Psychon Bull Rev (2016) 23:432–438
DOI 10.3758/s13423-015-0916-2

BRIEF REPORT

The visual system supports online translation invariance for object identification

Jeffrey S. Bowers[1] · **Ivan I. Vankov**[1] · **Casimir J. H. Ludwig**[1]

Published online: 25 August 2015

Abstract The ability to recognize the same image projected to different retinal locations is critical for visual object recognition in natural contexts. According to many theories, the translation invariance for objects extends only to trained retinal locations, so that a familiar object projected to a nontrained location should not be identified. In another approach, invariance is achieved "online," such that learning to identify an object in one location immediately affords generalization to other locations. We trained participants to name novel objects at one retinal location using eyetracking technology and then tested their ability to name the same images presented at novel retinal locations. Across three experiments, we found robust generalization. These findings provide a strong constraint for theories of vision.

Keywords Translation invariance · Translation tolerance · Object identification · Vision · Human visual perception and categorization · Object recognition · Perceptual categorization

Retinal images vary when an object is seen under different viewing conditions, including changes in orientation, viewing distance, illumination, and position in the visual field. Somehow the visual system must ensure that object recognition is invariant across these changes. Here we focus on how the visual system copes with translation across retinal positions. Can we recognize an object that we have only ever experienced in one part of the visual field in a novel retinal location? For object recognition to be tolerant to changes in retinal position, do we need to have experienced the object in all possible retinal locations?

There are three general views regarding how we achieve translation tolerance (see Fig. 1). According to Hypothesis 1, tolerance is largely postvisual. That is, our ability to identify familiar objects across a wide range of eccentricities is achieved by learning multiple high-level object representations for the same object at different retinal locations, and then linking these representations to a common postvisual code, as in Fig. 1A (e.g., Afraz & Cavanagh, 2008; Dehaene, Cohen, Sigman, & Vinckier, 2005; Kravitz, Kriegeskorte, & Baker, 2010; Ullman, 2007). According to Hypothesis 2, robust translation tolerance occurs within the visual system. That is, a given object has a common high-level visual representation, but contacting this representation requires training of the mapping between an object in different retinal locations and its higher-level visual representation, as is depicted in Fig. 1B (e.g., Cox & DiCarlo, 2008; Dandurand, Hannagan, & Grainger, 2013; Di Bono & Zorzi, 2013). A critical prediction of both hypotheses is that an object cannot be identified when it is projected to a novel retinal location that is distal from trained locations. Finally, according to Hypothesis 3, translation tolerance occurs within the visual system and is computed online. That is, the visual system maps a given object projected to different retinal locations to a common (single) high-level visual representation, regardless of its retinal location. An object can thus be identified at a novel location, even when this new location is quite distal from the locations at which the object was trained (e.g., Biederman, 1987).

The claim that translation tolerance is computed online is often described as the "standard view" (e.g., Kravitz et al., 2010). Early research appeared to support this view. For example, Biederman and Cooper (1991) showed that long-term priming for familiar objects is equally robust following a

✉ Jeffrey S. Bowers
j.bowers@bristol.ac.uk

[1] School of Experimental Psychology, University of Bristol, 12a Priory Road, Bristol BS8-1TU, UK

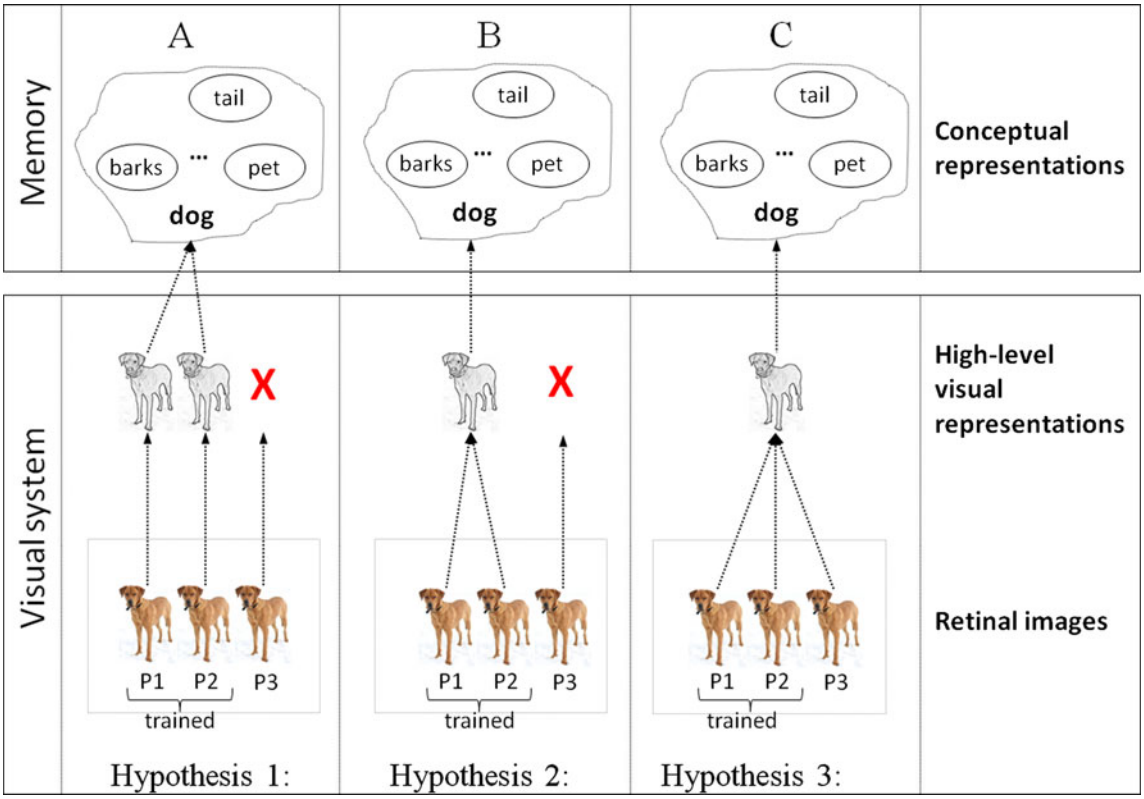

Fig. 1 Illustration of three hypotheses regarding translation invariance. In panel A, translation invariance is largely postvisual. According to this view, the visual system learns many high-level object representations for a given image, with different representations at different retinal locations. Translation invariance is achieved by mapping these distinct visual representations to a common conceptual code (or alternatively, name code). In panel B, translation invariance occurs within the visual system, but only at trained retinal locations. That is, the visual system learns to map a given image projected to different retinal locations onto a common (single) high-level visual perceptual code, but these mappings require training. In Hypothesis 3, translation invariance occurs within the visual system and is computed "online." That is, the visual system maps a given image projected to different retinal locations to a common (single) high-level perceptual code, even when the image is projected to a novel retinal location

study-to-test change in retinal location. Similarly, in masked-priming studies with words, robust and equal priming has been observed when the prime and targets are presented at the same or at different retinal locations (Bowers & Turner, 2005). An alternative account of these findings, however, is that the priming reflected postvisual processes, such as common name codes (Kravitz et al., 2010). Indeed, a number of studies designed to minimize the role of postvisual processing have failed to observe robust (or any) translation tolerance in priming and perceptual-learning tasks (e.g., Dill & Fahle, 1997; Kravitz et al., 2010; McAuliffe & Knowlton, 2000; Newell, Sheppard, Edelman, & Shapiro, 2005; for a review, see Kravitz, Vinson, & Baker, 2008). A common conclusion from these and related studies is to reject the online account of tolerance and instead adopt Hypothesis 1 or 2.

However, a number of design features of the latter studies make it difficult to draw any strong conclusions. First, the objects are typically flashed briefly at a given retinal location at study and/or test (e.g., Dill & Fahle, 1997; Kravitz et al., 2010; McAuliffe & Knowlton, 2000), and it is possible that tolerance requires more extended sampling of the object. Indeed, other invariances, including left–right reflection and picture-plane rotation invariance, are only manifested in priming tasks when participants are able to attend to the objects for longer periods of time (e.g., Thoma, Davidoff, & Hummel, 2007). Second, many of these studies used images that were highly unlike real objects (e.g., Dill & Fahle, 1997), or that were extremely similar to one another (e.g., Cox & DiCarlo, 2008). Distinguishing between highly similar patterns may rely more heavily on low-level visual representations (Ahissar & Hochstein, 2004), and low-level codes are more retinotopically constrained.

In order to provide a strong test of online translational tolerance, it is necessary to adopt conditions in which (i) more extended sampling is possible; (ii) the items are more object-like; (iii) the objects differ from one another in more than some fine perceptual detail; and (iv) postvisual codes cannot contribute to performance. To this end, we used eyetracking so that objects could be presented for a longer duration at controlled retinal locations. In addition, participants were trained on a set of novel objects that differed in configural properties rather than in fine details. The question was whether

participants could identify these objects in novel retinal locations after training in one location.

We used a naming task that required unique identification of the objects. The naming task does engage postvisual processes, but these codes could not support translation tolerance in our experiments, given that the objects were novel. To see why, consider a newly learned visual object representation that is tightly bound to a specific retinal location (as in Hypotheses 1 and 2). By definition, this representation cannot be accessed when the object is projected to very different retinal locations. And if the newly learned visual representation cannot be accessed, then this object cannot be named, given that the name codes is linked to newly learned visual representation (see Fig. 1). This contrasts with previous Biederman studies (e.g., Biederman & Cooper, 1991; Fiser & Biederman, 2001) that had assessed translation tolerance for familiar objects. In this situation, postvisual codes are contacted regardless of the locations of the object at study and test (e.g., the semantics and name of the category "piano" can be accessed wherever a piano is projected on the retina, given the participants' past history with pianos), and accordingly, postvisual cues could indeed have contributed to priming in these studies, even if the visual representations were tightly bound to retinal position.

Experiment 1

Materials and method

Participants Ten participants took part in Experiment 1a, and another ten participated in Experiment 1b. All participants were paid £10 for their time.

Stimuli and equipment We took six objects from Tarr and Pinker (1990) that included similar local features but that differed in their overall configurations (see Fig. 2). We chose these objects because they cannot be identified or distinguished from one another on the basis of their parts, but need to be identified as complete objects. No object was the mirror image of another, and they were all nonsymmetrical along the vertical plane. Each object was assigned a spoken name ("Q," "V," "C," "S," "D," and "J").

Presentation of the stimuli was managed by a MATLAB (MathWorks, Inc.) program using Psychophysics Toolbox 3.08. The objects were displayed on a ViewSonic G225f 21-in. CRT monitor, running at 85 Hz with a spatial resolution of 1,024 × 768 pixels and viewed at a distance of 57 cm, at which the objects extended 5° × 5° of visual angle. In Experiment 1a, the horizontal eccentricity from the center of the object to a central fixation cross was 5.5°. Experiment 1b was a replication, with an increased separation between the study and test locations (6.5°) and also a tighter region around fixation (see below). Participants were instructed to focus on a central

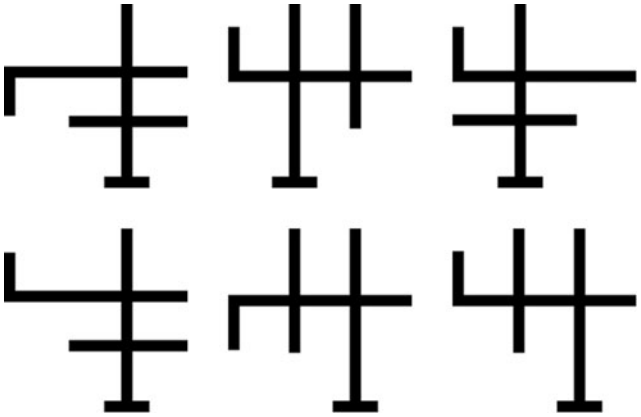

Fig. 2 The six objects used across all experiments

fixation cross while their eye movements were tracked at 1000 Hz using the EyeLink 2 k system (SR Research, Ltd.). Each time the system detected that the participants' gaze had moved sufficiently far away from fixation (more than 2° in Exp. 1a or 1.5° in Exp. 1b), it replaced the object with a mask that prompted the participant to return to the fixation point. The time during which a mask was shown was not included in the total presentation time of the stimulus.

Procedure and design In Experiments 1a and 1b, the participants focused on a centrally located fixation cross while an object was displayed in either the left or the right part of the screen. The experiment started with 30 familiarization trials, during which a random sequence of the novel objects shown in Fig. 2 was presented, along with their spoken names ("Q," "V," "C," "S," "D," and "J"), with the object–name mappings counterbalanced across participants. Participants were instructed to learn the object names. Then they completed a training phase. On each training trial, an object was displayed for 2 s and the participant attempted to retrieve its name. Written feedback was then provided, along with a repetition of the object and the object name.

An example of a training trial is shown in Fig. 3. The training lasted until the participant had managed to correctly name 24 novel objects in a row. The average numbers of trials needed to complete the training sessions were 112 in Experiment 1a and 92 in Experiment 1b. One person did not manage to complete the training in 150 trials and was replaced by another participant. During familiarization and training, all of the objects were presented either at the left or the right side of fixation, with locations counterbalanced across participants.

Participants who successfully completed the training phase then completed the test phase, which included 54 test trials in which the position of the objects was manipulated. For 18 of the test trials, the object was presented at the same position as in training (*same* condition); for another 18 trials, the objects was presented in the center of the screen (*center*); and for the remaining 18 trials, the object was presented at the opposite

Fig. 3 Schematic diagram of one training trial in Experiment 1. Participants fixated a fixation cross for 1 s, and then a single object was displayed for 2 s, at a horizontal eccentricity of 5.5° in Experiment 1a and 6.5° in Experiment 1b. Participants attempted to retrieve the object name and then received feedback, in the form of a written response followed by the display of the object and its name presented auditorily

side of the screen (*opposite*). There was no feedback in the testing phase, and the order of the test trials was randomized.

Results

As can be seen in Fig. 4, the participants were able to reliably name the trained objects when they were presented at novel positions, either in the center or at the opposite side of the screen. For example, in Experiment 1a, the average accuracy rates were 92 % in the *same*, 81 % in the *center*, and 71 % in the *opposite* condition. These results are inconsistent with the critical predictions of both Hypotheses 1 and 2, according to which performance should have been at chance (16.67 %).

In all of the conditions, the average accuracy was far above the chance level (Cohen's *d*s: same, 8.26; center, 5.06; opposite, 3.69). The effect sizes of the differences between experimental conditions were lower: same–center, $d = 1.06$; same–opposite, $d = 1.78$; center–opposite, $d = 0.73$. A similar pattern of results was found in Experiment 1b, with performance far above chance in all conditions: same, $d = 6.54$; center, $d = 4.28$; opposite, $d = 2.27$. Given the increased eccentricity of the objects and the stricter fixation conditions, this serves to highlight the robustness of the translation effects. The effect sizes of the differences between conditions were lower than in Experiment 1a: same–center, $d = 0.46$; same–opposite, $d = 1.01$; center–opposite, $d = 0.62$.

In these two experiments, both the retinal and spatial locations (in screen coordinates) of the objects differed between training and test. In Experiment 2, we assessed to what extent the reduced performance was due to changes in the spatial location. Objects were always presented in the center of the screen, but the location of the fixation point varied between study and test. In this way, the retinal location varied, but the spatial location was held constant.

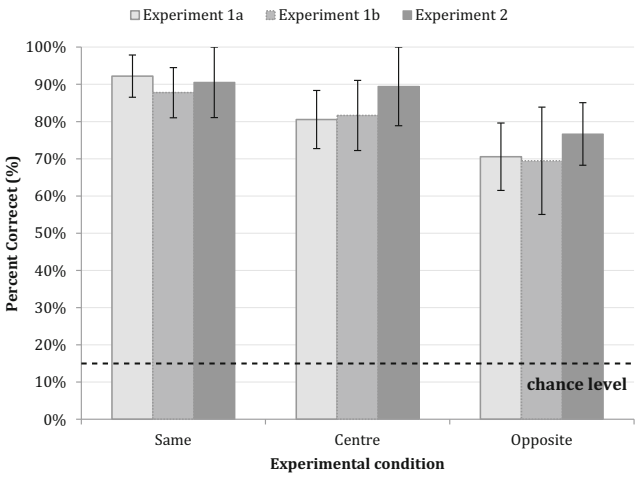

Fig. 4 Percent correct object naming in Experiments 1 and 2, as a function of the study–test display conditions. Chance level is 16.7 %. The error bars represent confidence intervals

Experiment 2

Materials and method

Participants Ten participants took part in each of the experiments and were paid £10 for their time.

Stimuli and equipment The same equipment and stimuli were used as above. The horizontal eccentricity of the fixation point was 6.5°, and each time that a participant's gaze was more than 1.5° away from fixation, the object was replaced with a mask that prompted the participant to return to the fixation point.

Procedure and design The experimental procedure was similar to the one described above. The only difference was that, during training, the novel objects were centered at the middle of the screen and the fixation point was located either on the left or the right side of the screen, with locations counterbalanced across participants. During testing, the fixation point was displayed either at the training position (*same*), at the center of the screen (*center*), or at the opposite side of the screen (*opposite*) in a randomized order. The average number of trials to complete the training session was 114.

Results

The results of Experiment 2 are also shown in Fig. 4. The pattern was similar to the one we found in Experiments 1a and 1b. Accuracy was above chance in all of the experimental conditions: same, 91 %, Cohen's $d = 4.86$; center, 89 %, $d = 4.29$; opposite, 77 %, $d = 4.45$. Performance was better in the *same* than in the *opposite* condition, $d = 0.96$, but not than in the *center* condition, $d = 0.07$. We also observed a significant reduction between the *center* and *opposite* conditions, $d = 0.83$.

In the final experiment, we attempted to eliminate the effect of retinal location by presenting four of the six objects in multiple retinal locations during study, with the remaining two objects being projected to only one location. Previous work on perceptual learning has suggested that tolerance may be improved when a number of similarly complex objects are experienced at multiple retinal locations (e.g., Xiao et al., 2008).

Experiment 3

Materials and method

Participants Ten participants took part and were paid £10 for their time.

Stimuli and equipment The same equipment and stimuli were used as above, but in this case, three novel objects were displayed side by side on the screen, with one to the left, one at the middle, and one to the right. The fixation cross was presented in the middle position. The horizontal eccentricity of the objects presented in the periphery was 5.5°. Each time that the participants' gaze was more than 2° away from fixation, the objects were replaced with a mask that prompted the participant to return to the fixation point.

Procedure and design The experiment started with a familiarization phase in which triplets of novel objects were displayed and their names were presented auditorily, one by one, from left to right. The two critical objects were always presented at the same retinal location—one to the left of fixation, and one to the right. The critical objects varied across participants, but the same names were used in all cases. The remaining four objects were presented at all positions. The familiarization phase was followed by a training phase in which the participants saw triplets of objects and named them from left to right. Feedback was provided after each trial. A response was regarded as incorrect if any of the three objects was named incorrectly. Training was completed when participants had managed to respond correctly on ten trials in a row.

The experiment ended with a test phase in which all of the objects were presented at all positions. Thus, four experimental conditions were formed: *same* (the object was trained in the periphery only and was tested at the trained position), *center* (the object was trained in the periphery and was tested in the center position), *opposite* (the object was trained in the periphery and was tested in the opposite peripheral position), and *control* (the object was trained at all positions). There was no feedback at test.

Results

The results of Experiment 3 are displayed in Fig. 5. In all conditions, accuracy was far above the chance level: same, 95 %, Cohen's $d = 11.43$; center, 80 %, $d = 3.54$; opposite, 80 %, $d = 2.85$; control, 92 %, $d = 9.45$. However performance was reduced in the *center* relative to the *same* condition, $d = 1.08$, as well as in the *opposite* relative to the *same* condition, $d = 0.68$. There was no difference between the *center* and *opposite* conditions, $d = 0.00$.

Once again, performance was reduced when objects were presented at different retinal locations during training and test. This highlights a retinotopic contribution of object learning that is difficult to eliminate and was observed even when control novel objects were trained at multiple locations.

General discussion

The results are clear-cut: After participants had learned to name novel objects at one retinal location, they were able to identify and name the same objects at other retinal locations with a high degree of accuracy. These results rule out all

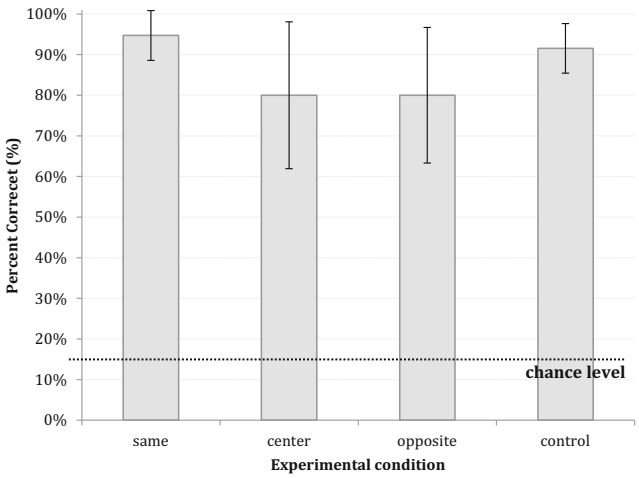

Fig. 5 Percent correct object naming in Experiment 3 for the critical and control objects, as a function of study–test conditions. The critical objects were studied at one location, and the control objects were studied at all locations. Chance level is 16.7 %. The error bars represent confidence intervals

theories that assume that there is little translation tolerance for objects within the visual system (Hypothesis 1) or that robust tolerance within the visual system requires training a given object at a given retinal location (Hypothesis 2), as we outlined in Fig. 1. Instead, the results lend some support to "online" theories of tolerance, in which high-level object codes are represented independently of retinal location, such that generalization is possible following an encounter with an object at a single location (Hypothesis 3).

What should be made of the drop in performance following a study-to-test retinal change? Does this lend some support to theories according to which visual objects codes are tightly bound to retinal location (Hypothesis 1) or to theories that claim that tolerance needs to be explicitly trained (Hypothesis 2)? Not at all. In fact, our findings are in striking contrast with the past work that has been used to support these hypotheses. For instance, Cox and DiCarlo (2008) argued for Hypothesis 2, on the basis of the finding that a rhesus monkey was at chance at identifying a novel object following a translation from +2° to –2° from fixation (and the fact that the responses of inferotemporal neurons showed the same selectivity in two monkeys). By contrast, in our Experiment 3, participants were at ~80 % accurate in naming novel objects following a shift of 13° (when chance was 16.7 %).

Furthermore, all theories of visual object identification agree that vision is mediated by a hierarchical system in which the early representations are tightly bound by retinal location (e.g., simple cells in V1). Thus, the effect of location may reflect the fact that performance was supported, in part, by low-level visual learning that boosted performance in the same-location condition. For instance, the feature ⌜ occurs in two of the novel objects (see Fig. 2), and if participants learned to map this lower-level feature to the object names, this could contribute to performance in a retinotopically constrained fashion. Note that the low-level features of our objects did not reliably predict the names of the objects—rather, the overall configurations of the features defined the objects—so it is not possible to explain the high accuracy in naming across conditions on this basis. Nevertheless, learning these low-level features might have boosted performance in the *same* condition, as we observed. Alternatively, the reduced performance across locations might have reflected a limitation of translation tolerance at the object level (with no contribution from low-level features), contrary to Hypothesis 3. According to this alternative hypothesis, translation tolerance is indeed limited, but the tolerance is much greater than has commonly been claimed (as in Hypotheses 1 and 2).

An obvious question remains: Namely, why did we obtain robust translation tolerance when most previous work has reported much more restricted tolerance? We can only speculate, but many methodological differences from the previous studies may explain the contrasting results. For example, Cox and DiCarlo (2008) trained two monkeys to discriminate briefly flashed objects that differed in a subtle visual detail over the course of 30–60 training sessions, in which each object was presented in each location over 20,000 times. This study was unlike the present experiments in that the objects were briefly flashed at study and test and differed only in subtle visual detail, and because of its greatly extended training phase. Any or all of these factors may have contributed to the different results. Indeed, DiCarlo and Maunsell (2003) suggested that the extensive training of objects at specific retinal locations may narrow the receptive fields in macaque inferotemporal cortex.

Similarly, as we noted earlier, many of the studies that reported little translation invariance with humans were also based on flashed objects at study and test and required participants to distinguish between objects that differed in visual detail. Again, these factors may have contributed to the limited translation invariance. Indeed, recent studies with humans have highlighted how training conditions can dramatically impact on the translation tolerance of low-level perceptual learning (e.g., learning to discriminate between subtle variations in contrast and orientation), with some studies showing little tolerance (e.g., Karni & Sagi, 1991) and others showing robust tolerance (e.g., Xiao et al., 2008). Accordingly, we think it is likely that the different translation results obtained with objects reflect something about the specific study and test conditions rather than about the subject populations (e.g., monkeys or humans). The most critical point, however, is that the visual system can support robust translation tolerance under some conditions that are arguably more ecologically valid.

The present findings are important because they provide a challenge for most theories and models of visual recognition. For example, many neural network models of word and object identification that claim to support translation tolerance have implemented specific versions of Hypothesis 1 or 2 (Dandurand et al., 2013; Di Bono & Zorzi, 2013)—that is, tolerance was achieved by training the models with objects (in these cases, words) in all possible spatial locations. What the authors did not test was whether their models could generalize and identify an object at an untrained location. Almost certainly the answer is "no," since no mechanisms were included that could achieve this capacity. Our findings highlight the need to develop processes that can support more robust translation tolerance (e.g., Földiák, 1991) that have been omitted in most current theories and models of object and word recognition.

Author note This research was supported by Leverhulme Grant Number RJ5538, awarded to J.S.B.

References

Afraz, S.-R., & Cavanagh, P. (2008). Retinotopy of the face aftereffect. *Vision Research, 48,* 42–54.

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences, 8,* 457–464.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94,* 115–147. doi:10.1037/0033-295X.94.2.115

Biederman, I., & Cooper, E. E. (1991). Evidence for complete translational and reflectional. *Perception, 20,* 585–593.

Bowers, J. S., & Turner, E. L. (2005). Masked priming is abstract in the left and right visual fields. *Brain and Language, 95,* 414–422.

Cox, D. D., & DiCarlo, J. J. (2008). Does learned shape selectivity in inferior temporal cortex automatically generalize across retinal position? *Journal of Neuroscience, 28,* 10045–10055.

Dandurand, F., Hannagan, T., & Grainger, J. (2013). Computational models of location-invariant orthographic processing. *Connection Science, 25,* 1–26.

Dehaene, S., Cohen, L., Sigman, M., & Vinckier, F. (2005). The neural code for written words: A proposal. *Trends in Cognitive Sciences, 9,* 335–341. doi:10.1016/j.tics.2005.05.004

Di Bono, M. G., & Zorzi, M. (2013). Deep generative learning of location-invariant visual word recognition. *Frontiers in Psychology, 4,* 635.

DiCarlo, J. J., & Maunsell, J. H. (2003). Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position. *Journal of Neurophysiology, 89,* 3264–3278.

Dill, M., & Fahle, M. (1997). The role of visual field position in pattern-discrimination learning. *Proceedings of the Royal Society B, 264,* 1031–1036.

Fiser, J., & Biederman, I. (2001). Invariance of long-term visual priming to scale, reflection, translation, and hemisphere. *Vision Research, 41,* 221–234.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation, 3*(2), 194–200.

Karni, A., & Sagi, D. (1991). Where practice makes perfect in texture discrimination: Evidence for primary visual cortex plasticity. *Proceedings of the National Academy of Sciences, 88,* 4966–4970.

Kravitz, D. J., Kriegeskorte, N., & Baker, C. I. (2010). High-level visual object representations are constrained by position. *Cerebral Cortex, 20,* 2916–2925. doi:10.1093/cercor/bhq042

Kravitz, D. J., Vinson, L. D., & Baker, C. I. (2008). How position dependent is visual object recognition? *Trends in Cognitive Sciences, 12,* 114–122.

McAuliffe, S. P., & Knowlton, B. J. (2000). Long-term retinotopic priming in object identification. *Perception & Psychophysics, 62,* 953–959.

Newell, F. N., Sheppard, D. M., Edelman, S., & Shapiro, K. L. (2005). The interaction of shape- and location-based priming in object categorisation: Evidence for a hybrid "what + where" representation stage. *Vision Research, 45,* 2065–2080. doi:10.1016/j.visres.2005.02.021

Tarr, M. J., & Pinker, S. (1990). When does human object recognition use a viewer-centered reference frame? *Psychological Science, 1,* 253–256.

Thoma, V., Davidoff, J., & Hummel, J. E. (2007). Priming of plane-rotated objects depends on attention and view familiarity. *Visual Cognition, 15,* 179–210. doi:10.1080/13506280500155627

Ullman, S. (2007). Object recognition and segmentation by a fragment-based hierarchy. *Trends in Cognitive Sciences, 11,* 58–64.

Xiao, L.-Q., Zhang, J.-Y., Wang, R., Klein, S. A., Levi, D. M., & Yu, C. (2008). Complete transfer of perceptual learning across retinal locations enabled by double training. *Current Biology, 18,* 1922–1926.