Synthese (2021) 198:4457–4491
https://doi.org/10.1007/s11229-019-02351-1

Binding and differentiation in multisensory object perception

E. J. Green[1]

Received: 29 September 2018 / Accepted: 27 July 2019 / Published online: 1 August 2019

Abstract
Cognitive scientists have long known that the modalities interact during perceptual processing. Cross-modal illusions like the ventriloquism effect show that the course of processing in one modality can alter the course of processing in another. But how do the modalities interact in the specific domain of object perception? This paper distinguishes and analyzes two kinds of multisensory interaction in object perception. First, the modalities may bind features to a single object or event. Second, the modalities may cooperate when differentiating an object or event from its surroundings. I critically evaluate evidence for various forms of multisensory binding. I then consider the case for multisensory differentiation. I argue that existing evidence for multisensory differentiation is inconclusive. I highlight ways that the issue might be empirically resolved.

Keywords Object perception · Multisensory perception · Binding · Perceptual organization · Object files

1 Introduction

Many objects and events are perceptible through more than one modality. We can both see and touch a glass, and we can both see and hear the glass shatter when it falls to the floor. But how do the modalities interact when we apprehend particular objects and events in the world? And what is the nature of the representations that result from these interactions?

This paper distinguishes and analyzes two kinds of multisensory interaction in object perception. The first kind is *multisensory binding*. Perception may determine that a single object or event is being perceived through two distinct modalities, and represent the object or event as possessing features perceived through both modalities.

✉ E. J. Green
ejgr@mit.edu

[1] Department of Linguistics and Philosophy, Massachusetts Institute of Technology, Cambridge, USA

This is what occurs if, say, we perceive a glass as both transparent and smooth. The second kind is *multisensory differentiation* of an object in space or time. Specifically, perception may use information from multiple modalities either to segregate an object from its spatial surroundings or to reidentify an object over time. For instance, perception might draw on tactual input to determine how far a partially occluded object extends in space, or it might use sound of a previously visible object that has now disappeared from view to reidentify it and update its location. Binding has received a fair amount of attention from philosophers, while differentiation has been less discussed. As I'll argue below, there are important unresolved questions about multisensory differentiation, and the issue has important implications for the architecture of perceptual processing.

The structure of the paper is as follows. Section 2 outlines basic tenets of the *object file* framework, which will be employed through much of the paper. Section 3 analyzes multisensory binding into two sub-tasks: First, perception must *identify* an object or event across modalities. Second, features from multiple modalities must be *integrated* with representations of multisensory identity. When this occurs, the features may also be integrated with *one another*. I first examine how perception might establish cross-modal identification, and then I explore several ways that unimodal feature representations might be modulated or transformed when they are integrated with representations from other modalities. This discussion results in an overall analysis of multisensory binding within the object file framework, summarized in Sect. 3.6. Section 4 turns to the issue of multisensory differentiation. I first suggest an empirical signature for multisensory differentiation. Multisensory differentiation must take place if we perceptually segregate or track an individual that would not be segregable or trackable through any single modality operating in isolation. Given this signature, I'll argue that the case for multisensory differentiation—at least as regards vision, audition, and touch—is presently inconclusive. I'll suggest some ways we might resolve the issue. Section 5 concludes by highlighting broader implications of multisensory object perception for the architecture and representational format of perception.

A caveat: The distinction between binding and differentiation is not intended as an exhaustive taxonomy of kinds of multisensory interaction in object perception. Other kinds are possible. For example, the modalities may interact in object recognition or categorization, or when selecting an object as the target of a motor action. I focus on the cases of binding and differentiation simply because they are two of the most fundamental abilities involved in object perception. To see this, note that recognition and categorization often depend on binding and differentiation, but not vice versa. To recognize an object as an avocado, plausibly you need to differentiate it from its surroundings and bind together its color, shape, and texture features. (Color or shape alone, for instance, may not be diagnostic of the object's category.) But we can differentiate and bind features to objects that we can't recognize, like abstract sculptures.

2 Object files

When a pigeon flies through your field of vision, you see it as a single thing that persists over time. You are also able to attribute a collection of features to it—you perceive its color, size, and motion, and these features seem to jointly characterize a single thing. The object file framework is a view about the perceptual representations that underlie these abilities.

Object files are representations that sustain reference to an individual (object or event) over time while also storing some of the individual's features (Kahneman et al. 1992; Green and Quilty-Dunn 2017). There is evidence that object files are composed of two separable constituents. First, there is a *singular constituent*, akin to a natural language demonstrative, which refers to an individual and continues to refer to it over time, despite changes in the individual's location or features.[1] Second, there is a *feature store*, which provides a temporary record of the individual's features. Some of these features can be retained even after the individual no longer possesses them. However, not every past feature is retained—only those that are selected for retention in visual working memory.[2]

Support for the singular constituent derives from studies that assess our ability to track objects through change. For example, multiple-object tracking studies have revealed that we can visually track about 4 target objects at a time, even as they move about randomly and their features change unpredictably (Pylyshyn and Storm 1988; vanMarle and Scholl 2003; Zhou et al. 2010). We can also see objects persist through change during apparent motion (Green 1986) and during the tunnel effect—where an object briefly passes behind an occluder and then appears to emerge on the other side (Flombaum and Scholl 2006). These findings suggest that perception can keep referring to an object despite changes in the features attributed to it. A natural proposal is that this capacity is based on a *separable singular constituent*—a constituent that picks out the object without representing any of its features, and can be maintained while feature representations are lost or updated.

Perhaps the strongest evidence for separable singular constituents comes from Bahrami's (2003) demonstration of change blindness during multiple-object tracking. Bahrami required subjects to track four targets among four distractors while also monitoring for changes in the targets' color (e.g., from red to green) or shape (e.g., from T-shaped to L-shaped). The data revealed that the participants could track targets at normal levels of accuracy while also failing to notice many of these feature changes. This was especially pronounced when the object's color or shape changed while it was briefly obscured by a mud splash. In this condition, color and shape change detection rates fell to about 60%. While this is still significantly above chance, tracking accuracy remained near 95%, indicating that there were a substantial number of cases in which a subject tracked a target successfully but failed to notice that it changed in color or shape.

[1] Pylyshyn (2007) labels this singular constituent a *FINST* or *visual index*.

[2] See Woodman and Vogel (2008) for evidence that we can selectively maintain just some of an object's features in visual working memory. See Hollingworth and Rasmussen (2010) for evidence that visual working memory is at least partly structured by object files.

Consider what this means for the representations responsible for tracking. Minimally, successful tracking requires a *correspondence* process, in which the perceptual system determines which objects perceived at time t_2 are continuations of the targets at time t_1. Suppose, then, that the object representations used in tracking contained no separable singular constituent. If so, then the only way for perception to *access* the representation of a target at t_1 would be to access a representation of at least some of the target's features. But note that this generalization must apply to the correspondence process as well. If an object at t_2 is deemed to be a continuation of a target at t_1, then a perceptual representation of at least some of the earlier target's features *needs* to be accessed. But Bahrami's (2003) findings indicate that this is not so. We are able to access a representation of an earlier target without accessing a representation of the target's features. This is just what happens when tracking is successful while feature changes go unnoticed. The most plausible explanation, I suggest, is that we can access a singular constituent that picks out the object without encoding its features.[3]

The feature-store component is supported by three strands of evidence. First, studies using the *object-reviewing paradigm* have shown that features are perceptually primed in an object-based manner. If a feature briefly appears on an object and then vanishes, we are subsequently faster to reidentify the feature if it reappears on the same object on which it initially appeared, even if the object has shifted location in the interim (Kahneman et al. 1992; Noles et al. 2005; Mitroff and Alvarez 2007). This is called the *object-specific preview benefit*. Second, there is evidence that object-specific feature stores are available for storage in visual working memory. If we are required to recall multiple features (e.g., shape and color), we are more accurate when they belong to the same object than when they belong to separate objects (Luck and Vogel 1997; Fougnie et al. 2010). Third, research using the *partial-repetition* paradigm has shown that when subjects are primed with an object instantiating a certain combination of features, they are subsequently *slower* to respond to an object that instantiates only one component of the earlier binding (Hommel 2004). For instance, if a subject is primed with a red X, then she will be slower to identify the shape of a green X than to identify the shape of a green O. As with the object-specific preview benefit, these partial repetition costs travel with objects as they move (Spapé and Hommel 2010).[4]

Importantly, there is evidence that object files are constructed not just in vision, but in audition as well. Zmigrod and Hommel (2009) found that the same partial repetition costs observed for visible objects could also be obtained for audible tones.

[3] Block (forthcoming) suggests that we can explain the capacity to track through change without appeal to separable singular constituents. Block's view is that there is a single iconic representation that picks out the object while also encoding some of its features (see also Carey 2009: p. 459). This representation comes to possess singular content by virtue of its functional role. Importantly, if Block's view is right, then accessing a visual representation of an object during tracking should require also accessing a representation of some of the object's features. It is incumbent on a proponent of this view to say *which* features this iconic representation depicts. Bahrami's (2003) findings suggest that it can't include shape or color. More generally, the evidence fits comfortably with the view that perceptual object representations contain separable singular constituents, while the iconic view can, at best, be carefully sculpted to accommodate the data.

[4] It might be wondered whether object files are used only for visual working memory, and not in online perception. However, there is evidence that this is not true. For instance, object files enable us to perceive objects as persisting across saccadic eye movements (Hollingworth et al. 2008), and holding objects in visual short-term memory interferes with the ability to visually track currently perceived objects (Fougnie and Marois 2009).

When subjects were primed with a soft, low-pitch tone, they were slower to respond to a loud, low-pitched tone than to a loud, high-pitched tone. Furthermore, there is evidence that visual and auditory tracking compete with one another for resources—it is harder to track a visual and an auditory target at the same time than to track a visual target on its own (Fougnie et al. 2018)—suggesting that a common mechanism may be involved in both tasks.

My working assumption in what follows will be that perception represents objects via object files, and that object files are at least sometimes constructed in modalities besides vision—specifically, in audition and touch.[5] This is of course an empirical issue. The soundness of the assumption must ultimately be gauged by its explanatory power when applied to object perception in non-visual modalities.

Section 3 explores multisensory interaction during perceptual binding, which I'll understand to involve the placing of feature representations within object files. Section 4 asks how the modalities might interact when segregating an object or reidentifying it over time. The nature of these processes determines, respectively, the conditions under which an object file is opened, and how the file is maintained after it is opened. Many of these issues do not *require* assuming the object file model. We certainly don't need to assume the reality of object files in order to think that multisensory binding takes place, or that the modalities can cooperate in object differentiation. Accordingly, I'm sure that much of what I'll say would carry over to other frameworks. But I'll leave the required adjustments for another time.

3 Multisensory binding

Let's say that a perceptual representation exhibits multisensory binding if it represents that a single object or event jointly possesses features F and G, where F and G are perceived through distinct modalities. O'Callaghan (2014, 2017) adduces a number of considerations in support of multisensory binding at the level of perceptual experience. It seems, for instance, that we can *non-inferentially judge* that an object has both visible and tactual features. It doesn't feel like we engage in post-perceptual inference when we decide that the tomato we perceive is both red and smooth. It seems like we do it just by endorsing the content of our perceptual experience.[6] Moreover, it seems that multisensory binding can sometimes be *illusory*. In the ventriloquism effect, we experience the ventriloquist's voice (an audible feature) as bound with the puppet's mouth movements (a visible feature), when this is not actually the case. O'Callaghan argues that this does not involve any illusion regarding audible or visible features themselves. Rather, we misperceive how these features are bound together.

[5] My discussion will not address object perception in the chemical senses. Nonetheless, many of the same issues are likely to arise in these cases. For some illuminating discussions of object perception in olfaction and taste, including the problems of figure-ground organization and feature binding, see Batty (2014), Carvalho (2014), and Millar (2017).

[6] Of course, the appeal to apparent non-inferentiality is not probative [as O'Callaghan (2014) acknowledges]. It's possible that we are simply *mistaken* in thinking that we judge the tomato to be both red and smooth merely by endorsing the contents of perception. Perhaps, for example, we do engage in post-perceptual inference, but the inference is unconscious.

Although doubts have been raised about some of these arguments (Briscoe 2017; Spence and Bayne 2014), I think there is a strong *prima facie* case for multisensory binding in perceptual experience [for further discussion of the issue, see Fulkerson (2011), O'Callaghan (2008, 2017), and Deroy (2014)]. However, my concern here is different. Even if multisensory binding is present in perceptual experience, this underdetermines the nature of the perceptual mechanisms that enable these experiences. I am interested in exploring the different kinds of multisensory interaction in perceptual processing that may subserve multisensory binding at the level of perceptual experience. Note, however, that some multisensory interactions (including interactions in binding or object differentiation) might happen purely unconsciously, without any phenomenal upshot. Previous papers have carefully explored the various ways that multisensory interactions in perceptual processing might be revealed in perceptual experience (Macpherson 2011; Deroy 2014; Spence and Bayne 2014; O'Callaghan 2017; Briscoe 2017). I'll have little to add to this issue here.

Multisensory binding requires, at minimum, two perceptual feats. First, perceptual processes must establish that the *same* object or event is being perceived through two or more modalities. That is, perception must achieve *cross-modal identification*. Second, perception must *combine* representations of cross-modal identification with feature representations from the modalities between which the identification has been made. This second feat raises the possibility of cross-modal effects that are *contingent* on cross-modal identification—multisensory interactions that only take place if cross-modal identification has been established. In what follows, I'll explore a variety of ways that both feats can be achieved. This will lead to an overall analysis of multisensory binding within the object file framework, summarized in Sect. 3.6.

3.1 Cross-modal identification

Why might we expect perceptual systems to establish cross-modal identification? The answer can be illustrated most clearly by considering the *function* of multisensory coordination.

The perception of properties through one modality is often influenced by the perception of properties through another modality. For instance, in the ventriloquism illusion, auditory perception of location is biased by visual perception of location. And in the McGurk effect, auditory perception of speech phonemes is biased by the visual perception of mouth shape and movement. However, while multisensory coordination is often demonstrated through illusions, it is typically useful for promoting accuracy and reducing noise. For example, if vision and touch produce independent noisy estimates of size, then perceptual systems can produce a more precise and accurate size estimate by taking a weighted average of the two unimodal estimates—that is, an average weighted by the reliability (inverse variance) of the two estimates (van Dam et al. 2014; Briscoe 2016).

But multisensory coordination is not always advantageous. In general, coordination is only advantageous if the information processed by two modalities derives from the same object or event. It is only beneficial to integrate visual and tactual information about size if vision and touch are converging on a single object. Otherwise, perception

runs the risk of biasing one or both estimates away from the true value. This is the case in spatial ventriloquism. Auditory localization is *misled* by vision because the modalities are in fact responding to separate events. To minimize these sorts of errors, we might expect the perceptual systems responsible for multisensory coordination to make reliable decisions about multisensory identity. Indeed, these decisions are reflected in leading computational models of multisensory coordination (Körding et al. 2007; Shams and Kim 2010).

But while it would plainly be *useful* for perception to represent multisensory identity, are such representations actually constructed? If so, how are they structured? Perhaps the most straightforward option is that when the perceptual system establishes multisensory identity, it *links* a pair of object files—one from each modality—via a representation of the identity relation. Thus, where a and b are the singular constituents of object files A and B, respectively, perception might link A and B via an explicit representation of the form <a = b>.

Following Recanati (2012), we can distinguish linking from a separate operation of *merging*. When files are linked, we represent that their referents are identical. However, information within each file remains clustered independently from information in the other (though some information may be allowed to pass between the files). Merging, on the other hand, pools the information in A and B into a single file. Either the information from one file is transferred to the other, or a new file is constructed and the information from both files is fed into it. If merging occurs, then no explicit representation of identity is needed. Rather, a *presumption* of identity is involved in collecting information from both modalities into a single file (see Recanati 2012: p. 42).[7]

When I speak of cross-modal identification without qualification, I intend to remain neutral between the linking and merging models.[8] The current subsection considers evidence that perception establishes cross-modal identification. This evidence is

[7] It's easy to see why it might be prudent to link files rather than merge them (Recanati 2012: pp. 44–47). Suppose that you initially collect information about Mark Twain and Samuel Clemens into two separate files before being told that the men are identical. You think this testimony is reliable, but you are less than wholly confident in its truth. Merging the two files straightaway would be risky, because if the testimony turned out to be incorrect, one would then have to re-segregate the information within the merged file into those bits associated with Twain and those associated with Clemens. Similar considerations might support linking without merging within multisensory perception, but here the situation is perhaps less severe. Suppose that two unisensory object files are merged into a single multisensory file, but the perceptual system then receives information that the two objects are really distinct. It must then perform the extra task of re-segregation. However, this task may be considerably simpler than in the cognitive case *if*—but *only* if—feature representations within the multisensory file remain tagged according to the modality from which they originated. Note, however, that this would require perception to always retain the initial unimodal feature representations after multisensory binding. This will not always be the case if perception exhibits *constitutive* binding (see Sect. 3.5).

[8] The distinction between linked and merged object files is similar to Nudds' (2014) distinction between "cross-modal" object perception and "amodal" object perception. In Nudds' framework, cross-modal object perception occurs when we coordinate information between separate, modality-specific object representations, while amodal object representation involves constructing a single representation that binds features delivered through multiple modalities. Nevertheless, the two distinctions are not the same, because cross-modal object perception in the foregoing sense does not entail the construction of linked object files. It is possible, in principle, to *coordinate* the information contained within two modality-specific object files without representing their referents as *identical* (see O'Callaghan 2017 for this point).

consistent with either the linking or merging model. The next subsection considers evidence that specifically bears on the merging model.

The linking and merging models both make two predictions. First, we should expect cross-modal identification to be based on reliable cues to whether two modalities are really converging on the same individual. That is, we should expect the process to be *sensible* in light of the information available to perception. Second, we should expect that, once completed, cross-modal identification can influence various forms of multisensory coordination. These forms of coordination should be *stronger* or *more likely* when identity is established than when it isn't (Welch and Warren 1980). I contend that both of these predictions are borne out in the case of audio-visual coordination.

But before turning to this evidence, let me address a possible concern. One might think that cross-modal identification could *not* occur between vision and audition simply because the individuals perceived through these modalities belong to different ontological categories. Audition targets *sounds*, while vision targets *material objects*. Following others (O'Callaghan 2008; Nudds 2009), however, I think that this line of thought is mistaken. It's highly plausible that *events* involving material objects are perceptible through both audition and vision. We can both *see* and *hear* collisions, shatterings, and scrapings. Note that this does not require holding that such events are *identical* to sounds, although this is an available option (see, e.g., Casati and Dokic 2009). My claims here will be neutral about whether we hear external events because they are identical to sounds or we hear them *by way* of hearing the sounds they produce. All that matters is that we hear them.

What might the cues to cross-modal identification look like? In the audio-visual case, appropriate identification would need to be signaled by *synesthetic correspondences* between visible and audible features (Parise and Spence 2009). That is, there should be mappings between audible and visible features that our perceptual systems use to determine that the two modalities are converging on the same object or event. Such correspondences could be either hardwired or learned.

O'Callaghan (2014, 2017) has recently appealed to synesthetic correspondences in audio-visual speech perception in an argument for multisensory binding. Vatakis and Spence (2007) found that audio-visual coordination in the perception of spatiotemporal properties is enhanced when the visually perceived gender of a speaker matches the auditorily perceived gender of a concurrent speech stream. (The relevant sort of coordination was a version of the temporal ventriloquism effect, which I'll describe below.) O'Callaghan takes this to show that multisensory interactions in speech perception are guided by cross-modal identification. When gender features are matched across modalities, cross-modal identification is more likely, and multisensory coordination is boosted.

I agree that there is compelling evidence for the use of synesthetic correspondence in the case of speech perception, and that this supports cross-modal identification in speech perception. However, there are concerns with generalizing from the case of speech perception to audio-visual perception more generally. In particular, several theorists (including Vatakis and Spence themselves) have suggested that speech perception is unique, and may even constitute a distinctive 'mode of perception' (Tuomainen et al. 2005; Vatakis and Spence 2008; Briscoe 2017; although see Vroomen and Stekelenburg 2011). For example, Tuomainen et al. (2005) presented subjects with a

sine-wave speech stimulus that was ambiguous between speech and non-speech. They found that certain audio-visual interactions (viz., the McGurk effect) only occurred when the subject was able to recognize the stimulus as speech, and not otherwise. Tuomainen et al. take these results to demonstrate a "special speech processing mode, which is operational also in audio-visual speech perception" (B20). Briscoe (2017) similarly suggests that "audio-visual speech processing may be special" (8), and questions the extent to which synesthetic correspondence guides audio-visual coordination outside the domain of speech. If so, then it is possible that cross-modal identification occurs only in the special speech-processing mode.

In what follows, I argue that synesthetic correspondences are used outside the domain of speech perception, suggesting that cross-modal identification between vision and audition is a more general phenomenon. I'll focus on the case that I think provides the best evidence for this: the correspondence between audible amplitude envelope and visible collision.

A sound's amplitude envelope is, roughly, its change in intensity over time. Typical impact events, such as collisions or strikings, are associated with an abrupt rise followed by a gradual decay. Grassi and Casco (2009) refer to this as a *damped* envelope. This can be contrasted with sustained sounds, such as the sound of drawing a bow across a violin, whose amplitudes are more stable over time. If audio-visual coordination is guided by cross-modal identification, then we might expect it to be sensitive to the synesthetic correspondence between damped envelope and visible impact. For this is relevant to whether the modalities are converging on the same physical event.

It turns out that audio-visual coordination is guided by this correspondence. First consider the *sound-induced bouncing* effect. Suppose that two objects begin at opposite ends of a computer screen, gradually approach and eventually overlap one another, and finally two objects emerge following the overlap. In this case, two percepts are possible. Either we can see the objects *bounce off* one another, or we can see them *stream through* one another. Streaming percepts tend to predominate in normal circumstances. However, Sekuler et al. (1997) found that when a sound is played at the moment of overlap, subjects are more likely to perceive bouncing. This is a case of audio-visual coordination. Auditory input guides the visual perception of motion trajectory. The effect may be mediated by perceiving a causal relation between the two visible objects and the sound (see O'Callaghan 2015, 2017).

Subsequent work has shown that not just any sound will promote the bouncing percept. Rather, the sound must be consistent with an impact event. Grassi and Casco (2009) used two sounds: a damped sound (as described above), and a "ramped" sound, in which the damped amplitude envelope was reversed to produce a gradual rise followed by abrupt decay. Even though the two sounds were equated for average intensity and overall duration, only the damped sound led to an increase in bouncing percepts. Thus, the sound-induced bouncing effect is sensitive to the synesthetic correspondence between damped sounds and visible impact. This fits with the view that audio-visual coordination is based in part on cross-modal identification.

Next consider the *temporal ventriloquism* effect. In this phenomenon, auditory perception biases the visual perception of an event's temporal onset or duration. In an elegant demonstration of this effect, Morein-Zamir et al. (2003) asked subjects to simply judge which of two light flashes appeared first. They showed that when sounds

were presented before and after the visible flashes, temporal order judgments were more accurate, suggesting that the sounds had attracted the visible events, making them appear further apart in time. If, on the other hand, the sounds were presented *between* the flashes, then judgments were less accurate, suggesting that the sounds pulled the visible events closer together. (See also O'Callaghan (2017) and Nudds (2014) for discussion of the temporal ventriloquism effect.)

For present purposes, the critical question is whether temporal ventriloquism is guided by a process of cross-modal identification. Chuen and Schutz (2016) performed an important test of this issue. Chuen and Schutz observed that *if* temporal ventriloquism occurs, then the perceived temporal interval *between* an audible and a visible event should be *reduced*, because the sound attracts the perceived onset of the visible event (see also Vatakis and Spence 2007). Accordingly, experimenters can explore the factors that influence temporal ventriloquism by way of audio-visual temporal order judgments. If temporal ventriloquism is stronger for one audio-visual pair than for another, then audio-visual temporal order judgments should be less accurate in the former case. Chuen and Schutz exploited this idea. Their stimuli included the sound and sight of a played cello along with the sound and sight of a played marimba. The marimba produces a characteristic damped sound, while the cello produces a sustained sound. Subjects were shown a sight-sound pair and were asked to indicate their order of occurrence. Consistent with the idea that temporal ventriloquism is guided by cross-modal identification, subjects' judgments were more sensitive to temporal order when, say, the sight of a cello was paired with the sound of a marimba than when the same instrument was presented to both modalities.[9,10]

One might object that the foregoing evidence does not show that audio-visual coordination needs to rely on cross-modal *identification*. We can explain the data, the objector insists, if synesthetic correspondences are simply used to establish cross-modal *association*. On this alternative account, the mapping between damped amplitude envelope and visible impact is used to establish a *link* or *association* between the audible event and the visible event, but the events are not literally perceived as *identical* (see Fulkerson (2011) for roughly this proposal).[11] And when events are associated, multisensory coordination is boosted.

If audible and visible events are either identified or associated, let's say that they are "paired". Although the association account of pairing is difficult to rule out definitively,

[9] One might wonder whether it was really the correspondence between damped envelope and visible impact—rather than the higher-level correspondence between familiar instrument categories (*cello* vs. *marimba*)—that was responsible for this effect. However, other studies have found that when two familiar instruments with *similar* amplitude envelopes are used—namely, a guitar and a piano—instrument matching does *not* affect temporal order judgments (Vatakis and Spence 2008). Thus, the effects Chuen and Schutz observed were plausibly due to the low-level correspondence between damped envelope and visible impact, and not to category-level correspondence.

[10] There is also evidence that visually apparent *gesture* duration influences auditorily apparent *sound* duration, and that these interactions are also guided by cues to audio-visual event identity (Schutz and Lipscomb 2007; Schutz and Kubovy 2009).

[11] Fulkerson (2011) enlists the association account in support of a view about how to distinguish unisensory from multisensory experiences. His view is that unisensory experiences represent a collection of features in a single binding structure where they are represented as jointly characterizing a single object, while multisensory experiences involve an association among multiple such binding structures.

I believe that further evidence clearly favors the identification account. There are logical constraints on the identity relation that do not apply to association. Suppose that we hear a single sound at the same time as we see multiple visible flashes. Because identity is one-to-one, the sound can be deemed *identical* to at most one of the visible flashes. But nothing prevents the sound from being *associated* with two or more of them. Accordingly, if we find that the pairing of audible and visible events is subject to a one-to-one constraint, this suggests that pairing is subserved by identification, not mere association.

Van der Burg et al. (2013) directly tested for numerical constraints on audio-visual pairing. Subjects were shown a circular arrangement of 24 discs. Every 150 ms, a random subset of the discs changed color from black to white, or vice versa. At an unpredictable point during the trial, one of the change events was accompanied by a tone. The subject was told beforehand that the discs that changed color alongside the tone were the targets, and the task was simply to remember them. At the end of a trial, the subject was directed to a particular disc and asked whether it had been one of the targets. Note that there are two key steps involved in completing this task. First, the tone must be *paired* with the changing discs and not the others, so the subject can determine which discs are the targets. Second, the discs that changed must be retained in visual working memory.

Critically, Van der Burg et al. (2013) used subjects' hit rates and false alarms to estimate limits on the number of targets that could be remembered (see Cowan 2001). They found that capacity limits never exceeded 1 (the average across experiments was 0.75). In other words, given a single auditory cue, at most one visible disc could be remembered. This is consistent with two possibilities: First, there might have been a capacity-1 limit on audio-visual pairing (the first step), as predicted by the cross-modal identification account, but not the association account. Second, there might have been a capacity-1 limit on the number of discs that could be held in visual working memory (the second step). A further experiment ruled out the latter possibility. When the targets were instead indicated by a *visible* cue (a color change), capacity recovered to typical working memory levels (roughly 3–4 objects). This suggests that the capacity-1 limit specifically constrained audio-visual pairing. The cross-modal identification account predicts this constraint, while the association account does not.

I don't intend the foregoing evidence to conclusively settle the dispute between the identification and association accounts of audio-visual pairing. It's possible that there is a brute one-to-one constraint on cross-modal association, with no explicit representation of cross-modal identity (and also no merging of unisensory object files into multisensory object files). However, while the identification account explains the one-to-one constraint on audio-visual pairing, the association account leaves it unexplained. (After all, it can't be due to general working memory limits.) Thus, I believe that the evidence clearly puts the burden of proof on the association view. The identification account explains all of the data that the association account explains, plus more. So, absent strong additional reasons in favor of the association account, we should prefer the identification account.

Thus, there is compelling evidence that audio-visual coordination is guided by cross-modal identification. Coordination is stronger when cross-modal identification is possible. And we should note that the mapping between damped envelope and visible

impact is just one type of synesthetic correspondence. Examples could be multiplied. There is also evidence that audio-visual coordination is sensitive to correspondence between visible size and audible pitch—smaller objects are associated with high-pitched tones (Parise and Spence 2008, 2009)—to correspondence between visible movement and modulation in pitch—vertical ascension in space is associated with rising pitch (Jain et al. 2008)—and to correspondence in the internal temporal structure of visual and auditory stimulus sequences (Parise et al. 2013). It is thus highly plausible that perception makes cross-modal identifications, and that these are used to guide multisensory interaction.

3.2 File merging

But *how* is cross-modal identification achieved? The foregoing evidence is consistent with the view that perceptual object files are always wholly unisensory, but unisensory files can be linked given appropriate synesthetic correspondences. Linking could facilitate coordination between the feature representations stored in each file, while otherwise keeping the two files insulated from each other.

Merging seems to mark a deeper form of multisensory object perception. For, in this case, modality-specific object representations are combined into a single object representation that is proprietary to *neither* modality. The resulting file is multimodal through and through. It is available for processing in more than one modality. But how could we determine that two object files have been merged and not merely linked? Plausibly, we might draw on the very same kinds of evidence that led us to believe that *unisensory* information is consolidated into a single object file. If the very same paradigms that support *unisensory* object files also support *multisensory* object files, then this gives us reason to think that perception forms multisensory object files.[12]

One piece of evidence for multisensory object files involves the partial repetition paradigm introduced above. Recall the basic finding: When subjects are primed with a particular feature conjunction (e.g., red square), they are slower to respond to an object that repeats just *one* of the primed features (e.g., a green square) than to an object that repeats either both or neither of them. Importantly, Zmigrod et al. (2009; see also Zmigrod and Hommel 2010) found that the *same* pattern appears in multisensory contexts. For example, subjects primed with a red object accompanied by a high-pitched tone were slower to respond to a red object paired with a low-pitched tone than to a blue object paired with a low-pitched tone. Similar results were found for pairings of tones and tactual vibrations. Perhaps this is because participants formed object files whose feature stores combined audible and visible features.

A second piece of evidence for multisensory object files derives from Jordan, Clark, and Mitroff (2010), who used the object-reviewing paradigm. Jordan et al. first showed subjects a preview display containing objects at the top and bottom of the display.

[12] See also O'Callaghan (2017: pp. 161–165), who argues that if the very same evidence taken to support unisensory binding also support multisensory binding, then we should conclude, other things being equal, that multisensory binding occurs. I should note that O'Callaghan (2014, 2017) and Briscoe (2017) also mention two of the main studies discussed in this subsection (Zmigrod et al. 2009 and Jordan et al. 2010), so the evidence itself is fairly familiar. However, I aim to critically evaluate these experiments to figure out just what they show and what they leave unresolved.

Visible stimuli, such as a telephone and a dog, briefly appeared on the objects and then vanished. Next, the objects shifted to opposite sides of the display. Finally, an audible sound, such as a telephone ring or a bark, was presented from a speaker located either on the left or right side of the display, and the subject needed to report whether the sound matched either of the stimuli from the preview display. Jordan et al. found that responses were fastest when the sound matched the visible stimulus that had appeared on the object that now occupied the location where the sound originated. This seems to show that the object-specific preview benefit generalizes across modalities. A picture of a telephone primes a ringing sound, and does so in an object-specific manner. Jordan et al. conclude that object files "store object-related information in an amodal format that can be flexibly accessed across senses" (500).

Nonetheless, while these data provide interesting support for multisensory object files, I think there are reasons for caution. I'll consider the studies in turn.

First, while the Zmigrod et al. (2009) results indicate that perceptual priming is sensitive to multimodal feature combinations, we can't be certain that the pairs of features were genuinely bound to *objects*. In particular, because the priming and test stimuli were presented at the same location, it is possible that the multimodal feature pairs were bound to this location rather than to the objects that occupied it (compare Austen Clark's (2000) feature-placing model of binding). On this option, the multimodal combinations resulted in representations of the form: <red and high-pitched over there>, rather than <object O is red and high-pitched>. Luckily, there is an obvious way to resolve this issue. It needs to be determined whether multimodal partial repetition costs still appear when objects *shift location* during a trial. Recall that this datum *has* been produced for *unimodal* partial repetition costs (Spapé and Hommel 2010).

Fortunately, the Jordan et al. (2010) study avoids the object/location ambiguity because the object-specific preview benefits were observed across changes in location. Nonetheless, the study leaves another issue unresolved.

Let's say that an object file is *richly* multisensory if it can house any feature, regardless of modality, that is perceptually attributed to the object in question. The Jordan et al. findings indicate that there are some feature representations activated in response to both a visible picture (e.g., the telephone in the preview display) and an audible sound (e.g., the ringing sound in the test display), and that these representations can be entered into object files. But what are these feature representations representations *of*? A natural answer would be: high-level categories like *telephone*.[13] Note, however, that there is nothing distinctively *auditory* about this category—it can be identified through multiple modalities. Accordingly, the Jordan et al. (2010) findings leave the following possibility open: Object files can store at most a combination of low-level features delivered through a single modality together with certain high-level categories. However, some of these high-level categories can be perceived through other modalities as well, and when they are, object-specific preview benefits are produced. If this is right, then we should expect that an object file may store both *red* and *telephone*, but cannot store both *red* and *loud*. To the best of my knowledge, this possibility remains

[13] See Gordon and Irwin (2000) for further evidence that high-level categories can be housed in object files, and see Quilty-Dunn (2016) for discussion of the implications of this evidence for the format of perceptual object representations.

untested. Accordingly, it is possible to accommodate the Jordan et al. results without positing *richly* multisensory object files, although we do need to say that some of the contents of object files can be accessed via inputs to more than one modality.

To sum up: There is evidence for multisensory object files. Some of the feature representations within an object file can be activated by information delivered to multiple modalities. Nevertheless, it remains an open question whether object files are ever richly multisensory. Specifically, it remains to be determined whether object files can freely combine both low-level features and high-level categories across modalities.

Here's where we are. Cross-modal identification is a precondition for multisensory binding. To bind features from separate modalities as features of a single individual, perception must register that the same individual is being perceived through both modalities. So far, I've considered two views about how cross-modal identification might take place within the object file framework: the linking and the merging model. Evidence plainly supports *some* kind of cross-modal identification, although it is less decisive regarding which of the two models is right. I now shift to a different issue. Suppose that cross-modal identification has been established. Perception thus represents a single individual as presented to more than one modality. When unisensory feature representations are *integrated* with this information, this creates the possibility of *identity-contingent* multisensory interactions—interactions that only occur if cross-modal identification has been established. In what follows, I distinguish three grades of identity-contingent interaction during feature binding. I'll call these *non-cooperative*, *cooperative*, and *constitutive binding*.

3.3 Non-cooperative binding

The first possibility is that *no* modulation takes place when two features are cross-modally bound. Plausibly this will be true in certain cases. Suppose that an object's visible orientation provides no useful information about its tactually perceived texture. In that case, while the two features may be bound together into the same object file if merging occurs, this is the full extent of their interaction. Otherwise, it is just as if they were perceptually attributed to separate objects. Call this *non-cooperative binding*. This type of binding still presupposes cross-modal identification. It is just that cross-modal identification does not lead to changes in perceptual representation of the bound features.

We've already encountered compelling evidence against the non-cooperative model. Cross-modal identification does facilitate multisensory coordination—for instance, in the sound-induced bouncing effect and in temporal ventriloquism (Grassi and Casco 2009; Chuen and Schutz 2016). Thus, it's not true that unimodal feature representations are always unaffected by multisensory binding.

Nevertheless, it's important to keep in mind that some instances of multisensory binding may be fully non-cooperative. Thus, we shouldn't take a lack of coordination between unimodal feature representations to show that these representations aren't housed within the same object file, or within linked object files. Indeed, it's quite plausible that *unimodal* feature binding is often non-cooperative. This would even be the norm according to views on which vision analyzes separate feature dimensions,

like shape and color, along distinct pathways and only later combines them into single object representations (e.g., Treisman 1988). While the resulting features are bound, this is the extent of their interaction. Some cases of multisensory binding may work similarly.

3.4 Cooperative binding

Cooperative binding occurs if the perception of features in one modality influences the perception of features in another modality, *and* the nature of this influence is contingent upon having established that the same object or event is being perceived through both modalities.[14]

We've already encountered evidence for cooperative binding, which I won't rehearse. But cooperative binding can be further analyzed. de Vignemont (2014) draws a useful distinction between *additive* and *integrative* binding. Additive binding occurs when perception binds *non-redundant* features from separate modalities to the same object. Features are non-redundant if they fall along different dimensions.[15] An example would involve perceiving a tomato as both red and smooth. *Integrative* binding occurs when perception binds *redundant* features perceived through separate modalities to the same object. Features are redundant if they fall along the same dimension. For instance, during ventriloquism we might bind distinct visual and auditory estimates of an event's location or temporal onset. Or we might integrate visual and tactual estimates of an object's size (Ernst and Banks 2002).

Either additive or integrative binding can be cooperative. Indeed, this appears to be the norm for integrative binding. When perception binds visible and audible estimates of location, these are not merely attributed to the same object. Rather, the two estimates bias one another in a manner determined by their relative reliability (Alais and Burr 2004; van Dam et al. 2014). This is to be expected. When the features delivered by two modalities fall along the same dimension, binding them creates the possibility of conflict. Because a single object can occupy at most one location, the perceptual system faces pressure to resolve discrepancies between visual and auditory localization

[14] Cooperative binding has important parallels to O'Callaghan's (2017) notion of coordinated multisensory awareness. However, a key difference should be noted. As O'Callaghan points out, it is possible for the perception of features across modalities to be coordinated even if those features are not perceptually bound (see his discussion of conflict resolution across the modalities on pp. 158–160). Thus, while cooperative binding trivially requires binding, coordinated multisensory awareness does not. I note one further difference of emphasis between the current approach and that of O'Callaghan (2017). O'Callaghan treats multisensory binding as a strictly stronger grade of multisensory interaction that mere multisensory coordination, because it "marks the point at which perceptual awareness can no longer be characterized in modality-specific terms" (160). I agree with this observation, but would add that there are other respects in which coordination is more demanding than binding. As the discussion of non-cooperative binding highlights, collecting feature representations from two modalities into the same object file doesn't require substantive coordination between these representations.

[15] I borrow the redundant/non-redundant terminology from de Vignemont (2014: p. 130). de Vignemont also counts features as non-redundant when they are features of different parts of an object. .

in order to maintain a coherent representation of the world. So we should expect integrative binding to be cooperative.[16]

However—and this is an important point—cooperation is not *restricted* to cases of integrative binding. Even if two modalities process separate dimensions, values along one dimension can still be *informative* about values along the other if the dimensions happen to be correlated in the perceiver's environment.[17] For instance, suppose that certain colors render certain textures more likely. Then it would make sense for texture processing to take color information into account, even though there would be no obvious *conflict* in representing a highly improbable combination of texture and color. One potential case of cooperative additive binding involves influences of perceived pitch on perceived texture. Jousmäki and Hari (1998) had subjects rub their palms together while a microphone recorded the sound that was produced. The sound was played to subjects through headphones either unaltered or adjusted in pitch or intensity. Jousmäki and Hari found that when the sound was increased in pitch, subjects perceived their hands as drier or more "paper-like". This phenomenon was plausibly perceptual as it was dependent on fine-grained temporal characteristics of the stimuli. A small 100 ms temporal displacement between the tactual sensation and the sound significantly diminished the effect. Patently, pitch and texture fall along separate dimensions. So if both features were bound to the same object or event (admittedly, the data do not conclusively settle this) their binding was additive, not integrative.

Of particular interest, there is evidence that additive binding may *become* cooperative through perceptual learning. Over the course of 500 training trials, Ernst (2007) introduced subjects to a correlation between luminance and stiffness—for instance, darker objects could tend to be stiffer than brighter objects. After training, subjects were given a discrimination task that required them to judge which of two comparison stimuli was distinct from a standard stimulus in either luminance or stiffness. Ernst reasoned that the brightness-stiffness correlation might affect subjects' discrimination performance. Here's why. Suppose that dark objects tend to be stiff while bright objects tend to be soft. If perception makes use of this correlation, then the two dimensions should bias one another. For instance, a soft and dark object should tend to appear both somewhat stiffer and somewhat brighter than it really is. Now suppose that a standard stimulus is intermediate in both brightness and stiffness. Then it should be more difficult to discriminate this standard stimulus from a comparison stimulus that is both softer and darker than from a comparison stimulus that is both stiffer and darker. Why? Because the differences in the former case should be attenuated due to the cross-modal bias: The stiff and bright stimulus should appear both somewhat softer and somewhat darker, assuming the learned correlation exerts a cross-modal influence. And this is what was found. Following training, subjects became worse at making just

[16] Still, non-cooperative integrative binding is at least possible, if not actual. A sensory system wholly untroubled by cross-modal discrepancies might, for instance, bind visual and auditory estimates of location but not adjust either estimate in light of the other.

[17] For this reason, it is somewhat misleading when de Vignemont writes: "Additive binding is a matter of all or nothing. Either the features are bound or not. By contrast, integrative binding is a matter of degree. The information can be more or less bound" (132). de Vignemont's central case for treating integrative binding as a matter of degree is that sensory estimates from two modalities can bias one another to a greater or lesser degree (e.g., in the rubber hand illusion or the ventriloquism effect). But cross-modal biases can occur—and, indeed, can be a matter of degree—*either* in additive binding *or* in integrative binding.

those discriminations that would have been expected to become more difficult given the correlation encountered during training. But brightness and stiffness are independent dimensions. Thus, assuming that the properties were bound, they were additively bound, not integratively bound.

Thus, cooperation plausibly occurs in both additive binding and integrative binding. When we bind features following cross-modal identification, the perception of features through one modality can influence the perception of features through another modality, even if the features fall on different dimensions.

3.5 Constitutive binding

There is, however, a further operation that is only possible in the integrative case. Occasionally, when two modalities represent features along the same dimension, the unimodal feature representations may be *discarded* in favor of a single multisensory representation. In this case, the resulting feature representation is *constitutively* multisensory. It is not proprietary to any one modality, and it contains no unimodal constituents.[18] And it is able to take part in the algorithms of more than one modality. I'll call this *constitutive binding*.

We saw earlier that it is possible to transition from non-cooperative binding to cooperative binding through perceptual learning. If constitutive binding occurs, then it stands to reason that a similar transition from cooperative to constitutive binding may be possible.

But does constitutive binding occur? Some cases of integrative binding are probably not constitutive. During spatial ventriloquism, arguably we retain separate visual and auditory representations of location. The primary evidence for this is that while vision biases the auditory perception of location, the bias is often only partial (Bertelson and Radeau 1981; Briscoe 2016: p. 124). One hears the puppeteer's voice as originating somewhere *between* its actual origin and the location of the seen puppet. This accounts for the palpable sense of conflict or tension. We perceive the very same event as having discrepant locations—a heard location and a seen location. Likewise, when a subject touches a vibrating loudspeaker while also hearing a sound slightly displaced from the touched speaker, the sound is often heard to originate somewhere near the touched speaker. But, again, the bias is not complete. There remains a slight discrepancy between heard and felt locations (Pick et al. 1969; Welch and Warren 1980). Thus, the auditory and proprioceptive location estimates are probably not constitutively bound, even if they are cooperatively bound.

What sort of evidence *could* establish constitutive binding? I'll consider the two forms of evidence that, to my mind, come closest to this.

One way to motivate constitutively multisensory feature representations would be to show that instances of the very same feature representation can be produced via inputs to either of two modalities. For instance, if either visual input alone or tactile

[18] I suspect that this is what de Vignemont (2014) has in mind when she suggests that multisensory binding can be so strong that "the experiences melt into each other" (130). While de Vignemont casts this as a general feature of integrative binding, I would disagree. I suggest that it occurs only when integrative binding is also constitutive.

input alone can activate the same feature representation, then it would be natural to conclude that the representation is neither constitutively visual nor constitutively tactual. Rather, it is a multimodal representation shared between the two modalities.

This raises the issue of how to tell whether the same feature representation is producible via inputs to more than one modality. Some have taken *multisensory adaptation* to show this. Konkle et al. (2009) found that exposure to a given motion direction in vision could induce a motion aftereffect in touch, and vice versa. If a subject saw upward motion during adaptation, then she was more inclined to feel a neutral stimulus as moving downward. Adaptation to motion direction has also been demonstrated between vision and audition (Jain et al. 2008; Berger and Ehrsson 2016). Similarly, it has been found that rate perception adapts between vision and audition. Exposure to a 5 Hz visual stimulus led a 4 Hz auditory stimulus to be perceived as slower, and vice versa (Levitan et al. 2015).

In a discussion of the upshot of their results, Konkle and Moore (2009) write: "[P]rocessing tactile motion depends on circuits that were previously adapted by visual motion processing. Similarly, the processing of visual motion depends on circuits adapted by tactile motion. Crossmodal motion aftereffects reveal that visual and tactile motion perception rely on partially shared neural substrates" (480). This conclusion may seem to be bolstered by physiological evidence showing that moving tactual stimuli elicit activation in visual motion processing areas (Beauchamp et al. 2007). Similarly, Berger and Ehrsson (2016) conclude: "Visual and auditory motion perception rely on shared neural representations" (5).

If visual and tactile motion perception rely on the very same neural substrates, then it is quite plausible that they use the same representations. However, such results should be interpreted with caution. There are two issues with the transition from cross-modal adaptation to constitutively multisensory feature representations.

First, it is difficult to rule out the possibility that cross-modal adaptation aftereffects take place at the level of post-perceptual decision, rather than perception per se (Storrs 2015). For instance, it may be that after seeing upward visual motion, a subject is simply more likely to *judge* that a neutral tactual stimulus is moving downward, even if her underlying perceptual state is unaffected by the adaptation. In the case of unimodal visual adaptation, it is common to rely on *retinotopically specific* adaptation aftereffects to rule out the post-perceptual judgment hypothesis (e.g., Block 2014). For if adaptation aftereffects are accentuated at the retinotopic location of the adaptation stimulus, this strongly suggests that at least some of the adapted neural regions responsible for the effect are located within early, retinotopically organized visual areas. However, no retinotopic specificity was reported in the studies just mentioned.[19]

[19] Although I am unaware of any cross-modal adaptation study that reveals retinotopically specific aftereffects, there is some evidence for *spatiotopically* specific cross-modal aftereffects—aftereffects specific to a particular region of external space. Anobile et al. (2016) found that after a 6-second period of rapid finger-tapping, subjects underestimated the numerosity of a visible array of dots. Conversely, after a same-length period of slow tapping (and so fewer taps), subjects tended to overestimate the numerosity of the visible array. Critically, these effects were *only* found when the visible stimulus was presented to the same side of space as the tapping movement. For instance, if the subject tapped on the left side—with either her right or her left hand—an adaptation aftereffect was only found for visible stimuli also presented to the left. If this aftereffect were due purely to a post-perceptual decision strategy, it is difficult to see why it would be so strongly sensitive to this spatial match. Nonetheless, while the Anobile et al. study provides compelling

Still, the physiological data showing tactually produced activation of paradigmatically visual brain areas may go some way toward alleviating this concern.

The second and more serious worry enlists the distinction between causation and constitution. Two perceptual processes may be *causally* connected without using the *same representations*. To overlook this would be to commit what Adams and Aizawa (2010) call the "coupling-constitution fallacy". Suppose we grant that the same feature representations can be produced via inputs to vision or touch. This does not entail that these feature representations are constitutively multimodal. For the datum could also be explained by appeal to causal relationships among separate unimodal representations. Suppose, for instance, that whenever a unimodal visual representation of upward motion is activated, this causes activation of an accompanying tactual representation of upward motion. If so, we might expect prolonged exposure to visible upward motion to produce a tactual motion aftereffect. However, the aftereffect would not be produced by adaptation to any constitutively multimodal motion representation, but rather by adaptation to the tactual representation that is activated *alongside* vision. Thus, adaptation evidence does not definitively establish constitutive binding.

There is a second form of evidence that may be taken to support constitutive binding. To understand it, however, we'll need some background. Critically, if constitutive binding occurs, then we might expect it to produce changes in patterns of *perceptual discriminability*. Here's how this would work (see Hillis et al. (2002) or van Dam et al. (2014) for details). Suppose that the perceptual system starts out with separate unimodal estimates of an object's location. Because sensory processes are biased and noisy, these two estimates are likely to differ from one another. Now suppose that when the estimates are combined, they are discarded in favor of a single, constitutively multisensory representation of location. If so, the perceptual system should *lose access* to the initial unimodal estimates. This creates the possibility of *multimodal metamers*. If two distinct pairs of unimodal inputs elicit the *same* constitutively multisensory representation, then the two pairs should become *perceptually indistinguishable*. And this should be revealed in the subject's discrimination performance—she should lose her ability to discriminate the two metameric stimulus pairs from one another.

Prsa et al. (2012) investigated whether multimodal metamers occur following integration of visual and vestibular information about self-rotation. Subjects were seated in a rotating chair while viewing a display affixed to the chair. The display contained a dot pattern that could dynamically change to produce the visible appearance of self-rotation. Using this setup, Prsa et al. presented subjects with inconsistent combinations of vestibular and visual input. Subjects were given an "oddball" discrimination task in which they had to choose which of three successive rotations was different from the others. Prsa et al. explain what we should expect if visual and vestibular information about self-rotation are constitutively bound:

Footnote 19 continued
evidence for a visual adaptation aftereffect produced through non-visual adaptation, the precise mechanism remains unclear. Specifically, it is unclear whether the adaptation transfer was between touch and vision or between motor systems and vision. Further, the study remains susceptible to the second concern (causation vs. constitution) discussed below.

> [D]ifferent cue combinations can theoretically give rise to the same fused percept, since they would differ only in terms of information that is lost. For example, a perceived rotation size S borne out by a whole body rotation of size $S + \Delta$ paired with an equally reliable visual cue simulating a rotation of size $S - \Delta$ can be indistinguishable from a true rotation of size S produced by both stimuli (Prsa et al. 2012: p. 2282).

In their analysis, Prsa et al. compared the predictions of multiple models, but two are most critical: one model in which subjects only had access to the fused multimodal estimate, and one in which they had access to three estimates—both unimodal estimates alongside a multimodal estimate. Prsa et al. found that subjects' discrimination performance was best explained by the first model, in which discrimination was based solely on a single multimodal representation of self-rotation. This is because their discrimination performance was *worse* than would be expected under the second model. The authors conclude: "Visual and vestibular idiothetic cues are individually discarded after being fused into a single percept" (2289).[20]

The Prsa et al. (2012) findings provide compelling evidence for constitutive binding. Still, an alternate possibility should be flagged. The constitutive binding account suggests that when visual and vestibular estimates are integrated, they are obligatorily *replaced* by a single, multimodal representation of self-motion. However, another option is that the two estimates are instead *obligatorily updated* to *coincide* with one another. On the latter story, the perceptual system retains separate visual and vestibular representations of self-motion even after visual-vestibular integration, but these estimates are simply redundant. Nevertheless, while this alternative account is available, considerations of parsimony may favor the constitutive binding account. Other things being equal, we should not expect the perceptual system to maintain separate redundant estimates if these confer no computational advantages (e.g., Prsa et al. 2012: pp. 2289–2290).

3.6 Interim conclusion

It may be helpful to summarize the taxonomy of multisensory binding just laid out. I've distinguished two subtasks involved in multisensory binding. First, perception must establish cross-modal identification. Second, the features delivered through separate modalities must be integrated with information about cross-modal identification. There are two ways that cross-modal identification could take place: Separate unisensory object files could be linked, or they could be merged. If they are linked, then the referents of their singular constituents are represented as standing in the identity relation. If they are merged, then the perceptual system opens a single object file into which features delivered through both modalities are entered. Once cross-modal identification has been achieved, perceptual representations of bound features may be modified in various ways. First, it is possible that unimodal feature processing is left unaltered, but the outputs of these processes are bound to a single object or

[20] An earlier study due to Hillis et al. (2002) found that the opposite pattern held when combining visual and haptic information about size. Participants' responses indicated that they retained access to unimodal size estimates.

event. Second, it is possible that separate unimodal feature representations become coordinated, in which case the modalities bias one another—the perception of features in one modality is causally sensitive to the perception of features in another modality. Finally, it is possible that perception discards unimodal feature representations in favor of a constitutively multisensory representation—a representation that is available for processing within multiple modalities and contains no unimodal constituents.

I have been working under the assumption that perception represents objects by means of object files, and that these are the representational mechanisms responsible for multisensory binding. Certain parts of the above taxonomy are unlikely to carry over to other frameworks. In particular, I suspect that alternative views might have no room for the distinction between linking and merging models of cross-modal identification. This is because these models assume a clear separation between one object file and another: In the case of linking, two object files are maintained after cross-modal identification, while in the case of merging, only one is maintained. If these are to be genuinely distinct states of affairs, the relevant files must exist.

On the other hand, certain features of the above taxonomy are likely to apply to any analysis of the mechanisms of multisensory binding, regardless of whether object file theory is correct. I highlight two. First, if perception is capable of multisensory binding, then we can ask how binding affects the representation of features within the modalities across which binding is established. Are unimodal feature representations unaffected, or does binding enhance their coordination? Some version of the distinction between non-cooperative and cooperative binding is thus likely to transcend the object file account. Second, if separate modalities can independently represent the same feature (e.g., location or shape) prior to binding, then we can ask whether binding causes these modality-specific representations to be collapsed into one. The distinction between constitutive and non-constitutive binding, then, is likely to arise on any account of multisensory binding, regardless of whether it trades in the currency of object files.

4 Multisensory differentiation

So far, I've focused on the ability to bind features from multiple modalities to a single object or event. However, binding features to a single individual arguably requires an even more basic perceptual achievement. We must *differentiate* the individual from its surroundings in space or time. We can't perceive a tomato as jointly red and smooth unless we perceive the tomato (though, of course, we need not perceive it *as* a tomato).[21] And, plausibly, we don't perceive the tomato at all unless we differentiate it from its surroundings (e.g., Dretske 1969; Siegel 2006). When I look at a

[21] My claim is not that object differentiation is required for binding *tout court*. Rather, it is that when we bind multiple features *to an object*, we must differentiate the object to which they are bound. Certain kinds of feature binding may occur without object differentiation. Certain kinds of feature binding may contribute to object differentiation. However, I hold that in such cases the features are not bound to the object whose differentiation they facilitate. More precisely, if features F and G are bound prior to the differentiation of

uniformly white square, I don't perceive an arbitrary section consisting just of the left eighth of the square, although I do of course perceive an individual that contains this section as a part.

4.1 Differentiation

As I'll understand it here, *differentiation* includes both (i) segregation of an individual from its spatial surroundings, and (ii) tracking or reidentifying an individual over time. These processes both contribute to determining the spatiotemporal boundaries of an object—the region it carves out in spacetime. To segregate an object or event is to determine its extension in space, and to reidentify an object or event is to determine (in part) its extension over time.

Objects can be differentiated more or less determinately. A perceptual system might produce only a very coarse-grained representation of the spatial boundaries of an object, or it might be noncommittal about which of two currently perceived objects is the continuation of an object perceived earlier. Although I said above that binding features to an object requires differentiating the object, I do not claim that the object must be differentiated to a maximal degree of determinacy. Moreover, certain kinds of feature binding (e.g., binding color and shape) may require only the segregation of an object from its spatial surroundings. But other kinds, like binding an object's motion with its gradual shrinking in size, seem to require reidentification or tracking as well. One must perceive a single thing persisting from time 1 to time 2 as both moving and shrinking in the interim. Which kinds of differentiation are required for various kinds of feature binding is an interesting issue that I won't attempt to adjudicate here.

It might be objected that segregation and tracking do not comprise a natural or unified perceptual capacity. Perhaps the two abilities recruit distinct perceptual processes governed by wholly distinct principles, and should be studied independently. I regard this as an open possibility, and I don't want to prejudge the issue. However, in the context of multisensory perception there are important parallels between spatial segregation and temporal reidentification that support a unified treatment. Let me explain.

When perception parses a scene into objects, it does so in accordance with certain principles of perceptual organization. The Gestalt psychologists first systematized many of these rules a century ago (see Wagemans et al. 2012). Note that in saying that perception "accords" with the perceptual organization principles, there is no presumption that these principles are explicitly represented anywhere within perception. Perceptual systems need only undergo computational transitions that conform to them.

Remarkably, while most principles of perceptual organization were originally formulated to characterize visual object perception, similar principles have been found within auditory and tactual perception. For example, just as vision tends to group elements that are similar in brightness, size, or orientation into a single object, audition tends to group sounds that are similar in pitch, loudness, or timbre into a single sound

Footnote 21 continued
object O, then F and G are not bound to O, though they may be bound to something else. (Perhaps they are bound to a spacetime region that O partially occupies (Clark 2000), but I take no stance on the matter.).

stream (Bregman 1990).[22] And touch tends to group together elements that are similar in surface texture (Chang et al. 2007; Gallace and Spence 2011).

Importantly, these parallels between vision and audition transcend the spatial/temporal divide. The similarity principle in auditory grouping characterizes the perception of sound streams that unfold over time, while in the case of vision the principle standardly characterizes the perceptual grouping of elements at a single time. (The latter is a form of spatial segregation: segregation of an individual—the perceptual group—from its surroundings.) This suggests that certain perceptual organization strategies are used both for visual segmentation in space and for auditory tracking over time. Thus, it is not implausible for an investigation of object differentiation in the multisensory context to encompass both spatial segregation and temporal reidentification/tracking. Of course, we should remain open to the possibility that certain abilities I've included under the 'differentiation' label (e.g., purely visual tracking) really are functionally independent from the rest. This, however, is just the sort of fact that close study of multisensory interaction in object differentiation might reveal.

In what follows, I ask how and whether the modalities interact during object differentiation. This is an important issue for understanding the architecture of perception. For it concerns which perceptual processes admit multisensory input, and how early multisensory interactions can take place. Are multisensory interactions put off until after the scene has been unimodally parsed into objects and decisions have been made about which currently perceived objects are continuations of earlier ones, or do the modalities coordinate at these stages as well?

4.2 Mere convergence

One possibility is that separate modalities exhibit *no* interaction during object differentiation. Segregation in space and reidentification over time are carried out independently in the various modalities. (Of course, a mixed view is possible too. Perhaps spatial segregation is purely unimodal, while reidentification is multisensory, or vice versa.) Separate modalities may, however, *converge* on a common object or event. If this happens, then cross-modal identification and multisensory binding can take place. But cross-modal interactions, on this account, are delayed until after unimodal object differentiation is complete. Call this the *mere convergence account*. The mere convergence account is consistent with all the forms of multisensory binding discussed above.[23]

[22] All three modalities also follow a version of the rule of good continuation. In vision, we tend to group dots or line segments together when they are collinear, or lie along a smooth curve. Similar phenomena have been reported in tactual perception (Gallace and Spence 2011: p. 555). Likewise, in audition we tend to group successive sounds if they can be linked by a smooth rate of change in frequency. Suppose that tone A gradually descends in frequency before being obscured by a loud noise. If tone B begins immediately following the noise, but can be linked along a smooth frequency path with tone A, this promotes grouping the two into a single stream, which Bregman (1990) calls a "glide".

[23] Consider, for example, constitutive binding. It might be that vision and touch do not cooperate at all when differentiating objects from their surroundings, but may happen to converge on a common object. When this happens, the two unimodal object files are consolidated into a single multimodal object file, and certain unimodal feature representations within the files are replaced by constitutively multimodal feature representations (i.e., constitutive binding). In this case, constitutive binding would occur only after unimodal

Consistent with the mere convergence account, some kinds of multisensory interaction have been claimed to occur only after unimodal perceptual organization has been established. Keetels et al. (2007) investigated this issue in the case of temporal ventriloquism. Using a version of the paradigm discussed earlier (Morein-Zamir et al. 2003), Keetels et al. presented subjects with a pair of visible flashes and asked them to report which one had come first. The flashes were accompanied by two tones, which could either sandwich the flashes or occur simultaneously with them. Recall that if temporal ventriloquism takes place, then temporal order judgments should be more accurate in the former case than in the latter. Keetels et al. examined whether the temporal ventriloquism effect was sensitive to unimodal auditory grouping. The two tones were flanked by a series of tones with which they could either group or fail to group. They could be surrounded, for instance, by tones of either the same or different frequency. Critically, Keetels et al. found that temporal ventriloquism *only* occurred when the two critical tones (those that could interact with the light flashes) did *not* group with the surrounding tones. Evidently, if the tones were part of a larger sound stream then they could not temporally attract the flashes. The authors concluded that auditory temporal grouping (i.e., the differentiation of a temporally extended auditory event) occurs prior to—and indeed can *prevent*—temporal ventriloquism.[24]

These findings fit with a view on which perceptual processing remains unimodal until each modality has delivered its own parsing of the scene into objects and events. After unimodal object differentiation takes place, multisensory identity can be established, and this in turn facilitates cross-modal interactions like temporal ventriloquism and the sound-induced bouncing effect.[25]

Nevertheless, while the mere convergence account can explain some of the available data, it is highly unlikely that multisensory interactions are always put off until after unimodal object differentiation is complete. For one thing, neurons as early as primary visual cortex receive inputs from auditory and somatosensory areas (see Ghazanfar and Schroeder 2006; Murray et al. 2016 for review). These early cross-modal connections are systematic and coherent. In one fMRI study Vetter et al. (2014) found that it was possible to decode the category of a heard sound (e.g., bird sounds or traffic noises) on the basis of activity in primary visual cortex, even though the subjects were blindfolded. Other studies have shown that activity in primary visual cortex is systematically modulated by the sound-induced flash illusion, in which a single flash is illusorily perceived as two flashes due to the simultaneous presentation of two beeps (for the original illusion, see Shams et al. 2000; for neuroimaging findings, see Watkins et al. 2006). These early multisensory phenomena suggest that multisensory interaction probably does not await the conclusion of unimodal object differentiation.

Footnote 23 continued
object differentiation takes place. Thus, constitutive binding is compatible with the mere convergence account.

[24] These results confirm earlier work on the sound-induced bouncing effect: Watanabe and Shimojo (2001) found that the effect was reduced when the tone played at the moment of visible collision was placed amidst a series of tones with which it could be grouped.

[25] But *why* should auditory grouping prevent the temporal ventriloquism effect? Here is a possible story: When a tone is grouped into a larger sound stream, it is no longer perceived as an individual audible event. Rather, it is treated as an undifferentiated temporal part of a larger audible event (the whole stream). Accordingly, it is no longer a candidate for identity with the relevant visible event (e.g., a flash or collision). And because multisensory identity can't be established, multisensory coordination does not take place.

However, a weaker version of the mere convergence account may survive this type of evidence. One might concede that interactions among modalities occur at the earliest levels of perceptual processing, but suggest that there are wholly unimodal processes within perception that operate in parallel with certain cross-modal interactions. And it remains possible that segregation and tracking mechanisms are unimodal processes of this sort. They access information from just one modality and perform algorithms proprietary to that modality. Early cross-modal connections do not refute this view. We need to determine the *function* of these connections—specifically, whether they contribute to object differentiation.

4.3 Multisensory differentiation: an empirical signature

To decide whether the modalities ever coordinate during object differentiation, we need to determine the *empirical signatures* of such coordination. What would we expect to observe if the modalities do work together during object differentiation?

I suggest the following as an empirically sufficient condition for multisensory differentiation. The modalities must coordinate during object differentiation if it is possible to perceptually differentiate an individual that could *not* be differentiated through any single modality operating on its own.[26] Recall that differentiation includes both the segregation of an individual from its spatial surroundings, and the tracking or reidentification of an individual over time. Either of these abilities might involve multisensory interaction.

As regards segregation in space, it might be discovered that the perceived spatial boundaries of an object can be determined only by combining information across the senses. For example, suppose an object is partially occluded. One half can be seen, while the occluded half can only be felt. Perception might combine information from vision and touch to determine how the object completes behind the occluder. As regards reidentification over time, it might be that the temporal persistence of an object, or the temporal unfolding of an event, can be determined only by combining information across the senses. Suppose that you see a dog pass behind a wall and then hear a series of barks while it is occluded. These barks might be used to reidentify the dog and to update the perceptual representation of its location. This reidentification would not be possible on the basis of vision or audition alone.

Two clarifications before continuing. First, the claim that perception exhibits multisensory differentiation is significantly stronger than the claim that certain objects are perceptible through more than one modality. The latter claim is compatible with the mere convergence account. Specifically, it is compatible with the view that the processes responsible for differentiating objects and events from their spatiotemporal surroundings are wholly unimodal, but just happen to pick out the same objects in certain cases.

[26] Charles Spence seems to have a similar idea in mind with the notion of *intersensory gestalten* (Spence 2015; Spence et al. 2007). Spence characterizes intersensory gestalten as a "multisensory (or supramodal) organization (or structure), which, importantly, isn't present in any of the constituent sensory modalities when considered individually" (2015: p. 646).

Second, the notion of multisensory differentiation has an important analogue in the case of feature perception. O'Callaghan (2017) and Briscoe (2019) have recently argued for the existence of *novel multisensory features*. These are features that can only be perceived through multiple modalities operating in concert. For example, O'Callaghan cites flavor features (see also Macpherson 2011: pp. 449–450), whose perception seems to require the interaction of gustatory, olfactory, and somatosensory perception. Briscoe mentions the perception of location in egocentric space, which involves integration of information from multiple senses (e.g., vision, touch, and proprioception) to produce a representation of location in a non-modality-specific body-centered reference frame.

The analogy between multisensory differentiation and novel multisensory features is deliberate. However, the two issues should be examined separately. For even if there are novel multisensory features, this does not settle whether there is multisensory differentiation. Suppose, for example, that Briscoe is right that perceived egocentric location constitutes a novel multisensory feature. It could still be that the objects assigned locations in one's body-centered reference frame are always differentiated either through vision alone, audition alone, etc. This would be a case of novel multisensory features without multisensory differentiation. More generally, even if there are novel multisensory features, it is possible that the objects to which these features are attributed are differentiated unimodally.

4.4 Evaluating the evidence for multisensory differentiation

I now consider the evidence for multisensory differentiation. Multisensory differentiation occurs if the modalities cooperate when segregating an object in space or reidentifying an object over time. However, because most of the evidence that I'm aware of concerns the latter sort of interaction, this is where I will focus.

One attempt to demonstrate multisensory reidentification between vision and audition was arguably unsuccessful. Huddleston et al. (2008) presented subjects with a pair of illuminable LEDs along a vertical axis together with a pair of speakers along a horizontal axis, forming a circle. Together, these generated a series of light flashes and white noise bursts (e.g., light on top, noise to the right, light on bottom, noise to the left). The subject's task was to report whether the 'motion' produced by this audiovisual series was clockwise or counterclockwise. Note that if subjects perceive motion from the LEDs to the speakers, this suggests that they were able to reidentify an individual over time in a context where no single modality possessed adequate information to enable the reidentification.[27] However, while Huddleston et al. found that subjects could accurately judge the *order* in which the events occurred, the subjects did not report spontaneous percepts of *motion*. Huddleston et al. write:

[27] A caveat: The perception of motion needn't involve the perception of an *object* in motion, as the *phi* phenomenon illustrates. Motion might be perceived as a free-floating feature, or it could be placed at a spacetime region (e.g., "leftward motion thereabouts") rather than attributed to an object. Thus, even if the modalities interact in motion perception, this does not immediately show that they interact when reidentifying objects over time. Still, if subjects in the Huddleston et al. setup can perceive motion between the LEDs and the speakers, this at least would provide suggestive evidence for multisensory differentiation.

> Surprisingly, none of the subjects had an integrated percept of rotational motion. (…) Rather, all subjects reported a percept of lights moving from one LED location to the other and of sounds moving from one speaker to the other independently in each modality, even though they were sequentially presented in alternate modalities over time. (1212)

O'Callaghan (2017) argues that Huddleston et al. did not perform an optimal test for audio-visual apparent motion because white noise bursts and light flashes may not have enough in common to promote the percept of a persisting individual. Expanding on this line of thought, it would be interesting to know whether audible and visible stimuli that display the sorts of synesthetic correspondences known to promote percepts of audio-visual event identity are more easily linked via apparent motion (see Sect. 3.1). A further question is whether audio-visual apparent motion is more likely in an ecologically valid context in which the sound originates from behind a visible occluder (recall the dog-barking example from above). Nevertheless, it is at best an open question whether genuine audio-visual apparent motion is possible.

Studies of the visual-tactile case have met with more promising results. Harrar et al. (2008) had subjects sit at a table and place their index fingers inside small cup-shaped stimulators. These could emit a small probe into the index finger that felt like a gentle tap. Illuminable LEDs were mounted on top of the tactile stimulators. This set-up allowed Harrar et al. to compare characteristics of visual–visual, tactile–tactile, and visual-tactile apparent motion as a function of spatial distance and inter-stimulus interval (ISI: the time between successive lights or taps). They found that, at longer ISIs, the rated subjective quality of visual-tactile apparent motion did not differ significantly from visual–visual or tactile–tactile motion. Nevertheless, Harrar et al. found an important difference between unimodal and multimodal apparent motion. Unimodal apparent motion is known to conform to Korte's Law: With greater distances, the ISI for optimal apparent motion increases. Harrar et al. confirmed that visual–visual and tactile–tactile apparent motion followed this pattern.[28] But visual-tactile motion did not. Although the rated quality of visual-tactile motion was sensitive to ISI, changes in the distance between the subject's index fingers did not affect the optimal ISI (see also Harrar and Harris 2007).

Spence (2015) suggests that these results may be due to response bias. The fact that subjects *reported* perceiving visual-tactile apparent motion doesn't mean that they really perceived it. Indeed, this might explain why visual-tactile apparent motion didn't conform to Korte's Law: Subjects simply were not sensitive to the distance between their index fingers when deciding how to respond, although they were able to take ISI into account. On this view, the percept of the visual-tactile stimulus simply represented a succession of lights and flashes, with no linking motion.

However, subsequent findings cast doubt on the response bias interpretation. It is known that apparent motion within one modality can influence—or *capture*—apparent motion in another. Thus, suppose that a subject perceives an auditory apparent motion sequence that objectively moves from left to right. If she simultaneously perceives a visual apparent motion sequence that moves right-to-left, then she is more likely to inaccurately hear the auditory sequence move right-to-left as well (Soto-Faraco et al.

[28] However, in the tactile-tactile case, the relationship broke down for distances greater than 10 cm.

2004). Given this, Jiang and Chen (2013) investigated whether visual-tactile apparent motion [tested with essentially the same set-up as Harrar et al. (2008)] could exert capture effects on unimodal auditory apparent motion. They found that this was indeed the case. Of particular interest, the strength of the effect was *intermediate* between the capture effects exerted by tactile–tactile apparent motion and visual–visual apparent motion. Thus, visual-tactile apparent motion exerts a *stronger* effect on auditory motion than unimodal tactile motion does.

Note, moreover, that capture effects of this sort have typically been taken as evidence that a process belongs on the perception side of the perception/cognition divide (e.g., Scholl and Nakayama 2004). The reason is simply that, if a process exerts capture effects on other genuinely perceptual processes, then it is plausibly functionally integrated with them. And if a process is functionally integrated with other perceptual processes, then we should conclude, absent strong reasons to think otherwise, that it is a perceptual process as well.

Thus, I think there is a strong case that Harrar et al. (2008) have uncovered a genuine perceptual phenomenon, contra Spence (2015). But is it an instance of *multisensory object differentiation*? If so, here is what should happen: When an observer sees a light flash followed by a tap, she perceives this as a single individual moving from the location of the flash to the location of the tap. However, I am not convinced that *this* is true. Harrar et al. note that, according to their participants, multisensory apparent motion seems "more causal" than other kinds of apparent motion:

> [N]o subjects reported the sensation of a single light moving to or from the location of the touch in the multimodal condition. Instead, subjects in the visuotactile condition reported perceiving some type of multimodal apparent motion, but they often described it as being 'more causal' than the unimodal apparent motion. Our participants mainly interpreted their perception like a switch flicking on a light or like a cannon firing that was felt on one hand and then the flash from the landing explosive was seen on the other hand. (810)

This suggests an alternative interpretation. Subjects might have experienced a tactual event as *causing* a visible event, or vice versa. And perhaps the perceived direction of causality can exert capture effects on auditory motion. This interpretation may also help us understand why visual-tactile 'motion' doesn't follow Korte's Law. The perceptual system assumes that, the farther apart two locations are, the longer an object will take to move between them. But perhaps this constraint is absent—or at least relaxed—in the case of causal relations. We can perceive a flash as the cause of a tap, or vice versa, but our tendency to perceive this is less sensitive to the distance between the flash and the tap. Thus, while Harrar et al. (2008) have plausibly revealed an instance of genuinely perceptual interaction between vision and touch, it is not clear that this interaction qualifies as multisensory object differentiation. It is possible that vision and touch differentiated their objects independently, but the visible and tactual objects were then represented as standing in a causal relation.

The final case I'll consider involves interaction between audition and touch. Many musicians and dancers report that the experience of musical rhythm is richly multisensory. The "beat" is not merely heard—it is *felt*. This creates the potential for

multisensory differentiation. Can heard and felt events be grouped to form a multisensory rhythm with identifiable meter properties? Suppose that a perceiver encounters a musical rhythm composed of audible sounds together with tactual vibrations or taps. If these are perceived as a single extended stream, then the only way this could happen is by perceptually grouping information from audition and touch. We can construe such streams as complex events containing sounds and vibrations as parts (e.g., O'Callaghan 2016; Green 2018).

Huang et al. (2012) investigated this question. Subjects perceived a series of events consisting of sounds delivered through headphones together with vibrations delivered to the left index finger. Their task was to report whether the series had a 'march-like' or 'waltz-like' meter. Meter was marked by selectively "accenting" certain events in the series—increasing the intensity of the sound or tactual pulse. Huang et al. examined performance both in unimodal conditions and in bimodal conditions, allowing them to determine whether bimodal discrimination performance exceeded the levels that could be achieved through a single modality on its own. They found that this was the case. Participants could successfully discriminate march-like from waltz-like meters in conditions where the inputs provided to either modality taken individually were wholly ambiguous. The authors conclude: "The results demonstrate, we believe, for the first time that auditory and tactile input are grouped during meter perception" (10).

These results are compelling.[29] But do they decisively establish multisensory object differentiation? I think not, and the reason is straightforward. We cannot be sure that the subjects' apprehension of meter in the multisensory condition was genuinely *perceptual*. The fact that they could discriminate waltz-like from march-like audio-tactile meters doesn't imply that the discrimination was done within perception. Recall our discussion of Huddleston et al. (2008) above. Subjects in that study reported that they could figure out whether the implied direction of audio-visual motion was clockwise or counterclockwise, but they didn't *perceive* audio-visual motion. A similar issue arises here. Subjects might cognitively integrate information delivered through audition and touch to determine meter properties even if they don't perceive such properties.

What could settle this issue? I'll mention three potential findings that would furnish a more secure case for multisensory meter perception. First, we might find that there is *spatiotopic adaptation* to multisensory meter. For instance, we might find that after extended exposure to a march-like meter, subjects are more likely to perceive an ambiguous meter as waltz-like, and that the adaptation is specific to a particular region of space. Of course, any argument from such data would require the premise that retinotopic and spatiotopic adaptation are markers of perceptual processing, and so would only be as convincing as this premise (see Block 2014, forthcoming).

Second, we might find that multisensory meter is processed in *paradigmatically perceptual brain areas*. In this vein, Araneda et al. (2017) recently found that certain areas within the auditory dorsal stream (including the inferior parietal lobe and superior temporal gyrus) are selectively activated in response to rhythmic sequences regardless

[29] Indeed, despite expressing general skepticism about multisensory perceptual grouping, Spence (2015) concedes that "Huang et al.'s results can therefore be taken as providing support for the claim that audio-tactile musical meter perception constitutes one of the first genuinely intersensory Gestalten to have been documented to date" (647).

of whether the sequences are presented auditorily, tactually, or visually. However, it is unclear whether this result would carry over to a multimodal rhythm of the sort examined by Huang et al. (2012).

Third, we might find that multisensory meter influences *paradigmatically perceptual processes* (recall the discussion of capture effects above). For instance, it might be found that the perception of temporal intervals is affected by whether those intervals are part of an ongoing rhythm. If the same result were found in a multisensory context, this would be strong evidence that multisensory rhythm and meter are recovered in perception.

To sum up: Multisensory differentiation is important to the basic architecture of perceptual systems. Which perceptual processes allow cross-modal interaction, and which are limited to information from a single modality? And, in particular, is object differentiation encapsulated from the information stored in other modalities? I've argued that current evidence provides a suggestive but inconclusive case for multisensory differentiation. Alternative interpretations are left open. I've suggested some ways the issue might be resolved.

5 Conclusion: multisensory constraints on architecture and format

This paper has distinguished and analyzed two forms of multisensory object perception: multisensory *binding* and multisensory *differentiation*. Separate modalities may bind features to the same object, or they may cooperate in differentiating an object from its surroundings in space or time. I'll conclude by highlighting some ways that multisensory object perception may inform our understanding of the architecture and format of perception.

It is often noted that multisensory interactions refute a view of the sense modalities as wholly independent, encapsulated modules. This view was never very attractive, but it is certainly dead now (see Driver and Spence 2000).[30] However, there are still important questions about the architecture of perception that may be framed using the modular framework. Specifically, even if the modalities taken as a whole are unencapsulated from one another, it remains possible that particular perceptual processes within a modality are encapsulated from information held in other modalities. One such issue, I've urged, is whether object differentiation within a modality is encapsulated from the information held in other modalities. This would be perfectly consistent with holding that other perceptual processes—perhaps even some that unfold prior to or alongside object differentiation—access information from multiple modalities. And it is consistent with the view that if two modalities happen to converge on the same object, features from both modalities can be bound to it.

Multisensory object perception is also important for understanding the format of perceptual representations. If certain object or feature representations are *shared* across modalities (see, for instance, the merging model of cross-modal identification or the constitutive model of feature binding), then they must have a format suitable for

[30] Even though the perceptual modalities are unencapsulated with respect to one another, perception as a whole may still be encapsulated from cognition (see Firestone and Scholl 2016). I haven't said anything about the latter issue.

composing with representations from multiple modalities, and they must be able to participate in the algorithms of more than one modality. Perceptual object representations are thus not like ordinary pictures, which only depict visible properties. Of course, this doesn't settle whether they are iconic or depictive in some more abstract sense—that is a larger issue. However, it is important that debates about the structure and format of perceptual object representations not overlook multisensory perception, and the common tendency toward overly vision-centric models needs to be resisted.[31]

References

Adams, F., & Aizawa, K. (2010). Defending the bounds of cognition. In R. Menary (Ed.), *The extended mind* (pp. 67–80). Cambridge, MA: MIT Press.

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology, 14*(3), 257–262.

Anobile, G., Arrighi, R., Togoli, I., & Burr, D. C. (2016). A shared numerical representation for action and perception. *eLife, 5,* e16161.

Araneda, R., Renier, L., Ebner-Karestinos, D., Dricot, L., & De Volder, A. G. (2017). Hearing, feeling or seeing a beat recruits a supramodal network in the auditory dorsal stream. *European Journal of Neuroscience, 45*(11), 1439–1450.

Bahrami, B. (2003). Object property encoding and change blindness in multiple object tracking. *Visual Cognition, 10*(8), 949–963.

Batty, C. (2014). Olfactory objects. In D. Stokes, M. Matthen, & S. Biggs (Eds.), *Perception and its modalities* (pp. 222–246). Oxford: Oxford University Press.

Beauchamp, M. S., Yasar, N. E., Kishan, N., & Ro, T. (2007). Human MST but not MT responds to tactile stimulation. *The Journal of Neuroscience, 27*(31), 8261–8267.

Berger, C. C., & Ehrsson, H. H. (2016). Auditory motion elicits a visual motion aftereffect. *Frontiers in Neuroscience, 10,* 559.

Bertelson, P., & Radeau, M. (1981). Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. *Perception and Psychophysics, 29*(6), 578–584.

Block, N. (2014). Seeing-as in the light of vision science. *Philosophy and Phenomenological Research, 89*(3), 560–572.

Block, N. (forthcoming). *The border between seeing and thinking*.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: MIT Press.

Briscoe, R. E. (2016). Multisensory processing and perceptual consciousness: Part I. *Philosophy Compass, 11*(2), 121–133.

Briscoe, R. E. (2017). Multisensory processing and perceptual consciousness: Part II. *Philosophy Compass, 12*(12), e12423.

Briscoe, R. E. (2019). Bodily awareness and novel multisensory features. *Synthese*. https://doi.org/10.1007/s11229-019-02156-2.

Carey, S. (2009). *The origin of concepts*. Oxford: Oxford University Press.

Carvalho, F. (2014). Olfactory objects. *Disputatio, 6*(38), 45–66.

Casati, R., & Dokic, J. (2009). Some varieties of spatial hearing. In M. Nudds & C. O'Callaghan (Eds.), *Sounds and perception: New philosophical essays* (pp. 97–110). Oxford: Oxford University Press.

Chang, D., Nesbitt, K. V., & Wilkins, K. (2007). The Gestalt principles of similarity and proximity apply to both the haptic and visual grouping of elements. In W. Piekarski & B. Plimmer (Eds.), *Conferences in research and practice in information technology* (Vol. 64, pp. 79–86). Ballarat, Australia: The Australian Computer Society.

[31] Thanks to Alex Byrne, Jake Quilty-Dunn, Tyler Wilson, and two anonymous reviewers for helpful comments on earlier drafts of this paper. Thanks also to audiences at the 2018 Winter Perception Workshop at UCSD and at Oberlin College, where portions of this material were presented.

Chuen, L., & Schutz, M. (2016). The unity assumption facilitates cross-modal binding of musical, non-speech stimuli: The role of spectral and amplitude envelope cues. *Attention, Perception, & Psychophysics, 78*(5), 1512–1528.

Clark, A. (2000). *A theory of sentience*. Oxford: Oxford University Press.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24,* 87–185.

de Vignemont, F. (2014). Multimodal unity and multimodal binding. In D. J. Bennett & C. Hill (Eds.), *Sensory integration and the unity of consciousness* (pp. 125–150). Cambridge, MA: MIT Press.

Deroy, O. (2014). The unity assumption and the many unities of consciousness. In D. J. Bennett & C. Hill (Eds.), *Sensory integration and the unity of consciousness* (pp. 105–124). Cambridge, MA: MIT Press.

Dretske, F. I. (1969). *Seeing and knowing*. Chicago, IL: University of Chicago Press.

Driver, J., & Spence, C. (2000). Multisensory perception: Beyond modularity and convergence. *Current Biology, 10*(20), R731–R735.

Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch. *Journal of Vision, 7*(5), 1–14.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415,* 429–433.

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences, 39,* 1–19.

Flombaum, J. I., & Scholl, B. J. (2006). A temporal same-object advantage in the tunnel effect: Facilitated change detection for persisting objects. *Journal of Experimental Psychology: Human Perception and Performance, 32*(4), 840.

Fougnie, D., Asplund, C. L., & Marois, R. (2010). What are the units of storage in visual working memory? *Journal of Vision, 10*(12), 27–27.

Fougnie, D., Cockhren, J., & Marois, R. (2018). A common source of attention for auditory and visual tracking. *Attention, Perception, & Psychophysics, 80*(6), 1571–1583.

Fougnie, D., & Marois, R. (2009). Attentive tracking disrupts feature binding in visual working memory. *Visual Cognition, 17*(1–2), 48–66.

Fulkerson, M. (2011). The unity of haptic touch. *Philosophical Psychology, 24*(4), 493–516.

Gallace, A., & Spence, C. (2011). To what extent do Gestalt grouping principles influence tactile perception? *Psychological Bulletin, 137*(4), 538–561.

Ghazanfar, A. A., & Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends in Cognitive Sciences, 10*(6), 278–285.

Gordon, R. D., & Irwin, D. E. (2000). The role of physical and conceptual properties in preserving object continuity. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 26*(1), 136–150.

Grassi, M., & Casco, C. (2009). Audiovisual bounce-inducing effect: Attention alone does not explain why the discs are bouncing. *Journal of Experimental Psychology: Human Perception and Performance, 35*(1), 235–243.

Green, M. (1986). What determines correspondence strength in apparent motion? *Vision Research, 26*(4), 599–607.

Green, E. J. (2018). A theory of perceptual objects. *Philosophy and Phenomenological Research*. https://doi.org/10.1111/phpr.12521.

Green, E. J., & Quilty-Dunn, J. (2017). What is an object file? *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axx055.

Harrar, V., & Harris, L. R. (2007). Multimodal Ternus: Visual, tactile, and visuo-tactile grouping in apparent motion. *Perception, 36*(10), 1455–1464.

Harrar, V., Winter, R., & Harris, L. R. (2008). Visuotactile apparent motion. *Perception and Psychophysics, 70*(5), 807–817.

Hillis, J. M., Ernst, M. O., Banks, M. S., & Landy, M. S. (2002). Combining sensory information: Mandatory fusion within, but not between, senses. *Science, 298*(5598), 1627–1630.

Hollingworth, A., & Rasmussen, I. P. (2010). Binding objects to locations: The relationship between object files and visual working memory. *Journal of Experimental Psychology: Human Perception and Performance, 36*(3), 543–564.

Hollingworth, A., Richard, A. M., & Luck, S. J. (2008). Understanding the function of visual short-term memory: Transsaccadic memory, object correspondence, and gaze correction. *Journal of Experimental Psychology: General, 137*(1), 163–181.

Hommel, B. (2004). Event files: Feature binding in and across perception and action. *Trends in Cognitive Sciences, 8*(11), 494–500.
Huang, J., Gamble, D., Sarnlertsophon, K., Wang, X., & Hsiao, S. (2012). Feeling music: Integration of auditory and tactile inputs in musical meter perception. *PLoS ONE, 7*(10), e48496.
Huddleston, W. E., Lewis, J. W., Phinney, R. E., & DeYoe, E. A. (2008). Auditory and visual attention-based apparent motion share functional parallels. *Perception and Psychophysics, 70*(7), 1207–1216.
Jain, A., Sally, S. L., & Papathomas, T. V. (2008). Audiovisual short-term influences and aftereffects in motion: Examination across three sets of directional pairings. *Journal of Vision, 8*(15), 7:1–13.
Jiang, Y., & Chen, L. (2013). Mutual influences of intermodal visual/tactile apparent motion and auditory motion with uncrossed and crossed arms. *Multisensory Research, 26*(1–2), 19–51.
Jordan, K. E., Clark, K., & Mitroff, S. R. (2010). See an object, hear an object file: Object correspondence transcends sensory modality. *Visual Cognition, 18*(4), 492–503.
Jousmäki, V., & Hari, R. (1998). Parchment-skin illusion: Sound-biased touch. *Current Biology, 8*(6), R190.
Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology, 24,* 175–219.
Keetels, M., Stekelenburg, J., & Vroomen, J. (2007). Auditory grouping occurs prior to intersensory pairing: Evidence from temporal ventriloquism. *Experimental Brain Research, 180*(3), 449–456.
Konkle, T., & Moore, C. I. (2009). What can crossmodal aftereffects reveal about neural representation and dynamics? *Communicative & Integrative Biology, 2*(6), 479–481.
Konkle, T., Wang, Q., Hayward, V., & Moore, C. I. (2009). Motion aftereffects transfer between touch and vision. *Current Biology, 19*(9), 745–750.
Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLoS ONE, 2*(9), e943.
Levitan, C. A., Ban, Y. H. A., Stiles, N. R., & Shimojo, S. (2015). Rate perception adapts across the senses: Evidence for a unified timing mechanism. *Scientific Reports, 5,* 8857.
Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*(6657), 279–281.
Macpherson, F. (2011). Cross-modal experiences. *Proceedings of the Aristotelian Society, 111*(3), 429–468.
Millar, B. (2017). Smelling objects. *Synthese*. https://doi.org/10.1007/s11229-017-1657-8.
Mitroff, S. R., & Alvarez, G. A. (2007). Space and time, not surface features, guide object persistence. *Psychonomic Bulletin & Review, 14*(6), 1199–1204.
Morein-Zamir, S., Soto-Faraco, S., & Kingstone, A. (2003). Auditory capture of vision: Examining temporal ventriloquism. *Cognitive Brain Research, 17*(1), 154–163.
Murray, M. M., Thelen, A., Thut, G., Romei, V., Martuzzi, R., & Matusz, P. J. (2016). The multisensory function of the human primary visual cortex. *Neuropsychologia, 83,* 161–169.
Noles, N. S., Scholl, B. J., & Mitroff, S. R. (2005). The persistence of object file representations. *Perception and Psychophysics, 67*(2), 324–334.
Nudds, M. (2009). Sounds and space. In M. Nudds & C. O'Callaghan (Eds.), *Sounds and perception: New philosophical essays* (pp. 69–96). Oxford: Oxford University Press.
Nudds, M. (2014). Is audio-visual perception 'amodal' or 'cross-modal'? In D. Stokes, M. Matthen, & S. Biggs (Eds.), *Perception and its modalities*. Oxford: Oxford University Press.
O'Callaghan, C. (2008). Seeing what you hear: Cross-modal illusions and perception. *Philosophical Issues, 18*(1), 316–338.
O'Callaghan, C. (2014). Intermodal binding awareness. In D. Bennett & C. Hill (Eds.), *Sensory integration and the unity of consciousness* (pp. 73–103). Cambridge, MA: MIT Press.
O'Callaghan, C. (2015). The multisensory character of perception. *The Journal of Philosophy, 112*(10), 551–569.
O'Callaghan, C. (2016). Objects for multisensory perception. *Philosophical Studies, 173,* 1269–1289.
O'Callaghan, C. (2017). Grades of multisensory awareness. *Mind and Language, 32*(2), 155–181.
Parise, C., & Spence, C. (2008). Synesthetic congruency modulates the temporal ventriloquism effect. *Neuroscience Letters, 442*(3), 257–261.
Parise, C. V., & Spence, C. (2009). 'When birds of a feather flock together': Synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PLoS ONE, 4*(5), e5664.
Parise, C. V., Harrar, V., Ernst, M. O., & Spence, C. (2013). Cross-correlation between auditory and visual signals promotes multisensory integration. *Multisensory Research*, *26*, 307–316.
Pick, H. L., Jr., Warren, D. H., & Hay, J. C. (1969). Sensory conflict in judgments of spatial direction. *Perception and Psychophysics, 6,* 203–205.

Prsa, M., Gale, S., & Blanke, O. (2012). Self-motion leads to mandatory cue fusion across sensory modalities. *Journal of Neurophysiology, 108*(8), 2282–2291.
Pylyshyn, Z. W. (2007). *Things and places: How the mind connects with the world*. Cambridge, MA: MIT Press.
Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision, 3*(3), 179–197.
Quilty-Dunn, J. (2016). Iconicity and the format of perception. *Journal of Consciousness Studies, 23*(3–4), 255–263.
Recanati, F. (2012). *Mental files*. Oxford: Oxford University Press.
Scholl, B. J., & Nakayama, K. (2004). Illusory causal crescents: Misperceived spatial relations due to perceived causality. *Perception, 33*(4), 455–469.
Schutz, M., & Kubovy, M. (2009). Causality and cross-modal integration. *Journal of Experimental Psychology: Human Perception and Performance, 35*(6), 1791–1810.
Schutz, M., & Lipscomb, S. (2007). Hearing gestures, seeing music: Vision influences perceived tone duration. *Perception, 36*(6), 888–897.
Sekuler, R., Sekuler, A. B., & Lau, R. (1997). Sound alters visual motion perception. *Nature, 385,* 308.
Shams, L., Kamitani, Y., & Shimojo, S. (2000). Illusions: What you see is what you hear. *Nature, 408*(6814), 788.
Shams, L., & Kim, R. (2010). Crossmodal influences on visual perception. *Physics of Life Reviews, 7*(3), 269–284.
Siegel, S. (2006). How does visual phenomenology constrain object-seeing? *Australasian Journal of Philosophy, 84*(3), 429–441.
Soto-Faraco, S., Spence, C., & Kingstone, A. (2004). Cross-modal dynamic capture: Congruency effects in the perception of motion across sensory modalities. *Journal of Experimental Psychology: Human Perception and Performance, 30*(2), 330–345.
Spapé, M. M., & Hommel, B. (2010). Actions travel with their objects: Evidence for dynamic event files. *Psychological Research PRPF, 74*(1), 50–58.
Spence, C. (2015). Cross-modal perceptual organization. In J. Wagemans (Ed.), *The Oxford handbook of perceptual organization* (pp. 639–654). Oxford: Oxford University Press.
Spence, C., & Bayne, T. (2014). Is consciousness multisensory? In D. Stokes, M. Matthen, & S. Biggs (Eds.), *Perception and its modalities* (pp. 95–132). Oxford: Oxford University Press.
Spence, C., Sanabria, D., & Soto-Faraco, S. (2007). Intersensory Gestalten and crossmodal scene perception. In K. Noguchi (Ed.), *Psychology of beauty and Kansei: New horizons of gestalt perception* (pp. 519–579). Fuzanbo International: Tokyo.
Storrs, K. R. (2015). Are high-level aftereffects perceptual? *Frontiers in Psychology, 6,* 157.
Treisman, A. (1988). Features and objects: The fourteenth Bartlett memorial lecture. *The Quarterly Journal of Experimental Psychology, 40A*(2), 201–237.
Tuomainen, J., Andersen, T. S., Tiippana, K., & Sams, M. (2005). Audio–visual speech perception is special. *Cognition, 96*(1), B13–B22.
van Dam, L. C. J., Parise, C. V., & Ernst, M. O. (2014). Modeling multisensory integration. In D. J. Bennett & C. Hill (Eds.), *Sensory integration and the unity of consciousness* (pp. 209–229). Cambridge, MA: MIT Press.
Van der Burg, E., Awh, E., & Olivers, C. N. (2013). The capacity of audiovisual integration is limited to one item. *Psychological Science, 24*(3), 345–351.
VanMarle, K., & Scholl, B. J. (2003). Attentive tracking of objects versus substances. *Psychological Science, 14*(5), 498–504.
Vatakis, A., & Spence, C. (2007). Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli. *Perception and Psychophysics, 69*(5), 744–756.
Vatakis, A., & Spence, C. (2008). Evaluating the influence of the 'unity assumption' on the temporal perception of realistic audiovisual stimuli. *Acta Psychologica, 127*(1), 12–23.
Vetter, P., Smith, F. W., & Muckli, L. (2014). Decoding sound and imagery content in early visual cortex. *Current Biology, 24*(11), 1256–1262.
Vroomen, J., & Stekelenburg, J. J. (2011). Perception of intersensory synchrony in audiovisual speech: Not that special. *Cognition, 118*(1), 75–83.
Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., et al. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. *Psychological Bulletin, 138,* 1172–1217.

Watanabe, K., & Shimojo, S. (2001). When sound affects vision: effects of auditory grouping on visual motion perception. *Psychological Science, 12*(2), 109–116.
Watkins, S., Shams, L., Tanaka, S., Haynes, J. D., & Rees, G. (2006). Sound alters activity in human V1 in association with illusory visual perception. *Neuroimage, 31*(3), 1247–1256.
Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88*(3), 638–667.
Woodman, G. F., & Vogel, E. K. (2008). Selective storage and maintenance of an object's features in visual working memory. *Psychonomic Bulletin & Review, 15*(1), 223–229.
Zhou, K., Luo, H., Zhou, T., Zhuo, Y., & Chen, L. (2010). Topological change disturbs object continuity in attentive tracking. *Proceedings of the National Academy of Sciences, 107*(50), 21920–21924.
Zmigrod, S., & Hommel, B. (2009). Auditory event files: Integrating auditory perception and action planning. *Attention, Perception, & Psychophysics, 71*(2), 352–362.
Zmigrod, S., & Hommel, B. (2010). Temporal dynamics of unimodal and multimodal feature binding. *Attention, Perception, & Psychophysics, 72*(1), 142–152.
Zmigrod, S., Spapé, M., & Hommel, B. (2009). Intermodal event files: Integrating features across vision, audition, taction, and action. *Psychological Research PRPF, 73*(5), 674–684.