Multimodal binding as mereological co-constituency

Jonathan Cohen*

> Inflation is as violent as a mugger, as frightening as an armed robber and as deadly as a hit man.
>
> —Ronald Reagan, 1978

1 Introduction

Perceptual features are said to be bound when they are bundled together and distinguished from other features. Thus, in a canonical (unimodal, visual) example, if the stimulus contains a green triangle and a red square, ordinary visual perception organizes the *greenness* and *triangularity* together into one bundle, organizes the *redness* and *squareness* together into a second bundle, and treats the two distinct bundles as separate wholes. We know this binding occurs because, without it, visual systems would be unable to distinguish (as they obviously can) the configuration just described (*greenness* bundled with *triangularity*, *redness* bundled with *squareness*) from the distinct configuration consisting of a green square and a red triangle (*greenness* bundled with *squareness*, *redness* bundled with *triangularity*).[1]

On the dominant theoretical description (henceforth, “the convergence conception”), binding amounts to the representation of features as convergently exemplified by one and the same object—a so-called sensory individual. This understanding of binding relies on (and provides important support for) the idea that perception does not only register distal properties instances, but (in some modalities, at least some of the time) connects with and attributes properties to individual entities.[2] These sensory individuals serve as points of

*Department of Philosophy, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0119, cohen@ucsd.edu

[1] This is an instance of Jackson's 1977 Many Properties Problem.

[2] Note that this thesis, as stated, says very little about the nature of the individual entities involved; in particular, it should not be understood as committing to the perceptibility of bare particulars.

contact between mind and world—distal individual *res* that perception connects us to, and that anchor our *de re* cognitive contact with things outside our own heads. Moreover, they are the loci of perceptual attribution, serving as targets for focal attention and/or perceptual demonstration, and underpinning *de re* reference in thought and language. Importantly, sensory individuals' role as the loci for perceptual attribution grounds their role in explaining perceptual binding: the conception of binding as convergent attribution of features to one sensory individual doesn't get off the ground without the assumption that perception attributes its features to such individuals.

The convergence conception is, as noted, the canonical model of unimodal binding. It is endorsed, typically explicitly, by such disparate figures as Treisman (1998), Clark (2000), Shams et al. (2000), Campbell (2002), Treisman (2003), Cohen (2004), Matthen (2005), Pylyshyn (2007), and Dickie (2010). It is fair to say that the convergence conception of binding has proven itself an important and fruitful contribution to the contemporary understanding of unimodal perception.

However, research in recent decades has moved beyond unimodal cases to focus on a broad class of multimodal cases in which features from distinct perceptual modalities are bundled in what appears to be something like the manner seen in unimodal binding. Thus, visual and auditory features appear to be bundled in such much-discussed cases as the McGurk effect (McGurk and MacDonald 1976), the ventriloquist illusion (Pick et al. 1969; Vroomen and de Gelder 2000, 2004), the motion-bounce effect (Sekuler et al. 1997), and the sound-induced flash illusion (Shams et al. 2000, 2002). The cutaneous rabbit illusion (Geldard and Sherrick 1972) and rubber hand illusion (Botvinick and Cohen 1998) provide evidence of tactile/visual bundling. Similarly, there is evidence of bundling between olfaction and vision (Kuang and Zhang 2014), olfaction and gustation (Dalton et al. 2000; Spence et al. 2014), vision and gustation (Morrot et al. 2001), audition and gustation (Spence 2015), and more. In view of these findings, it is natural to wonder how and whether to extend theoretical ideas about perceptual binding, first developed with respect to unimodal cases, to multimodal perception.[3]

In what follows I will survey a range of answers to this question by working through what we might think of as (with apologies to Quine) three grades of multimodal involvement. The first grade is convergence; I'll argue that, despite what its proponents have claimed, the convergence conception is not straightforwardly applicable in multimodal settings, and that recent

[3] As we'll see in §§2–3, many theorists have taken the view that the bundling in these multimodal cases is sufficiently close to the unimodal paradigms to merit being treated as instances of binding. And, perhaps obviously, in the absence of a commitment to a theoretical conception of binding of a sort that would prejudge the question of the paper, one might reasonably question that choice. For example, one might instead reserve 'binding' for unimodal cases, and describe the multimodal cases as instances of 'multimodal integration' (arguably a more standard term for them in any case).

Of course, there's no point in fighting over the label. I propose to apply 'binding' to multimodal cases, with the caveat that this terminological choice should be understood to open, rather than close off, the substantive question of what theoretical account of the notion we should adopt.

attempts to overcome this difficulty are unsuccessful (§2). Next I'll consider critically a second grade of multimodal involvement—a revisionary conception of multimodal binding in terms of associative linking (§3); I'll argue that, while prominent objections against this view are unconvincing, it leaves important phenomena unexplained. This will lead me to offer a novel, deflationary, third grade of multimodal involvement, which I call the mereological co-constituency view; I'll present this account of multimodal binding, and argue that it answers to the empirically motivated needs that multimodal binding has been called on to address (§4). Finally, I'll briefly draw some morals (§5).

Before I begin, a few qualifications concerning scope and aims are in order. First, in what follows I won't worry about the difficult and important question of how to individuate perceptual modalities. All examples discussed will be drawn from the traditional, Aristotelian, five modalities; I will assume that these are modalities, and that they are distinct (though possibly not in all explanatory contexts). Second, though the reasons I will advance for a deflationary/non-traditional treatment of intermodal binding are, in principle, applicable to intramodal/unimodal binding as well, I think this view is much less well motivated by the empirical facts about intramodal cases; in particular, considerations about the ontological diversity of feature-bearing objects, which I take to be central to motivating the move away from the convergence conception, do not arise in such cases. Third, though below I'll be engaged in advocacy for the co-constituency account as a powerful and general conception of multimodal binding, it would be inappropriate to insist at the outset that there has to be a single best model that applies to every case. If, at the end of the day, co-constituency earns its keep merely as one component of a well-motivated pluralism about multimodal binding, then so be it. Finally, nothing in what I say should be taken to deny that we multimodally perceive distal objects and events. There's no question that we have perceptually-informed representations of individual objects and events bearing properties, and that we have multiple, distinct modal channels that contribute to these representations. Nothing I say below is intended to conflict with these facts.

2 Convergence

To fix intuitions, consider a canonical multimodal case of perceptual bundling, such as the McGurk effect. You see video of a speaker uttering the syllable /ga/ while simultaneously hearing an audio recording of the speaker uttering the syllable /ba/; as a result you perceive the syllable /da/.[4] On its face, this appears to be an instance of informational interaction between bundled visual and auditory features—note that you experience things quite differently, and indeed the modification to /da/ disappears, when you apprehend them in separate, successive (hence, unbundled) unimodal presentations.

[4] For an example, see https://youtu.be/yJ81LLxfHY8.

If this and the other phenomena cataloged above are to be regarded as instances of perceptual binding, we need a model for thinking about binding in a multimodal setting. An extremely natural starting answer presents itself: why not simply extend the convergence model, familiar from discussions of unimodal binding, to multimodal cases? Recall that the convergence account construes binding as the convergent attribution of distinct features to one and the same sensory individual. In principle, and at this level of description, the convergence model seems just as applicable to cases of multimodal as unimodal binding; the difference will lie merely in whether the distinct features attributed to one sensory individual are extracted by multiple modalities or one. Applying this model to the specific phenomenon of the McGurk effect, the claim would be that vision and audition convergently attribute their own features to one single individual (say, an utterance event).

There is a wrinkle. In an important paper, Nudds (2014) points out that the extension just contemplated can take two quite distinct forms, which he labels 'amodal' and 'crossmodal' integration. Though both count as robust types of integration between information derived from distinct unimodal perceptual processes, they differ in the types of representations they envisage as resulting from the integrative process. In amodal cases, integration of unimodal representations results in the construction of a brand new, amodal representation of the distal target:

> [in amodal integration . . .] a number of initially distinct processing streams combine to produce a single amodal representation of an object, that represents it as having features—such as spatial and temporal features—that may have been perceived with more than one sense modality, as well as features—such as colour—that are modality specific. A single amodal object representation may represent an object as shaped, coloured, and as the source of a sound. It follows that the same kind of amodal-object representation plays a role in explaining our perceptual awareness of particular things perceived with any of the sense modalities. Both our visual perception of an object and our auditory perception of a sound source is explained by appeal to the same kind of amodal object representation (Nudds 2014, p. 173)

In contrast, on the crossmodal model, the result of integration is not the creation of a brand new amodal representation, but rather a coordination relation holding between unimodal representations that all converge on the same distal target:

> with respect to features that can be perceived with more than one sense, the visual-object representation and the auditory-object representation represent the object as having the same features, for example, as being at the same location, occurring at the same time, and so on, but the visual-object representation of the object will also represent it as having features that are specific to vision, and the

> auditory-object representation of the object will also represent it as having features specific to audition (Nudds 2014, p. 174).

These two models of multimodal integration are importantly different, and therefore useful to have on the table in attempting to come to grips with individual cases.[5] However, and notwithstanding their differences, it is important to see that both crossmodal and amodal integration are species of convergence: both are committed to construing binding as the attribution of features to a single distal object. The two models differ on the questions of how many representations (one or many) and what types of representation (amodal or unimodally specific) result from integration. But they agree that, whatever representations do so result have the role of attributing multiple features—namely, those features derived from the unimodal inputs to integration—to one and the same distal individual. Despite their differences, then, they agree that binding integration results in the convergent attribution of *many* features to *one* individual.

The convergence conception of multimodal binding—particularly as usefully precisified by Nudds—is a simple, straightforward, and well-articulated extension of ideas that have proven their merit in unimodal settings. It has been defended by leading theorists (e.g., Nudds 2014; O'Callaghan 2014, 2016, 2017; Green 2019).[6] Despite this, the application of the convergence to multimodal cases faces a *prima facie* worry concerning the diversity of sensory individuals from distinct modalities.

The concern arises from the reasonable assumption that if features from distinct modalities are convergently bound to one common individual, then that one individual must be an appropriate locus for feature bearing in both modalities. That is, if modalities m_1 and m_2 are to share one and the same sensory individual, a, it must be that m_1 and m_2 overlap in the types of objects that they recognize as sensory individuals (and, of course, that a falls into the overlap). But, on the face of things, it's not obvious that this constraint is satisfied by pairs of modalities that are associated with sensory individuals, and for which there is evidence of multimodal binding. On the contrary, it seems

[5] As Nudds points out, one needn't construe every instance of integration as conforming to the same model; moreover, both models are compatible with the existence of unimodal/unintegrated representations in perceptual systems. The point of such models is not to facilitate a once-for-all time description that will apply to all perceptual processing, but simply to make available well-formulated descriptive options.

[6] O'Callaghan's position is more nuanced than this glib attribution lets on. Though he has certainly endorsed convergence as a description of (some cases of) *property* binding, he explicitly follows Treisman in construing binding as applying to both properties and parts (e.g., O'Callaghan 2014, pp. 74-75, 78)—often using 'feature' to range over both; and his descriptions of *part* binding are more readily understood as co-constituency than convergence. (In contrast, my topic here is exclusively the binding of properties.) Moreover, once he allows a diversity of models under the broad heading of binding, it's open to him to extend a co-constituency story to individual cases of property binding as well (e.g., see the discussion of audiovisual property binding in O'Callaghan 2011, 395ff). In short, O'Callaghan is not a uniform proponent of convergence. (Thanks to O'Callaghan for discussion of these matters.)

quite likely that the individuals associated with different modalities often fail to overlap.

For example, consider the modalities of vision and audition, both of which have been held by many to attribute qualities to sensory individuals, and involving which many theorists have claimed that there is good evidence of intermodal binding. On the one hand, many authors hold that visual sensory individuals are (at least in canonical instances) ordinary material objects—cohesive, temporally persistent objects extended and bounded in space (e.g., Marr (1982); Spelke (1990); Cohen (2004); Matthen (2005, pp. 277-282); Pylyshyn (2007); Dickie (2010); Nanay (2013, p. 51); but cf. Clark (2000)). On the other hand, while many have been attracted by the idea that audition attributes qualities to sensory individuals, it hasn't seemed attractive to claim (and, as far as I'm aware, no one has claimed) that such auditory individuals are ordinary material objects. Instead, there are a number of proposals in the literature that construe auditory individuals as things other than material objects. Thus, O'Callaghan (2007, 2009) holds that auditory properties qualify medium-disturbing events (cf. Matthen 2005); Casati and Dokic (2009) believes they qualify monadic events befalling material objects; Sorensen (2008) and O'Shaughnessy (2009) think they qualify waves; Nudds (2009, 2010a,b) thinks they qualify structures exemplified by waves; Pasnau (1999), Kulvicki (2008), Cohen (2010), and Kulvicki (2014) think they qualify dispositional properties of objects; Soteriou (2018) thinks they can qualify all of these; and Young and Nanay (2020) think they qualify temporally extended, causally composite individuals built from the types of entities cited by the others. But, given the standard conception of the sensory individuals of vision as ordinary objects, every one of these views about auditory individuals would seem to threaten the possibility of overlap between the sensory individuals of vision and those of audition. Consequently, every one of them threatens the possibility of extending the convergence conception of binding to intermodal cases involving vision and audition.

Moreover, this concerns spreads to putative instances of multimodal binding involving other pairs of modalities. Thus, to choose another example, while it is at least initially plausible that the sensory individuals of touch may, at least in some cases, overlap with those of vision (Fulkerson 2014), this fact suggests (for the reasons just rehearsed) that the sensory individuals of touch are unlikely to overlap with those of audition. Or, again, if there are sensory individuals for olfaction and gustation (Lycan 1996; Matthen 2005; Batty 2014; Carvalho 2014; Smith 2015), it is highly unobvious that these are of the same type as either the ordinary material objects that serve as visual (and perhaps tactual) individuals or any of the candidate types proposed as auditory individuals.

It would seem, then, that the relatively slight degree of overlap between sensory individuals associated with distinct modalities poses a significant *prima facie* threat to the extension of the convergence view to multimodal instances of binding. What to do?

Defenders of the convergence conception (O'Callaghan 2008; Nudds 2009; O'Callaghan 2014, 2016; Green 2019) have attempted to answer this threat by proposing to understand multimodal sensory individuals as structured, mereologically complex individuals that have unimodal sensory individuals as constituents, but that can (*qua* complex wholes) be shared by multiple modalities. Thus, in an eloquent expression of the proposal, O'Callaghan writes that,

> Multisensory perceptual objects are mereologically complex individuals with hybrid structure. Some of their parts and features are perceptible through one sense, and some are perceptible through another. Each sense provides a partial perspective on the whole. The complex whole is perceptible as such through the coordinated use of multiple senses. The key is that perceiving a whole does not require perceiving each of its parts or features. Visuo-tactile objects include the material bodies on which vision and touch converge—a subset of visible objects. Audio-visual objects are environmental happenings that involve bodies and include sounds. Flavors are complex, and they are only fully perceptible using multiple senses; however, flavors are properties, attributed to things we ingest, rather than individuals (O'Callaghan 2016, p. 1287).[7]

Before we can determine whether this suggestion resolves the problem about the diversity of sensory individuals, and thereby clears the way for the extension of the convergence conception to multimodal instances of binding, two observations are required. The first is that, if they are to be of help in understanding binding, the mereological complexes at issue cannot be mere, unstructured bags of features (as the notion of mereological summation may bring to mind). Rather, these must be *structured* complexes, where the structure in question extends to part-whole relations, predicational relations, and possibly more. The second observation concerns the source and nature of this structure. As presented by O'Callaghan and others, the proposal appears to be that multimodal sensory individual should be identified with complex mereological entities licensed by the One True Metaphysical Theory of Mereology.[8] But that's potentially misleading. For, if the mereological complexes invoked here are to serve their purpose in defending an understanding of multimodal binding as convergent psychological attribution of features to single individuals, then those individuals must be not only metaphysically real, but also available for the psychological construction of perceptual representations (and, for forms of binding of which there is awareness, also available to conscious

[7] It's worth mentioning that the mereological conception of multimodal *sensory individuals*, *qua* objects of perception (as defended by O'Callaghan 2016), should be distinguished from both the mereological account of conscious multimodal *experiences*, *qua* complex phenomenal states (as defended by Bayne and Chalmers 2003; Bayne 2010) and the idea that some *unimodal* sensory individuals are themselves mereologically complex (as defended by Young and Nanay 2020).

[8] And, indeed, the proposal appears to depend on a kind of mereological realism that, philosophy being what it is, is denied by some (e.g., Rosen and Dorr 2002; Sider 2013). As this issue goes substantially beyond the scope of this paper, I'm prepared to grant as much metaphysical realism about mereology as is needed for present purposes.

awareness). As such, the sort of mereology required by the solution on offer is not the metaphysical mereology of parts and wholes, but a psychologized mereology of part/whole groupings made available by our perceptual-*cum*-cognitive endowments. Mere metaphysical mereological structure without a psychological reflection won't do.

Even granting these clarificatory points, I now want to point out that the proposed solution does not, as advertised, solve the diversity puzzle, or, therefore, save the convergence conception of multimodal binding. The problem is that, while the mereological complex proposal provides a common individual accessible to distinct modalities—viz., the psychologized mereological whole—that common individual is in many ordinary cases not a locus for the exemplification of features attributed by perceptual modalities. To see why, consider O'Callaghan's example of a clapping event, which he proposes to construe as a mereologically complex happening consisting of visible and audible components (O'Callaghan 2016, p. 1284). Suppose we accept this, and suppose we also allow that the modalities of vision and audition converge in connecting us with this mereologically complex individual in virtue of connecting us with distinct parts of this common whole. Perhaps we will also grant, therefore, that we see the whole (in virtue of seeing its visible components) and hear the whole (in virtue of hearing its audible components), hence that there is an individual that we both see and hear. Still, notwithstanding these claims, it is the ordinary material objects we see—not the mereological complex having material objects as constituents—that exemplify visible properties such as colors and shapes. Vision represents the *hands*, and not the happening of a clapping, as bearing a size and color. Likewise, audition represents the *sound*, and not the complex visual-*cum*-auditory happening of a clapping, as bearing the auditory qualities of volume and timbre. As such, while the mereological complex proposal may give us a way of saying that distinct modalities connect us to one and the same individual, it does not vindicate the convergence conception of multimodal binding as perceptual attribution of features from distinct modalities to one and the same individual.[9]

It would seem, then, that our *prima facie* challenge to the possibility of multimodal binding remains unanswered, and stands as a threat to the convergence conception (on any precisification).

3 Association

A quite different proposal for understanding multimodal binding, defended by Goldstone (1998), Fulkerson (2011), Bayne and Spence (2014), and Goldstone

[9] Response: Properties can be inherited by mereological complexes from their constituents. If the hands have a property F (say, a size) and are constituents of a mereologically complex clapping, then we can attribute F to the complex as well.

Counter-response: That sort of inheritance, construed so as to apply to the perceptible features of all mereologically complex perceptible items, is untenable. A mereologically complex figure might have components that are circular and square; if inheritance obtained in general, it would license the conclusion that the complex figure is circular and square. Surely that can't be right.

and Byrge (2015) (and possibly Connolly (2014a,b, 2019)) construes multimodal binding as a relation of "association" between distinct individuals figuring in unimodal attributions. On this conception, there is no single sensory individual associated with multiple modalities, and that bears the features supplied by each of them. Rather, each modality (or, for that system within a given modality) attributes features to its own individual; binding occurs when the numerically distinct individuals that result from such multiple perceptual attributions stand in (consciously-represented) relations of association:

> Multisensory perceptual experiences do not involve the direct predication of features onto individual perceptual objects. Instead, there is an association between experiences .... What we experience is a higher-order association between sensory experiences (Fulkerson 2011, p. 506).

It may be helpful to think of the association view as a weakening of the crossmodal convergence model: where the latter involves coordination between distinct unimodal representations converging on the same individual target, the association view involves coordination between distinct unimodal representations that need *not* converge on a common target.

To see what this amounts to concretely, consider a case of multimodal binding between an auditory attribution of a pitch and a visual attribution of a spatial location. Where, on the convergent attribution view, such binding requires attribution by both modalities to a common individual, the associative proposal treats binding as an instance of psychological association holding between one auditory individual bearing a pitch and a distinct visual individual bearing a location. Since, on the latter view, the only sensory individuals required are those associated with the individual modalities, there is no need for individuals spanning separate modalities. Consequently, the observed diversity of unimodal objects is not a challenge to this conception of binding, as it is for any version of the convergent attribution conception. This seems an important advantage.

Despite this, the association view confronts a few challenges. First, as O'Callaghan (2014, p. 84) notes, mere association won't account for the observation that subjects ordinarily display fallible conscious awareness of whether feature attributions from distinct modalities are bound or not. After all, though subjects are sometimes fallibly aware of associations holding between their psychological states, sometimes they are not. Consequently, if any old relations of association were sufficient for intermodal binding, we would expect to see less awareness of whether binding obtains than we do. To answer this, the association view should require for binding not only that there be higher-order associations between unimodal perceptual representations, but that these associations are fallibly available to conscious awareness.

But O'Callaghan objects that even this is not sufficient—that even consciously available relations of association are too weak to capture the phenomenon of experienced binding:

> ...if seeming to be associated or to tend to co-occur does not guarantee seeming to share a common object or source, then appearing merely as being associated or as tending to co-occur is too permissive to capture the relevant distinctions among the cases discussed above. For instance, a sound and an image may seem merely to be associated or to tend to co-occur without seeming perceptually to share a common source. A rough surface and a red surface may seem to be associated without their seeming perceptually to be one surface or to belong to one object. Mere associations thus do not suffice for an account of that to which one may be multimodally perceptually sensitive, and they do not suffice for an account of multimodal perceptual awareness (O'Callaghan 2014, pp. 84-85).

O'Callaghan's worry is well-taken: we can represent feature exemplifications as co-occurring (or even tending to co-occur) without thereby representing them as bound, so a mere co-occurrence account of binding overgenerates instances of binding.[10] However, this is a point that the proponent of the association view can accommodate, so long as she understands the notion of association in a way that distinguishes it from mere co-occurrence. Even on a classical construal of association, spatiotemporal co-occurrence is neither necessary (today's olfactory experience of Grandma's brand of perfume is now associated with the remembered, long past, experience of the same perfume, or with the also long past gustatory experience of the cookies Grandma used to bake) nor sufficient (if a subject is conditioned to associate the sound of a bell with an electric shock, the equally proximate and co-occurring color of the walls in the room need not be so associated) for association.[11] (As Fulkerson (2011, p. 506) writes, "The coordination involved in the auditory-visual case is often (though not always) sensitive to temporal and spatial continuity.")

For that matter, the association proposal need not be tied to such a narrowly classical associationist conception of psychological association. Thus, in developing his version of the proposal, Fulkerson (2011) endorses a more flexible understanding of association:

> The idea of an associative relation is intended as a general concept that can explain a wide variety of multisensory interactions... (494).
>
> This notion of an associative relation is meant to be a general means of characterizing the structure of a range of distinct sensory mechanisms relating perceptual experiences ... (497).

[10] On the other hand, O'Callaghan's formulation of the point so as to demand that an account of binding amount to a representation of common sourcehood, in particular, seems unjustified. This demand goes well beyond the data, and is tantamount to begging the question against weaker (non-convergence) theoretical accounts of binding.

[11] Indeed, some psychotherapeutic conditioning techniques (e.g., extinction therapy (Marks 1979), applied behavioral analysis (Lovaas 1987), exposure-based therapy (Huppert and Roth 2003)) specifically aim at gradually increasing the spatiotemporal distance between associated items. This aim would be incoherent if association relations were restricted to items in narrow spatiotemporal proximity.

> . . . multisensory experiences involve some higher-level relation between separate experiences. . . (504).

Understood in this more flexible way, the association view has resources to evade the problem we are considering.[12]

A potentially more worrisome overgeneration objection to the association account comes from Green (2019, pp. 15-16). Green notes that the association relation doesn't, by itself, come with numerical limitations: it can apply as easily between two items as between three, four, or even more (subject to limitations on working memory). Consequently, he suggests, an association account of binding, unless supplemented by further constraints (that would require their own motivation), would seem to predict more binding than we observe. Thus, for example, in a case where one sound is presented with multiple visible flashes, an association account should allow binding between the one sound and many of the presented flashes. In contrast, if we thought of binding as constrained to apply to features of one and the same individual (as on a convergence view of the type Green favors) we would expect that the logic of identity would effectively constrain the binding of the sound to just the visible flash that is an aspect of the very same sensory individual. And he points to results of van der Burg et al. (2013) suggesting that there are, indeed, one-to-one constraints on at least some cases of audible/visible binding. In these experiments,

> 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. . . . 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 (Green 2019, pp. 15-16).

To be clear, I do not believe that this challenge is completely decisive (nor does Green). For one thing, the challenge, as posed, rests on a single study, and concerns only a single form of multimodal binding. For another, even if adopting an association conception of binding leaves us without an explanation of the observed numerical limitations on binding in terms of the binding relation itself, we cannot rule out the possibility of independent explanations of these facts in terms of properties of working memory, attention, or something else. Alternatively, we might even accept these limitations as brute facts about perceptual systems. Still, I think it is fair to say that the association account

[12] In conversation, Fulkerson assures me that he always intended a more flexible construal than that found in classical associationists. Reinterpreting the association view in this more flexible way brings his position much closer to the view defended in §4.

of binding leaves us without an explanation of an important limitation on the phenomenon, and that, other things equal, it would be preferable to have such an explanation.

4 Binding and co-constituency

4.1 The mereological co-constituency view

I propose that the way forward lies in bringing together separate insights from two views already discussed—the mereological complex answer to our puzzle about the diversity of sensory individuals (§2), and the associative account of multimodal binding (§3).

In §2, I argued that mereological complexes built from unimodal constituents are unsuitable as targets of convergent feature attribution because these complexes (as opposed to their unimodal mereological components) do not bear perceptual features contributed by individual modalities—and, *a fortiori*, cannot bear *multiple* features contributed by *distinct* perceptual modalities. But this dissatisfaction gives us no reason for denying that there are mereological complexes, that perception represents mereological complexes of various types, or that represented mereological complexes have an important role to play in our understanding of multimodal perception.

On the contrary, there is ample reason for accepting that we are perceptually connected to such complexes, quite independently of one's views about sensory individuals and binding. It should be relatively uncontroversial (barring a mereological nihilism of the sort we set aside earlier) that the objects and events we perceive are mereological constituents of larger, more complex objects and events. An apple is a mereologically complex object composed of (among other constituents) the surface facing you and the surfaces not facing you. So, too, a car crash is a mereologically complex event composed of (among other constituents) both visible and audible components. Just as O'Callaghan points out, we perceive the wholes by virtue of perceiving their parts. You see the apple, the complex object, by seeing one of its constituents (its facing surface), even though you do not see others of its constituents (e.g., its non-facing surfaces, its insides). Similarly, you see the complex event of the car crash by seeing its visible aspects, even though you do not see its non-visible aspects; and you hear it by hearing its audible aspects, even though you do not hear its non-audible aspects. Moreover, and crucially for our purposes, a distinguished history of research in perceptual psychology dating from at least the Gestalt period makes clear that perception organizes units at various levels into complex arrangements of parts and wholes—e.g., those that comprise objects (e.g., Biederman 1987; Spelke 1990; Tarr and Bülthoff 1998), groups (Koffka 1935; Wertheimer 1938; Palmer 1999; Chang et al. 2007; Wagemans, Elder, et al. 2012; Wagemans, Feldman, et al. 2012; Goldstone and Byrge 2015), and events (Michotte 1946/63; Johansson 1973; Gibson 1979; Zacks and Tversky

2001; Zacks, Speer, et al. 2007). This is to say that these forms of mereological organization are not only real, but psychologically represented.

On the other hand, while we saw in §3 that the association account of multimodal binding may underconstrain/overpredict the phenomenon of binding, there was much to like about this proposal as well. In particular, it offered the hope of avoiding the need for convergence, and so sidestepping the challenge posed to convergence by the diversity of unimodal individuals.

The solution is to join these ideas—to recognize the reality of (psychologically represented) multimodal mereological complexes, and to understand multimodal binding as the coordinated representation of co-constituency, but without requiring that bound features convergently apply to any one individual.[13] I'll call the resulting position the (MEREOLOGICAL) CO-CONSTITUENCY view of binding.

To understand the view, consider a paradigmatic instance of multimodal binding—the visual-*cum*-auditory perception of a car crash. A car crash is a complex mereological event, which contains visible and audible aspects. When we perceive this complex event, perception builds a complex, multimodal representation of this structured target. The current proposal is that what binds visible and audible aspects in perception is not the convergent attribution of visible and audible features to any single entity, but the fact that perception coordinates its representations of visible and audible aspects, treating them as sisters/co-constituents (i.e., nodes dominated by a single node) of the complex, hierarchically organized, event of the car crash. What we've been calling "visible aspects" of the car crash are exemplifications of visual features (as it might be, colors, forms) by specifically visual sensory individuals (as it might be, ordinary material objects). Likewise, the crash's "audible aspects" amount to exemplifications of auditory features (as it might be, pitches, loudnesses) by specifically auditory sensory individuals (as it might be, medium-disturbing events). The co-constituency view says that these visual and auditory aspects are bound in this sense: both the exemplification of visual features by visual individuals and the exemplification of auditory features by auditory individuals are represented by perception as mereological constituents of one and the same complex event of the car crash. That is, visual and auditory features are bound by figuring in coordinated representations making attributions to parts of the same (represented) complex whole.

The co-constituency view has a number of attractive advantages.

[13] Qualification: Perhaps there are cases in which binding results in the attribution of novel features not exemplified in any unimodal perceptual episode. (O'Callaghan (2017, p. 174) offers the example of *mintiness*, which plausibly requires binding between olfactory, gustatory, and oral somatosensory features.) If so, the co-constituency account can treat such novel features as qualifying the complex (and not the unimodal individuals of any individual modality). Even so, such cases are not helpfully construed as instances of convergence. Here it is not that unimodal features from distinct modalities convergently qualify the complex. Rather, unimodal features from distinct modalities non-convergently qualify the distinct individuals of those modalities that are constituents of the complex, and this leads to the downstream attribution of numerically distinct, and novel, features to the complex.

First, it is ontologically and representationally conservative. It is committed only to the existence and representation of (i) unimodal features and unimodal individuals already motivated by discussions of unimodal binding, together with (ii) the mereologically complex events built from the latter that, as we have seen, all parties to the debate have reason to accept. In particular, it is worth emphasizing that, for reasons discussed in §2, the co-constituency theorist's reliance on a psychologically represented structure of part-whole relations is shared by convergence theorists, who invoke mereology as a response to concerns about the diversity of unimodal individuals; hence, this commitment is not a special burden for the co-constituency view.[14]

Second, though it allows that distinct perceptual modalities relate to a common entity (the mereological complex), the co-constituency view is not committed to construing binding as convergence—it does not require for binding that distinct modalities convergently attribute features to any single entity. As such, if distinct modalities predicate their features to distinct entities, this is no obstacle to the view's treating such features as bound. The diversity problem is not a problem for the co-constituency view.

Third, and unlike (the classical associationist construal of) the association view, the co-constituency view can allow for numerical limitations on binding arising from numerical limitations on the psychological representation of mereological structure. The crucial insight here is that, whatever restrictions one takes to constrain merely metaphysical mereological summation, there are clearly substantive psychological restrictions on the organization of represented parts into represented wholes. It is fair to say that these psychological restrictions are complicated, multifarious, and span a range of levels at which parts can be organized into wholes (as attested by the vast literature on perceptual grouping). However, it does seem, as a matter of empirical fact, that the grouping operations performed by our psychological mechanisms do not permit arbitrary/unrestricted recombination. Whether or not the scattered collection of the Queen's nose and the Eiffel Tower constitues a metaphysically legitimate mereological sum, these two objects won't be constituents of a perceptually represented whole (in ordinary cases). (This is something that can be tested by standard psychological grouping criteria such as object-specific preview effects, preferential looking times, etc.). Consequently the features of these objects won't ordinarily enter into perceptual binding relations. Against this backdrop, the one-to-one constraints uncovered by van der Burg et al. (2013) can be construed as further limitations on psychological grouping. Their results appear to support the idea that, when the psychological mechanisms underlying grouping build a whole connecting an auditory tone with one visually presented event, this operation inhibits an alternative grouping of that same tone with earlier or later visually presented events. Though it might be that classical association (and metaphysical mereology) lack analogous

[14] One possible account of such a generalized psychological representation of this structure might invoke event files (Hommel 2004); other options include appeals to a general notion of indexing, predictive coding, and more. My defense of co-constituency need not take sides on these (important) implementational matters.

constraints, the co-constituency view has no trouble accommodating the finding that relevant psychologically mediated grouping operations are, as a matter of fact, subject to numerical limitations.

Finally, the co-constituency view can be viewed as a conservative extension of the convergence conception: the former can explain the attractions of, and inherit the successes of, the convergence view, while succeeding in cases where the latter fails. Specifically, convergence can be seen a special, degenerate case of co-constituency—a case where co-constituents of a single represented whole happen to stand in the identity relation (as well as the mereological co-constituency relation). In such cases, perceptual feature F_1 applies to a_1, perceptual feature F_2 applies to a_2, and, whether these features and individuals are contributed by the same or different modalities, it turns out that $a_1 = a_2$; hence we can say that F_1 and F_2 convergently qualify the very same individual. However, given that a_1 can be counted a trivial constituent of any complex of which it itself is a constituent, it is (in a good, if trivial, sense) its own co-constituent. Consequently, such cases can just as easily be described within the co-constituency framework.

4.2 Co-constituency and the ties that bind

I now want to argue that the account of binding sketched in §4.1 explains the phenomena that have provided the strongest support for believing in multimodal binding, and so is plausibly up to the job of answering the theoretical needs for which that notion has been invoked.

A first point to make is that, on the co-constituency view, multimodal binding arises from a perceptual, rather than an entirely post-perceptual, representation of co-constituency, and so should be counted a substantively perceptual phenomenon. This is important because, while it might initially seem tempting to think of intermodal binding in exclusively postperceptual terms—say as a judgment that features F and G are linked, there are persuasive reasons for rejecting this proposal. One reason is that that there are illusions of intermodal binding (as distinguished from illusory representations of any of the features bound), in which perception erroneously represents binding even though postperceptual judgment does not (Cinel et al. 2002). Another is that illusory representations of binding can, in many cases, survive the addition of cognitively represented information that would undercut a fully postperceptual representation of binding: e.g., the McGurk illusion persists even after subjects learn about the effect and experience the auditory stimulus without the visual stimulus and vice versa.[15] A third, though perhaps weaker, reason is that perceptual judgments about intermodal binding and the perceptual guidance

[15] That is, this particular binding effect seems to be encapsulated from cognition in just the sense that is true of some other illusory perceptual states—say, that produced by the Müller-Lyer configuration (Fodor 1983). (This is not to say that all instances of binding are so encapsulated; on the contrary, and as discussed in note 17, evidence involving the motion-bounce illusion and other cases suggests that at least some instances of the representation of binding are penetrated by, i.e., are not encapsulated from, cognition.)

of action on the basis of intermodal binding seem to have an immediacy that some have thought incompatible with a postperceptual understanding of the phenomenon (O'Callaghan 2014). In any case, all of this is just what we should expect on a perceptual understanding of intermodal binding such as the co-constituency view.

A second consideration is that the co-constituency account allows for a wide range of intermodal informational interactions that go beyond the aggregation of information from multiple modalities, including those that many authors have appealed to explicitly to motivate the convergence conception (and, since these are often not distinguished, for accepting intermodal binding itself).[16] One paradigm of such an interaction occurs in the McGurk effect, where it is not only true that a visual feature is bundled together with an auditory feature, but that the fact of bundling results in modulation or modification of the auditory feature representation: bundling the visual representation of /ga/ with the auditory representation of /ba/ results not just in an experience combining copresent contributions from two modalities, but modification of the content of the auditory representation (from /ba/ to /da/). In another type of interaction, a feature representation in one modality cues or disambiguates a feature representation in another. Thus, for instance, in the motion-bounce (/sound-induced bouncing) illusion (Sekuler et al. 1997), a sound played at the moment when two visually perceived moving objects cross trajectories leads observers to disambiguate the visually ambiguous crossing event as a collision rather than as one object streaming through another: the proposal is that binding the auditory feature together with the visual feature results in a modification or reconstrual of the latter.[17] A third paradigm of such interactions is manifest in the finding that there are intermodal object-specific preview advantages and penalties (Zmigrod et al. 2009; Jordan et al. 2010) mirroring the object-specific preview advantages and penalties for unimodally perceived objects (Kahneman et al. 1992): following recognition of a feature in modality m_1, reaction times are lower for the recognition of a second and congruent feature in modality m_2, but higher for the recognition of a second and incongruent feature in m_2, when the second feature is bound to the first, whether or not $m_1 = m_2$.

The lesson from these cases is that the operation of grouping itself can have significant psychological consequences for grouped features. However, and crucially for our purposes, this lesson is agnostic between the convergence and co-constituency conceptions of binding: we can take the lesson on board whether we think of the features in question as applying to one and the same

[16] Thus: "The main source of empirical evidence for intermodal binding is that sensory systems interact and share information" (O'Callaghan 2014, p. 81); see also Nudds (2014), O'Callaghan (2015, 2017), and Green (2019). Theorists have used a varieties of labels to mark out these interactions. In the parlance of Ernst and Bülthoff (2004) these are instances of *cue integration*, as opposed to mere *cue combination*. de Vignemont (2014) makes a similar distinction between *integrative* and *additive* forms of binding.

[17] Results of Grassi and Casco (2010) show that this type of binding is penetrated by the presumably cognitive representation of feature congruence: the bounce interpretation of the visual motion is favored only when the auditory feature is consistent with an impact event (cf. Vatakis and Spence 2007).

individual (as per convergence) or not (as per co-constituency). Proponents of convergence will explain the data by saying that distinct feature representations can influence one another in interesting ways when they involve convergent attribution to a single individual: that the value assigned to one such feature can modify the value assigned to another, that there is facilitation/suppression between them, and so on. But none of this depends essentially on convergence, as such. Proponents of the co-constituency account will claim that such influence occurs between distinct feature representations when the latter stand in the relevant sort of co-constituency relation—that the perceptual attribution of F_1 to o_1 and F_2 to o_2 can exert these forms of mutual psychological influence even when $o_1 \neq o_2$.[18]

All of this is to say that the central features of multimodal binding—features that have not only convinced many writers to believe in multimodal binding, but that many have thought can only be understood as resulting from multimodal convergence—are equally explicable in terms of mereological co-constituency.

5 Conclusion

The co-constituency account of multimodal binding has much to recommend it. It avoids problems that beset competing views, provides illuminating accounts of phenomena that have compelled theorists to embrace multimodal binding, and highlights the importance of continuing research into the psychological mechanisms underlying various forms of perceptual grouping.

Despite these advantages, and as mentioned at the outset, there is no reason for insisting in advance that all multimodal binding must conform to the co-constituency model. It may be that the best approach involves the application of different models to different cases. This is, of course, a broadly empirical question that should not be prejudged from the armchair. My hope is that spelling out the co-constituency model, and comparing it with rivals, will aid in this assessment.

Finally, it is worth acknowledging that the co-constituency view has significantly deflationary, revisionist consequences regarding both binding and sensory individuals. Perhaps the view's most revisionary aspect (which it shares with the association view but not the convergence view) lies in its recognition of two different levels of sensory individuals, corresponding to two different roles that the latter have been called upon to play. On the one hand, the co-constituency view allows that there are whatever unimodal sensory individuals that are contributed by individual perceptual modalities, and that bear the perceptual features of those modalities. On the other hand, the view also allows that perception connects us with multimodal, complex entities, themselves built from unimodal constituents. These mereologically complex

[18] Note that, on this account, such interaction is no "mere causal influence" between otherwise independent entities (O'Callaghan 2014, p. 81; cf. Nudds 2014, p. 179); it is causal influence *between items that stand in psychologically significant co-constituency relations*.

entities won't, in general, bear perceptual features themselves, though their mereological constituents will; and, of course, these complex entities are at the heart of the view's explanation of binding relations between perceptual features.

Accepting these two distinct levels of sensory individuals amounts to a recognition that, in the multimodal setting (and apart from degenerately unimodal special cases), the role of bearing perceptual features comes apart from the role of accounting for binding. If accepting this lesson is the price of an adequate account of intermodal binding, we should pay it.[19]

References

Batty, Clare (2014), "Olfactory Objects," in *Perception and its Modalities*, ed. by Stephen Biggs, Dustin Stokes, and Mohan Matthen, Oxford University Press, pp. 222-245.

Bayne, Tim (2010), *The Unity of Consciousness*, Oxford University Press UK.

Bayne, Tim and David J. Chalmers (2003), "What is the Unity of Consciousness?" In *The Unity of Consciousness*, ed. by Axel Cleeremans, Oxford University Press.

Bayne, Tim and Charles Spence (2014), "Is Consciousnes Multisensory?" In *Perception and Its Modalities*, ed. by Dustin Stokes, Stephen Biggs, and Mohan Matthen, New York, USA: Oxford University Press, pp. 95-132.

Biederman, Irving (1987), "Recognition-by-components," *Psychological Review*, 94, pp. 115-147.

Botvinick, Matthew and Jonathan Cohen (1998), "Rubber hands' feel'touch that eyes see," *Nature*, 391, 6669, p. 756.

van der Burg, Erik, Ed Awh, and Christian N. L. Olivers (2013), "The Capacity of Audiovisual Integration Is Limited to One Item," *Psychological science*, 24, pp. 345-351, DOI: 10.1177/0956797612452865.

Campbell, John (2002), *Reference and Consciousness*, Oxford University Press.

Carvalho, Felipe (2014), "Olfactory Objects," *Disputatio*, 6, 38, pp. 45-66, DOI: 10.2478/disp-2014-0003.

Casati, Roberto and Jérôme Dokic (2009), "Some Varieties of Spatial Hearing," in *Sounds and perception: New philosophical essays*, ed. by Matthew Nudds and Casey O'Callaghan, Oxford University Press, Oxford, pp. 97-110.

Chang, Dempsey, Keith Nesbitt, and Kevin Wilkins (2007), "The Gestalt Principles of Similarity and Proximity Apply to Both the Haptic and Visual Grouping of Elements," in *Conferences in Research and Practice in Information Technology*, ed. by Wayne Piekarski and Beryl Plimmer, The Australian Computer Society, Ballarat, Australia, vol. 64, pp. 79-86.

[19] Thanks for discussion and comments from Kevin Connolly, Casey O'Callaghan, Will Davies, Matthew Fulkerson, Mohan Matthen, Bence Nanay, Ayoob Shahmoradi, Gedalyahu Wittow, and co-participants in a spring 2020 seminar on perceptual interactions.

Cinel, Caterina, Glyn Humphreys, and Riccardo Poli (2002), "Cross-Modal Illusory Conjunctions between Vision and Touch," *Journal of experimental psychology. Human perception and performance*, 28 (Nov. 2002), pp. 1243-66, DOI: 10.1037//0096-1523.28.5.1243.

Clark, Austen (2000), *A Theory of Sentience*, Oxford University Press, New York.

Cohen, Jonathan (2004), "Objects, Places, and Perception," *Philosophical Psychology*, 17, 4, pp. 471-495.

— (2010), "Sounds and Temporality," *Oxford Studies in Metaphysics*, 5, pp. 303-320.

Connolly, Kevin (2014a), "Making Sense of Multiple Senses," in *Consciousness Inside and Out: Phenomenology, Neuroscience, and the Nature of Experience*, ed. by Richard Brown, Springer.

— (2014b), "Multisensory perception as an associative learning process," *Frontiers in Psychology*, 5, p. 1095, DOI: 10.3389/fpsyg.2014.01095.

— (2019), *Perceptual Learning: The Flexibility of the Senses*, Oxford University Press, New York.

Dalton, P., N. Doolittle, H. Nagata, and P. A. S. Breslin (2000), "The merging of the senses: integration of subthreshold taste and smell," *Nature Neuroscience*, 3, 5, pp. 431-432, DOI: 10.1038/74797.

Dickie, Imogen (2010), "We Are Acquainted with Ordinary Things," in *New Essays on Singular Thought*, ed. by Robin Jeshion, Oxford University Press, pp. 213-245.

Ernst, Marc and Heinrich Bülthoff (2004), "Merging the Senses into a Robust Percept," *Trends in cognitive sciences*, 8 (May 2004), pp. 162-169, DOI: 10.1016/j.tics.2004.02.002.

Fodor, Jerry A. (1983), *The Modularity of Mind*, MIT Press, Cambridge, Massachusetts.

Fulkerson, Matthew (2011), "The Unity of Haptic Touch," *Philosophical Psychology*, 24, 4, pp. 493-516, DOI: 10.1080/09515089.2011.556610.

— (2014), *The First Sense: A Philosophical Study of Human Touch*, MIT Press, Cambridge, Massachusetts.

Geldard, F. A. and C. E. Sherrick (1972), "The Cutaneous "Rabbit": A Perceptual Illusion," *Science*, 178, 4057, pp. 178-179.

Gibson, James Jerome (1979), *The Ecological Approach to Visual Perception*, Houghton Mifflin, Boston.

Goldstone, Robert L. (1998), "Perceptual learning," *Annual Review of Psychology*, 49, 1, pp. 585-612, DOI: 10.1146/annurev.psych.49.1.585.

Goldstone, Robert L. and Lisa A. Byrge (2015), "Perceptual learning," in *The Oxford Handbook of the Philosophy of Perception*, ed. by Mohan Matthen, Oxford University Press, Oxford, pp. 812-832.

Grassi, Massimo and Clara Casco (2010), "Audiovisual bounce-inducing effect: When sound congruence affects grouping in vision," *Attention, Perception, & Psychophysics*, 72, 2, pp. 378-386.

Green, E. J. (2019), "Binding and Differentiation in Multisensory Object Perception," *Synthese*, DOI: 10.1007/s11229-019-02351-1.

Hommel, Bernhard (2004), "Event files: feature binding in and across perception and action," *Trends in Cognitive Sciences*, 8, 11, pp. 494-500, DOI: https://doi.org/10.1016/j.tics.2004.08.007.

Huppert, Jonathan and Deborah Roth (2003), "Treating Obsessive-Compulsive Disorder with exposure and response prevention," *The Behavior Analyst Today*, 4 (Jan. 2003), DOI: 10.1037/h0100012.

Jackson, Frank (1977), *Perception: A Representative Theory*, Cambridge University Press, New York.

Johansson, Gunnar (1973), "Visual perception of biological motion and a model for its analysis," *Perception & Psychophysics*, 14, 2, pp. 201-211.

Jordan, Kerry E., Kait Clark, and Stephen R. Mitroff (2010), "See an object, hear an object file: Object correspondence transcends sensory modality," *Visual Cognition*, 18, 4, pp. 492-503, DOI: 10.1080/13506280903338911.

Kahneman, Daniel, Anne Treisman, and Brian J Gibbs (1992), "The reviewing of object files: Object-specific integration of information," *Cognitive Psychology*, 24, 2, pp. 175-219, DOI: 10.1016/0010-0285(92)90007-O.

Koffka, Kurt (1935), *Principles of Gestalt Psychology*, Harcourt, Brace, & World, New York.

Kuang, Shenbing and Tao Zhang (2014), "Smelling directions: Olfaction modulates ambiguous visual motion perception," *Scientific Reports*, 4, 1, p. 5796.

Kulvicki, John (2008), "The Nature of Noise," *Philosophers' Imprint*, 8, 11, pp. 1-16.

— (2014), "Sound Stimulants: Defending the Stable Disposition View," in *Perception and Its Modalities*, ed. by Dustin Stokes, Stephen Biggs, and Mohan Matthen, Oxford University Press, Oxford, pp. 205-221.

Lovaas, Ole Ivar (1987), "Behavioral treatment and normal educational and intellectual functioning in young autistic children." *Journal of consulting and clinical psychology*, 55, 1, pp. 3-9.

Lycan, William G. (ed.) (1996), *Consciousness and Experience*, MIT Press, Cambridge, Massachusetts.

Marks, Isaac (1979), "Exposure Therapy for Phobias and Obsessive-Compulsive Disorders," *Hospital Practice*, 14, 2, pp. 101-108, DOI: 10.1080/21548331.1979.11707486.

Marr, David (1982), *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, W. H. Freeman, San Francisco.

Matthen, Mohan (2005), *Seeing, Doing, and Knowing: A Philosophical Theory of Sense Perception*, Oxford University Press, Oxford.

McGurk, Harry and John MacDonald (1976), "Hearing lips and seeing voices," *Nature*, 264, 5588, pp. 746-748.

Michotte, Albert (1946/63), *The Perception of Causality*, Methuen, London.

Morrot, Gil, Frédéric Brochet, and Denis Dubourdieu (2001), "The Color of Odors," *Brain and Language*, 79, pp. 309-320.

Nanay, Bence (2013), *Between Perception and Action*, Oxford University Press, Oxford.

Nudds, Matthew (2009), “Sounds and Space,” in *Sounds and perception: New philosophical essays*, ed. by Matthew Nudds and Casey O’Callaghan, Oxford University Press, Oxford, pp. 69-96.

— (2010a), “What Are Auditory Objects?” *Review of Philosophy and Psychology*, 1, 1, pp. 105-122.

— (2010b), “What Sounds Are,” in *Oxford Studies in Metaphysics: Volume 5*, ed. by Dean Zimmerman, Oxford University Press, pp. 279-302.

— (2014), “Is audio-visual perception ‘amodal’ or ‘crossmodal’?” In *Perception and its modalities*, ed. by Stephen Biggs, Mohan Matthen, and Dustin Stokes, Oxford University Press, New York, chap. 6, pp. 166-188.

O’Callaghan, Casey (2007), *Sounds*, Oxford University Press, Oxford.

— (2008), “Seeing What You Hear: Cross-Modal Illusions and Perception,” *Philosophical Issues*, 18, 1, pp. 316-338, DOI: 10.1111/j.1533-6077.2008.00150.x.

— (2009), “Sounds and Events,” in *Sounds and Perception: New Philosophical Essays*, ed. by Matthew Nudds and Casey O’Callaghan, Oxford University Press, pp. 26-49.

— (2011), “XIII—Hearing Properties, Effects or Parts?” *Proceedings of the Aristotelian Society*, 111, 3, pp. 375-405, DOI: 10.1111/j.1467-9264.2011.00315.x.

— (2014), “Intermodal Binding Awareness,” in *Sensory Integration and the Unity of Consciousness*, ed. by David J. Bennett and Christopher S. Hill, MIT Press, pp. 73-103.

— (2015), “The Multisensory Character of Perception,” *Journal of Philosophy*, 112, 10, pp. 551-569, DOI: 10.5840/jphil20151121035.

— (2016), “Objects for Multisensory Perception,” *Philosophical Studies*, 173, 5, pp. 1269-1289, DOI: 10.1007/s11098-015-0545-7.

— (2017), “Grades of Multisensory Awareness,” *Mind and Language*, 32, 2, pp. 155-181, DOI: 10.1111/mila.12137.

O’Shaughnessy, Brian (2009), “The Location of a Perceived Sound,” in *Sounds and perception: New philosophical essays*, ed. by Matthew Nudds and Casey O’Callaghan, Oxford University Press, Oxford, pp. 111-125.

Palmer, Stephen E. (1999), *Vision Science: Photons to Phenomenology*, MIT Press, Cambridge, Massachusetts.

Pasnau, Robert (1999), “What is Sound?” *Philosophical Quarterly*, 49, 196, pp. 309-324.

Pick Jr., Herbert L., David H. Warren, and John C. Hay (1969), “Sensory conflict in judgements of spatial direction,” *Attention, Perception, & Psychophysics*, 6, 4, pp. 203-205.

Pylyshyn, Zenon W. (2007), *Things and Places: How the Mind Connects with the World*, MIT Press.

Rosen, Gideon and Cian Dorr (2002), “Composition as a Fiction,” in *The Blackwell Companion to Metaphysics*, ed. by Richard Gale, Blackwell, pp. 151-174.

Sekuler, Robert, Allison B Sekuler, and Renee Lau (1997), “Sound alters visual motion perception,” *Nature*, 385, 6614, p. 308.

Shams, Ladan, Yukiyasu Kamitani, and Shinsuke Shimojo (2000), “What you see is what you hear,” *Nature*, 408, 6814, p. 788.

— (2002), “Visual illusion induced by sound,” *Cognitive Brain Research*, 14, 1, pp. 147-152.

Sider, Theodore (2013), “Against Parthood,” *Oxford Studies in Metaphysics*, 8, pp. 237-293.

Smith, Barry C. (2015), “The Chemical Senses,” in *Oxford Handbook of Philosophy of Perception*, ed. by Mohan Matthen, Oxford University Press, Oxford, pp. 314-352.

Sorensen, Roy (2008), “Hearing Silence: the Perception and Introspection of Absences,” in *Sounds and Perception: New Philosophical Essays*, ed. by Matthew Nudds and Casey O’Callaghan, Oxford University Press, Oxford, pp. 126-145.

Soteriou, M. (2018), “Sound and illusion,” in *Perceptual Ephemera*, ed. by Thomas Crowther and Clare Mac Cumhaill, Oxford University Press, Oxford, pp. 31-49.

Spelke, Elizabeth (1990), “Principles of Object Perception,” *Cognitive Science*, 14, pp. 29-56.

Spence, Charles (2015), “Eating with our ears: Assessing the importance of the sounds of consumption to our perception and enjoyment of multisensory flavour experiences,” *Flavour*, 4 (Jan. 2015), p. 3, DOI: 10.1186/2044-7248-4-3.

Spence, Charles, Barry Smith, and Malika Auvray (2014), “Confusing tastes and flavours,” in *Perception and its Modalities*, ed. by Dustin Stokes, Mohan Matthen, and Stephen Biggs, Oxford University Press, Oxford, chap. 10, pp. 247-274.

Tarr, Michael J. and Heinrich Bülthoff (1998), *Object Recognition in Man, Monkey, and Machine*, MIT Press, Cambridge, Massachusetts.

Treisman, Anne M. (1998), “Feature binding, attention and object perception,” *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, 353, pp. 1295-1306.

— (2003), “Consciousness and Perceptual Binding,” in *The Unity of Consciousness*, ed. by Axel Cleeremans, Oxford University Press, pp. 95-113.

Vatakis, Argiro and Charles Spence (2007), “Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli,” *Perception & psychophysics*, 69 (Aug. 2007), pp. 744-56, DOI: 10.3758/BF03193776.

de Vignemont, Frédérique (2014), “Multimodal Unity and Multimodal Binding,” in *Sensory Integration and the Unity of Consciousness*, ed. by David J. Bennett and Christopher S. Hill, MIT Press, Cambridge, Massachusetts, chap. 6, pp. 125-150.

Vroomen, Jean and Beatrice de Gelder (2000), “Sound enhances visual perception: cross-modal effects of auditory organization on vision,” *Journal of Experimental Psychology: Human Perception and Performance*, 26, 5, pp. 1583-1590.

Vroomen, Jean and Beatrice de Gelder (2004), “Ventriloquism and the Freezing Phenomenon,” in *The Handbook of Multisensory Processes*, ed. by G. A. Calvert, C. Spence, and B. E. Stein, MIT Press, Cambridge, MA, pp. 141-150.

Wagemans, Johan, James H. Elder, Michael Kubovy, Stephen E. Palmer, Mary A. Peterson, Manish Singh, and Rüdiger von der Heydt (2012), “A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization,” *Psychological bulletin*, 138 6, pp. 1172-1217.

Wagemans, Johan, Jacob Feldman, Sergei Gepshtein, Ruth Kimchi, James R. Pomerantz, Peter A van der Helm, and Cees van Leeuwen (2012), “A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations.” *Psychological bulletin*, 138 6, pp. 1218-1252.

Wertheimer, Max (1938), “Laws of organization in perceptual forms.,” in *A source book of Gestalt psychology*, ed. by W. D. Ellis, Kegan Paul, Trench, Trubner & Company, pp. 71-88, DOI: 10.1037/11496-005.

Young, Nick and Bence Nanay (2020), “Audition and Composite Sensory Individuals,” in *Sensory Individuals*, ed. by Aleksandra Mroczko-Wąsowicz and Rick Grush., Oxford University Press, Oxford.

Zacks, Jeffrey M., Nicole Speer, Khena Swallow, Todd Braver, and Jeremy Reynolds (2007), “Event Perception: A Mind/Brain Perspective,” *Psychological bulletin*, 133, pp. 273-93, DOI: 10.1037/0033-2909.133.2.273.

Zacks, Jeffrey M. and Barbara Tversky (2001), “Event structure in perception and conception,” *Psychological Bulletin*, 127, 1, pp. 3-21.

Zmigrod, Sharon, Michiel Spapé, and Bernhard Hommel (2009), “Intermodal event files: integrating features across vision, audition, taction, and action,” *Psychological Research*, 73, 5, pp. 674-684.