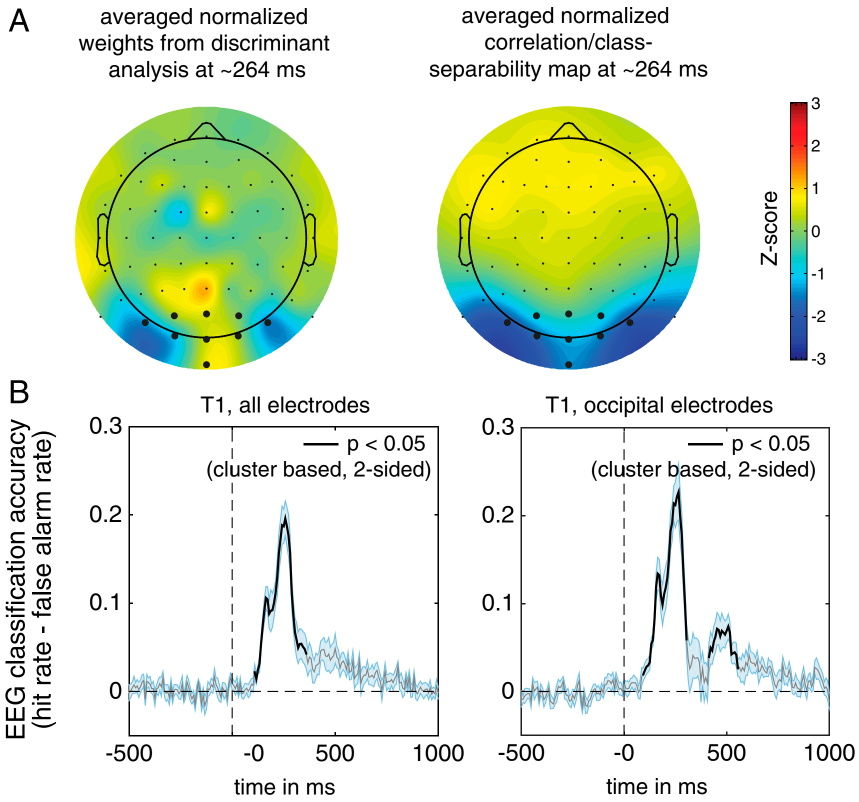
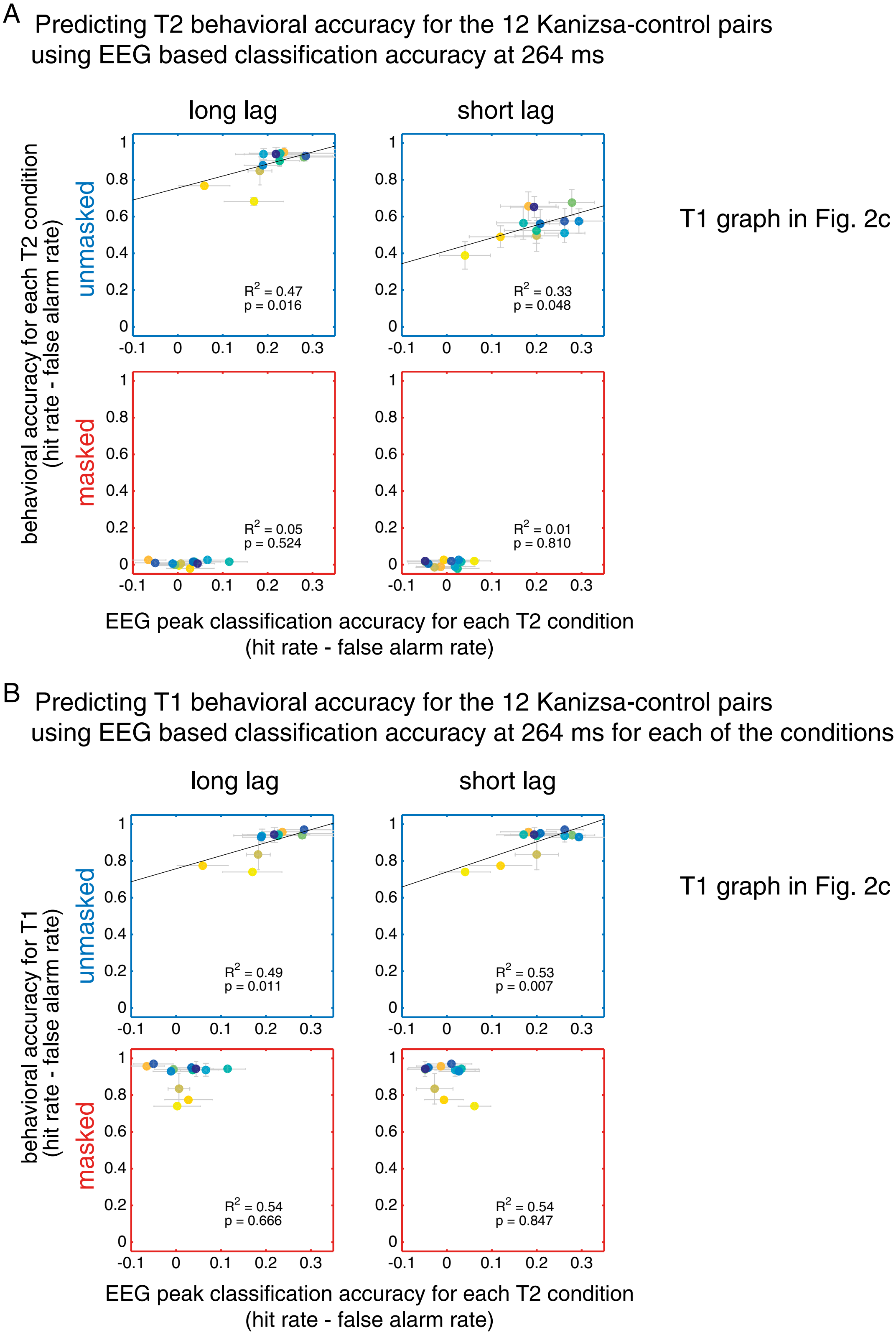


Fig. S4. Classifier weights when training on the 1-back RSVP task (*A*, *Left*) and the correlation class separability map (*A*, *Right*) at 264 ms. Line graphs are average ± SEM in light blue; thick black lines reflect $P < 0.05$, cluster-based permutation test. Because the signal is clearly occipital in nature, we compared T1 classification accuracy for all electrodes (*B*, *Left*) to classification accuracy for only the occipital electrodes (*B*, *Right*) PO7, PO3, O1, Iz, Oz, POz, PO8, PO4, O2; black dots in the topographic maps. Because the occipital electrodes result in superior performance, we used the occipital electrodes for the initial analyses (Figs. 2–4). Note, however, that using all electrodes and training on T1 (as in Fig. 5) did not substantially change the pattern of results.



A Predicting T2 behavioral accuracy for the 12 Kanizsa-control pairs using EEG based classification accuracy at 264 ms

T1 graph in Fig. 2c

B Predicting T1 behavioral accuracy for the 12 Kanizsa-control pairs using EEG based classification accuracy at 264 ms for each of the conditions

T1 graph in Fig. 2c

Fig. S5. Prediction of behavioral accuracy based on classifier performance in each of the four experimental conditions. (*A*) Behavioral accuracy within conditions based on classifier accuracy within those conditions. In both unmasked conditions, classification accuracy nicely predicts behavioral performance across the 12 Kanizsa–control pairs, albeit weaker in the short-lag AB condition. This is not surprising, given that access mechanisms are likely to dilute behavioral performance. (*B*) When using classifier performance to predict the uncontaminated T1 behavior, performance is invariably high in the unmasked conditions.

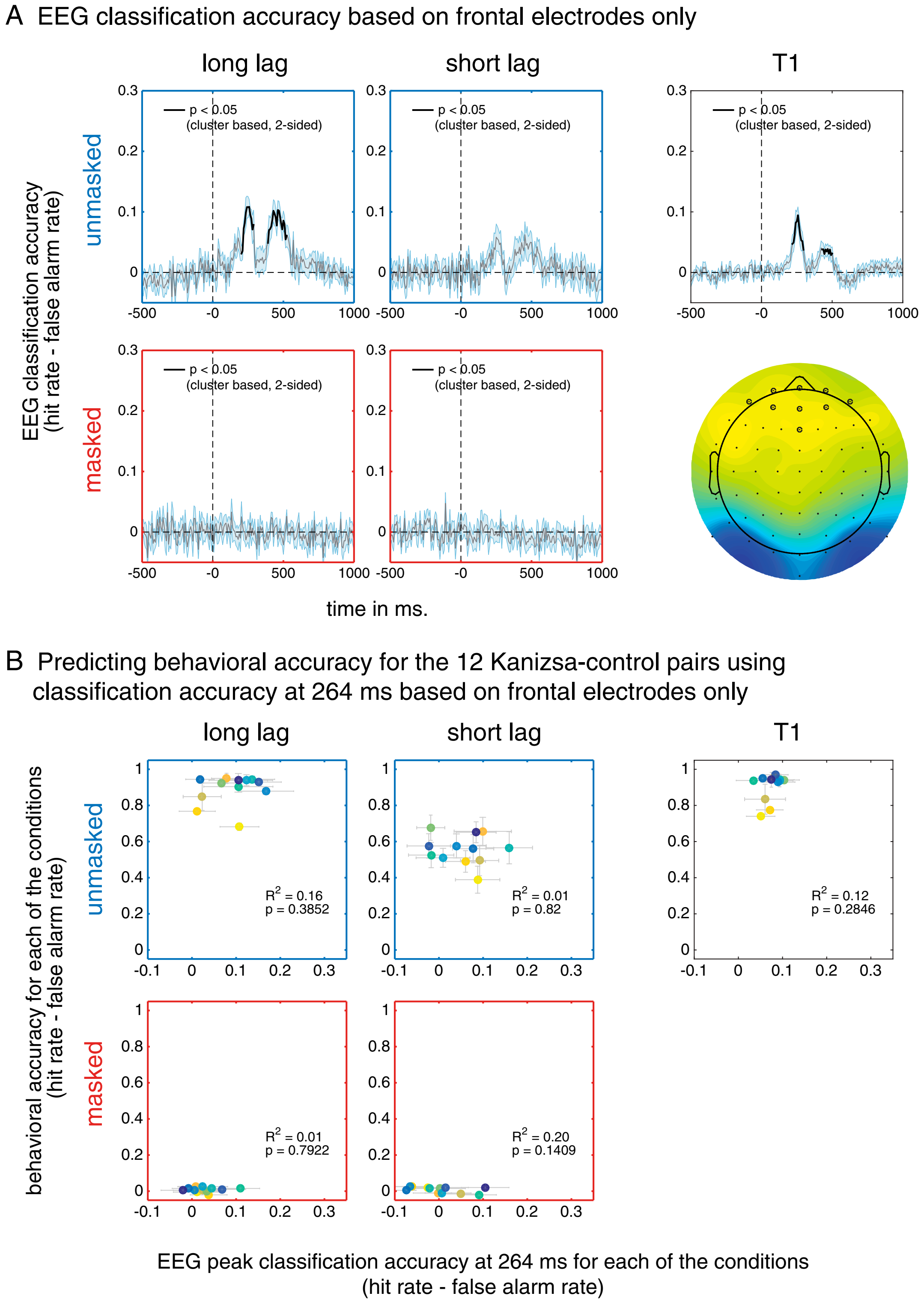


Fig. S6. Contribution of frontal electrodes to perceptual integration. Although the signal related to perceptual integration is clearly occipital in nature (Fig. S4), a control analysis was performed to determine whether frontal electrodes contribute to this signal. (*A*) Classification accuracy for the four experimental conditions as well as T1, using only frontal electrodes: Fp1, AF7, AF3, Fpz, Fp2, AF8, AF4, AFz, and Fz. *Right Bottom* shows the topographic correlation/class separability map when using all electrodes (see *SI Methods* for details), with the frontal electrodes highlighted using black dots. (*B*) The degree to which this signal predicts behavioral performance across the 12 Kanizsa–control pairs in the four experimental conditions as well as T1. The frontal signal is invariably unable to predict behavioral performance across the 12 Kanizsa–control pairs (Fig. 2*C* and Fig. S5).

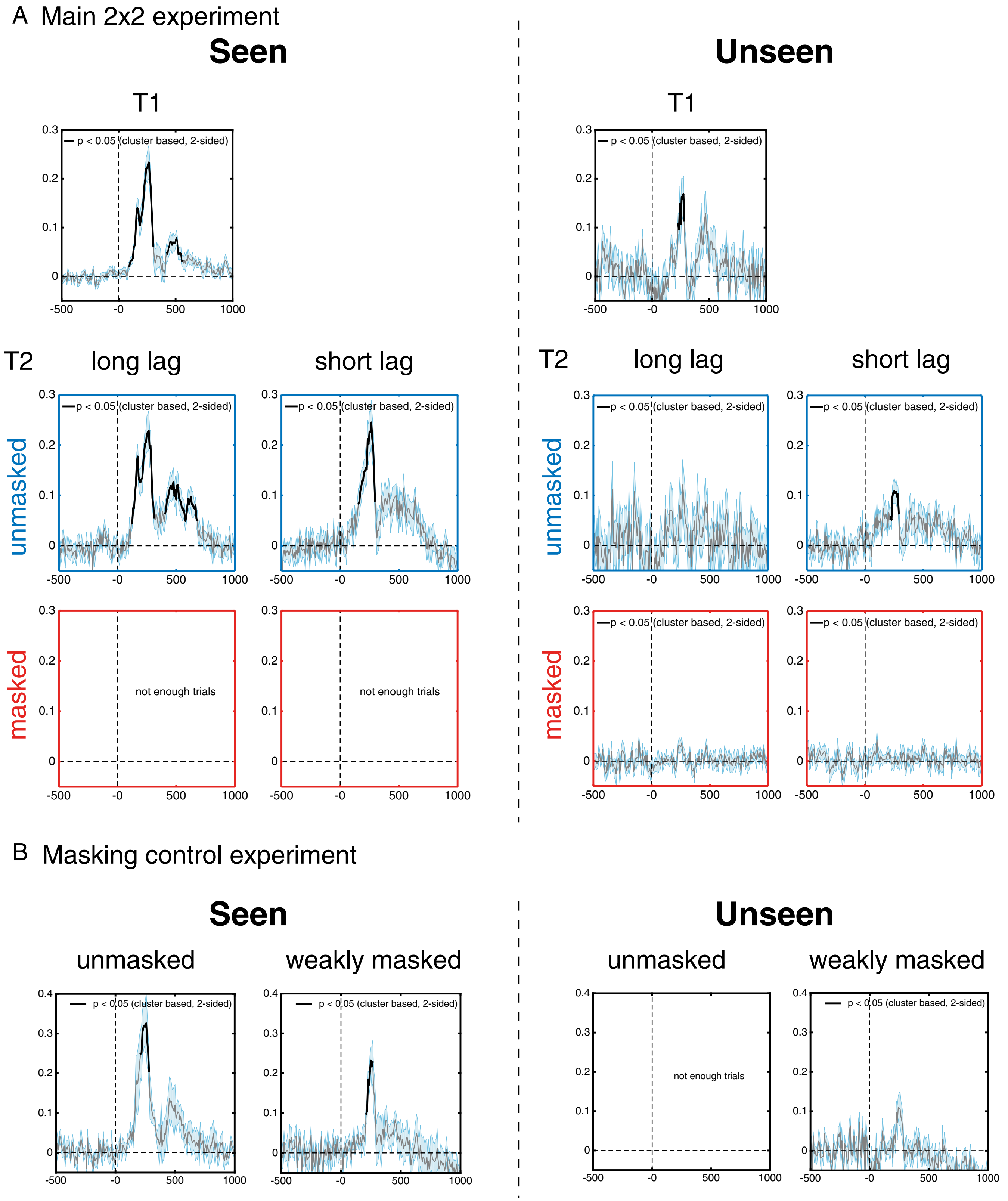


Fig. S7. Seen–unseen analysis. (*A*) Splitting the main experiment up according to behavioral decision. (*B*) Splitting the masking control experiment up according to behavioral decision. Please read *SI Results*, *Seen–Unseen Analysis*, for an explanation of the pitfalls associated with behavior contingent selection of neural data and proper interpretation. Results are consistent with the main text.

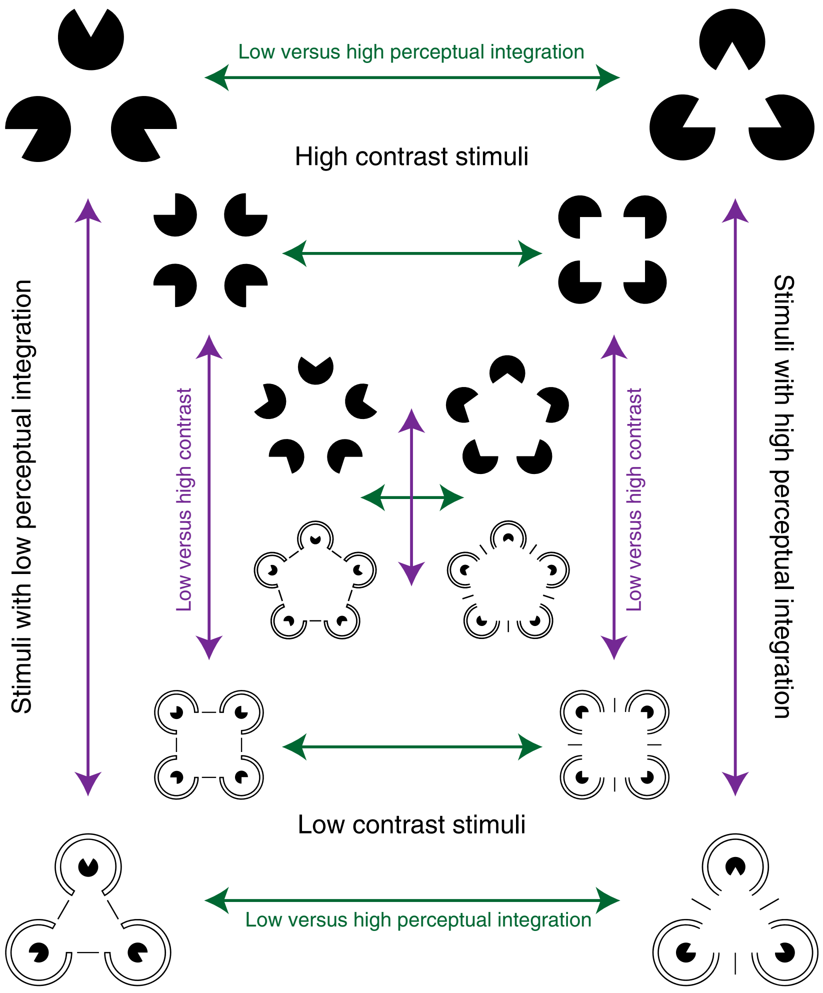

Low versus high perceptual integration
High contrast stimuli
Stimuli with low perceptual integration
Low versus high contrast
Low versus high contrast
Stimuli with high perceptual integration
Low contrast stimuli
Low versus high perceptual integration

Fig. S8. Contrast detection vs. perceptual integration. Stimuli used in the masking control analysis belonging to Fig. 3. Stimulus design was such that one could compare either in the contrast dimension or in the perceptual integration dimension, while collapsing orthogonally over the other dimension.

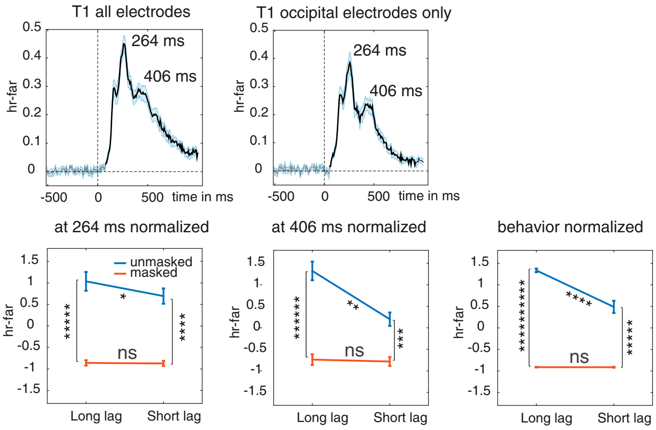


Fig. S9. Classification accuracy for all electrodes and occipital electrodes when training and testing on T1 (eightfold leave-one-out procedure). Line graphs are average ± SEM in light blue; thick black lines reflect $P < 0.05$, cluster-based permutation test. Given the contribution of response and decision mechanisms to the response, we now see a slight enhancement when using all electrodes compared with when using occipital electrodes only (Fig. S4). Bottom panels shows graphs for the normalized responses when training on T1 at 264 and 406 ms, and normalized responses obtained from behavior. ns, not significant ($P > 0.05$). $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, $****P < 10^{-4}$, $*****P < 10^{-5}$, $******P < 10^{-6}$, $**********P < 10^{-12}$.