Second Edition

Emotion Measurement

Edited by Herbert L. Meiselman

ELSEVIER

Emotion Measurement

Second Edition

Edited by

Herbert L. Meiselman

ELSEVIER

WP
WOODHEAD
PUBLISHING
An imprint of Elsevier

Woodhead Publishing is an imprint of Elsevier
The Officers' Mess Business Centre, Royston Road, Duxford, CB22 4QH, United Kingdom
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States
The Boulevard, Langford Lane, Kidlington, OX5 1GB, United Kingdom

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-0-12-821125-0

For information on all Woodhead Publishing publications visit our website at https://www.elsevier.com/books-and-journals

Publisher: Charlotte Cockle
Acquisitions Editor: Megan R. Ball
Editorial Project Manager: Andrea Dulberger
Production Project Manager: Surya Narayanan Jayachandran
Cover Designer: Mark Rogers

Typeset by TNQ Technologies

Chapter 2

Navigating the science of emotion

Lisa Feldman Barrett[a,b], **Christiana Westlin**[a]
[a]*Department of Psychology, Northeastern University, Boston, MA, United States;* [b]*Department of Psychiatry, The Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital/Harvard Medical, Charlestown, MA, United States*

Chapter outline

Abstract
The science of emotion is muddled with a multitude of hypotheses and assumptions about what emotions are and how they work. The standard narrative framework used in the science of emotion conceals meaningful variation within scientific traditions and obfuscates similarities, depriving both newcomers and scientists of effective conceptual tools for making scientific progress. In this chapter, we introduce a different narrative framework to reorganize the field according to the degree to which variation is hypothesized across instances that belong to the same emotion category (e.g., instances of anger), similarities across instances of different categories (e.g., instances of anger vs. instances of fear) and the causes of this variation. The newer narrative framework will help readers to make more informed measurement decisions, equipping them to properly evaluate theoretical assumptions and understand the consequences of their measurement choices.

Emotion Measurement. https://doi.org/10.1016/B978-0-12-821125-0.00030-3

2.1 Preliminary observations

Those who are new to the science of emotion have a common experience: when they crack open a textbook or review chapter, they are confronted with a *perplexing number* of claims and ideas about what emotions are and how they work; these ideas vary a great deal from one another in almost every way imaginable; theories disagree on the details of how an instance of emotion should be defined, on where to draw the boundaries for what counts as an emotion and what does not, on which emotion categories matter, on how emotions are different from related concepts like mood, reward, and motivation, and on how various phenomena such as facial movements, physiological changes, feelings and motor actions should be treated.

In an attempt to bring some order to this dizzying cornucopia of ideas and speculations, writers have created various particular narrative frameworks in which ideas (or "theories") are assembled into a few broader groupings, or categories, much like those that appear in this volume: basic emotion, appraisal and dimensional. By grouping some hypotheses together into a category, while separating others into different categories, writers attempt to digest variation in the hypotheses, inviting you (the reader) to ignore certain distinctions (i.e., the different assumptions and ideas within the same category) and to focus your attention on other distinctions (those that differentiate the categories). Any category is a grouping of items, events, objects, and even hypotheses about the nature of emotion, that are treated as similar *for some purpose* (Murphy, 2002). So, a category of scientific views is a grouping of hypotheses that are deemed similar for some purpose. In any science, the organizing principle for grouping ideas—the purpose—is determined by the writer's goal.

For the past century or so, the goal of many scholars has been to argue over whether certain emotion categories are natural kind categories with firm boundaries in nature (Barrett, 2006a, 2006b). Scientists seem unable to agree on which emotion categories are "basic" or exactly what the criteria for "basicness" is (for a discussion, see Ortony & Turner, 1990), but for the most part, the question of "basicness" can be translated into three related hypotheses: Is a given emotion category a collection of instances that are similar enough to one other in their features that the presence or absence of an instance of that emotion category be diagnosed in a unique, coordinated suite of observable changes in autonomic physiology, facial movements, body postures, vocalizations, and so on? Does the suite of changes issue from a common mechanism that evolved to deal with recurring challenges to our ancestors' ability to reproduce (i.e., fitness-related threats or opportunities)? And is the category part of a universal human nature (i.e., all neurotypical humans, barring illness, are born with the capacity to produce these emotional instances, and correspondingly, to recognize them in others)? This is a fairly standard narrative framework of theoretical views on the science of emotion which cleaves the various hypotheses into three categories – "basic, appraisal, and dimensional" — or sometimes two – "discrete versus dimensional" approaches. In our view, this standard narrative constitutes one of the largest barriers to scientific progress (Barrett, 2006a), leaving both newcomers and seasoned scientists ill-equipped to make informed decisions about how to measure instances of emotion.

First and foremost, the standard narrative framework is muddled by conceptual errors. When using it, writers typically refer to "emotions," thereby failing to distinguish between instances of emotion and categories of emotion (i.e., ignoring the token vs. type distinction in philosophy). Referring simply to "emotions" does not clarify whether a hypothesis refers to a specific instance or a group of instances. Of course, if you hypothesize that emotion categories are natural kinds, such that each instance within a category is highly similar to every other, both in observable, measurable features (i.e., the physiological changes, movements, vocalizations, feelings, and other properties) and in underlying causes (e.g., a neural circuit or pattern), then this failure to distinguish between a category (the type) and its individual instances (the tokens) is not much of a problem. But the distinction between tokens and types has always been important in the science of emotion, going all the way back to William James. Some theoretical frameworks, such as the theory of constructed emotion (which we discuss later in this chapter), hypothesize considerable variation in the instances of an emotion category (e.g., instances of anger, fear, sadness and happiness can be either pleasant or unpleasant; Wilson-Mendenhall, Barrett, & Barsalou, 2015; Harmon-Jones, Harmon-Jones, Abramson, & Peterson, 2009), as well as similarity in the features across categories (e.g., stereotypic instance of anger and fear categories in Western cultures are unpleasant and high in arousal; Barrett & Bliss-Moreau, 2009; Russell, 2003). As a consequence, the failure to honor the token-type distinction already makes it difficult to ask certain questions and test certain hypotheses, causing much confusion in the scientific literature on emotion.

Second, the standard narrative framework often characterizes the basic emotion and appraisal hypotheses as proposing that emotions are "discrete," meaning that emotion categories have discrete boundaries, which is then contrasted with the opposing hypothesis that emotional instances are best understood using "dimensions." This "discrete versus dimensional" dichotomy is an error, however. The so-called dimensional hypothesis is better understood as "constructionist" (again, this approach is described in some detail below (also see Barrett, 2017b for a review of constructionist approaches and Barrett & Russell, 2015a, 2015b for a review of psychological construction, which is one constructionist approach to emotion). Constructionist approaches do hypothesize the existence of emotion categories, but the nature of the categories are thought to differ both in terms of the variation in their features and in their underlying causes from the so-called "discrete" emotion hypotheses. Referring to constructionist hypotheses as "dimensional" occurs because a constructionist approach to the science of emotion advocates for the *necessity* of measuring the core affective features of emotional instances, such as the properties of "affect," i.e., general feelings of valence (pleasure and displeasure) and arousal (quiescent and still to highly activated); no theoretical view suggests that these features are sufficient for describing emotional instances. And the suggestion is that affective features should be measured on a continuum, rather than as present or absent (i.e., categorically), because it is hypothesized that they vary across the instances of the same emotion category and are often similar across instances of different categories (e.g., Touroutoglou, Lindquist, Dickerson, & Barrett, 2015; Wilson-Mendenhall, Barrett, & Barsalou, 2013; Wilson-Mendenhall et al., 2015). The mistaken assumption, then, is that a constructionist approach reduces all instances of emotion to affective dimensions (e.g., Cowen & Keltner, 2017). Sometimes writers go on to criticize constructionism for its inability to distinguish different emotion categories from one another when their stereotypes share the same affective features (e.g., in the standard narrative, for example, anger, fear, disgust and guilt are characterized as a high-arousal, negative emotion categories, as are a variety of other emotion categories), but, again, this is confusing necessity with sufficiency. Scientists who subscribe to a constructionist philosophy hypothesize that affective features, including valence and arousal, are necessary to best measure and describe an instance of emotion but do not propose that they are sufficient. In fact, a careful read of the literature reveals that no constructionist treatment of emotion has ever hypothesized that emotional instances can sufficiently be reduced to or *explained by* valence and arousal. Mischaracterizing constructionism as "dimensional" and then criticizing it for the limitations of a "dimensional approach" is not helpful to scientific clarity.[a] Nonetheless, references to the "dimensional" approach continue to appear in newly published papers on emotion, no matter how many times this error is explained.

(A further error in the standard narrative is that it characterizes valence and arousal as causal processes or mechanisms when the constructionist hypotheses, themselves, do not. Valence and arousal are descriptive properties, not mechanisms that cause anything. Neither are they unique to emotion (valence

a. The affective circumplex (Barrett, 2004; Russell, 1980; Russell & Barrett, 1999; see chapter: Theoretical Approaches to Emotion and Its Measurement, Fig. 2.2, which can also be found in Barrett & Russell, 1999, Fig. 2.1) is not an explanatory theory of emotion. It is a low dimensional, descriptive map that represents two properties or features of emotional instances. These are properties or features of experience, valence and arousal, themselves cannot be mechanistically reduced, and are emergent properties of more basic processes. One well-known constructionist hypothesis by Russell (2003) is agnostic on how affect (described as feelings of valence and arousal) is caused. Another approach, our own Theory of Constructed Emotion, formerly the Conceptual Act Theory (Barrett & Bliss-Moreau, 2009; Barrett et al., 2015), makes very specific hypotheses about how affect arises from more fundamental mechanisms in the brain and body.

and arousal are fundamental features of all thoughts, beliefs, memories, perceptions, and so on; simply put, they are descriptive features of consciousness; for a discussion, see Barrett & Bliss-Moreau, 2009; Wundt, 1897/1998).

In addition to the errors embedded in the standard framework, the very act of grouping various scientific views into categories does not do justice to their similarities and differences. The framework obscures some important variations within each theoretical category. For example, consider the "appraisal" category. One variety of appraisal views, which has been called *causal* or *classical appraisal views* (Barrett, Mesquita, Ochsner & Gross, 2007; Gross & Barrett, 2011), assumes that "appraisals" are literal cognitive mechanisms that cause the subjective evaluations that, in turn, either cause or constitute instances of emotion (Arnold, 1960a, 1960b; Frijda, 1986; Lazarus, 1966; Roseman, 2011; Sander, Grandjean, & Scherer, 2005; Scherer, 2005; see chapter: Theoretical Approaches to Emotion and Its Measurement). But another variety of appraisal views characterize "appraisals" as descriptive features, much like valence and arousal; for example, if an instance of emotion is associated with the appraisal of "novelty," then during that emotional instance, something is experienced as novel. This descriptive variant of the appraisal category (Clore & Ortony, 2000, 2008; Ortony & Clore, 2015) is agnostic on the mechanisms that produce the appraisals. So, whereas causal appraisal views would assume that an experience of novelty is caused by a literal novelty-detector in your brain (just as they mistakenly assume that the experience of pleasure and displeasure is caused by a literal valence detector in your brain), descriptive appraisal approaches don't assume that there is a single mechanism that causes novelty (or valence). These *constitutive* or *constructive appraisal views*, like constructionist views, do not assume that there is any parallelism between a mechanism and the resulting experience (i.e., there is no anger mechanism causing instances of anger, no "goal relevance" mechanism evaluating goal relevance so that a person experiences a situation as relevant to his or her goals, and so on).

The standard narrative framework not only conceals meaningful variation within a category of hypotheses about emotion, but it also obfuscates similarities across the categories. Perhaps the most important similarity is that basic emotion hypotheses and causal appraisal hypotheses both make *essentialist assumptions* about the nature of emotion. Essentialism is the belief that a category of instances referred to by the same word (such as anger, pride, awe, etc.) or phrase (e.g., "basic emotion hypotheses," "appraisal hypotheses," etc.) share a deep, immutable causal mechanism that gives them their nature (this is essentialism as described by John Locke). For example, both groupings assume that a psychological phenomenon is caused by a dedicated mechanism of the same name (e.g., in basic emotion views, an instance of fear is caused by a "fear" mechanism; in causal appraisal views, the experience of novelty is caused by a "novelty" mechanism). Furthermore, both groupings hypothesize a specific, dedicated underlying causal mechanism for each emotion category, either a population of dedicated neurons (for a review, see Tracy & Randles, 2011) or a particular configuration of appraisals (Scherer, 2009; see chapter: Theoretical Approaches to Emotion and Its Measurement), or a hypothetical mechanism, such as an affect program (Ekman & Cordaro, 2011; Tomkins & McCarter, 1964), which is an example of psychological essentialism (Medin & Ortony, 1989).

Basic emotion and causal appraisal views also share another version of essentialism: the belief that a group of instances share a "fingerprint" (i.e., a pattern of features that are frequent or typical in instances of the category so that there is one best "instance" that can be used to diagnose the presence of the category (like a prototype; this is the sort of essentialism advocated by Plato), such as a specific, synchronized pattern of measurable changes in the face, in the body, in behavior, etc.).[b] The pattern (the Platonic essence) is supposedly caused by the Lockean essence – the dedicated emotion circuit (in basic emotion views) or by the pattern of appraisals (in causal appraisal views).

It is important to mention that, in principle, these hypotheses allow for variation within an emotion category. A fingerprint can vary from one instance to another because of the oils and substances on your fingertips, the temperature of your skin, and the surfaces you touch, even though the underlying ridges on your skin are constant. Similarly, basic emotion and causal appraisal views allow for variation in movements of the face, in electrical signals of the autonomic nervous system, in acoustical changes of the voice, in voluntary movements of the body, and so on. Some randomness is expected, and other processes, independent of an instance of emotion itself, are also thought to account for variation, such as "display rules" or other regulatory strategies, such as suppression (Ekman & Cordaro, 2011; Gross, 2015; Matsumoto, Keltner, Shiota, Frank, & O'Sullivan, 2008; Roseman, 2011; Tracy & Randles, 2011). And even if the hypothesis of an emotion category's fingerprint is explicitly disavowed by writers, in practice they hypothesize that each category has a pattern that can be used to uniquely "recognize" its instances in the same way that a fingerprint uniquely identifies an individual person.[b] Appraisal views, moreover, were originally motivated to account for variation in emotional phenomena, and while in principle they acknowledge the

b. Consider the idea of a "fingerprint." The pad of your finger has ridges which do not change from instance to instance. The sweat, dirt, ink, or other substances on the ridges of your finger are transferred to the surface of a brass door handle, a wooden table, a piece of paper, or whatever you touch, leaving a print. Your fingerprint won't look exactly the same each time you touch a surface. Sometimes you might grip the handle with more pressure or less. Some surfaces are rough with strong traction, while others are smoother and allow your fingers to slip a bit. Sometimes your skin might be warmer and more pliable. Even though your fingerprint does not look identical each time, it looks similar enough and is unique to you, and only you, so that it can be used to identify you but not other people.

likelihood of such variation, their theorizing and research tends to focus on the presumed "basic" categories (e.g., Barrett, Ochsner, & Gross, 2007).

In science, essentialist assumptions often persist, even when scientists are unaware of endorsing them and despite accumulating evidence that they are false, most notably because of a phenomenon called "psychological essentialism" (Medin & Ortony, 1989); this form of essentialism allows people to posit a hypothetical or unseen cause in the absence of any evidence of what the cause might be (e.g., the "affect program" concept) (Ekman & Cordaro, 2011; Tomkins & McCarter, 1964). Psychological essentialism inoculates believers against disconfirming evidence, allowing them to continue to theorize about and believe in the existence of emotion essences, and to focus on the small proportion of published studies that support their existence, despite the even larger number of studies that disconfirm them (for a review, see Barrett, 2017a; for some examples, see Azari et al., 2020; Barrett, 2006a, 2012, 2013, 2018; Barrett, Adolphs, Marsella, Martinez, & Pollak, 2019; Barrett & Finlay, 2018; Barrett, Lindquist, et al., 2007; Brooks et al., 2017; Brooks, Chikazoe, Sadato, & Freeman, 2019; Clark-Polner, Johnson, & Barrett, 2017; Guillory & Bujarski, 2014; Gendron, Crivelli, & Barrett, 2018; Gendron et al., 2020; Hoemann et al., 2019, 2020; Jackson et al., 2019; LeDoux, 2015; Lindquist, Wager, Kober, Bliss-Moreau, & Barrett, 2012; Siegel et al., 2018; Touroutoglou, Lindquist, Dickerson, & Barrett, 2015).

The errors that are woven throughout the standard narrative framework in the science of emotion, the important differences that are obscured by grouping hypotheses together into theoretical categories, and the resulting essentialism that is often implicit or ignored, leave the science of emotion with an ineffective organizational framework for making scientific progress. Without a meaningful roadmap, newcomers find it difficult to identify what is known about emotional phenomena with any degree of certainty. Scientists themselves are paralyzed in a "disconfirmation dilemma" (Greenwald & Ronis, 1981) that makes accumulating knowledge about emotion almost impossible (except within theory-based silos that selectively review the evidence and fail to consider the broader empirical landscape). The linguist George Lakoff has called emotion an *essentially contested concept*: everyone agrees that emotions exist, but a variety of meanings are simultaneously employed, and scientific inquiry seems unable to settle the matter. Indeed, the science of emotion is struggling today with the very same dilemmas as it was a century ago (Barrett, 2017b; Gendron & Barrett, 2009). And the status of the empirical literature today is not that different: there are some experiments that support the basic emotion and causal appraisal hypotheses, but these are the tip of a much larger iceberg of scientific evidence that does not. As a consequence, despite tremendous investments of research time and money, emotions remain mysterious and deeply perplexing.

This confusion comes with a high price tag: studies don't replicate. Measures don't work as expected (i.e., they don't hang together as coordinated suites and they don't predict behavior as well as they could, or by some accounts, at all) and there is a persistent confusion about how to measure instances of emotion. This situation indicates the immediate need for a different scientific approach to explain what emotion categories are and how to measure emotional instances.

In this chapter, we offer a different approach—a narrative framework to reorganize the field according to whether or not hypotheses assume the existence of emotion essences. Hypotheses about emotion vary in the extent to which they indulge in essentialism and in the type of essentialism they incorporate, if at all, and understanding these distinctions provides a more useful roadmap for the science of emotion, including ways to measure emotional instances. Denying the existence of emotion essences does not mean denying the existence of emotions (Barrett, 2012). Instead, it means acknowledging the importance of tremendous variation in emotional life, whether across cultures, across individuals or even within an individual across contexts. It means designing experiments that make it possible to observe this variation both within and across emotion categories. It also means testing hypotheses about the causes of emotional instances without the need for emotion essences.

For the remainder of the chapter, we introduce and discuss this newer narrative framework — identifying whether a theoretical view is assuming essentialism or not —to allow you to make sense of the rest of this volume's chapters in terms of the underlying assumptions that are so rarely revealed and openly evaluated. Hopefully, this will equip you to properly evaluate the remaining chapters in this volume, as well as understand the consequences of the assumptions (whether they are explicitly stated or not). Within the next sections, we not only consider the various theoretical approaches and their degree of essentialism embedded within, but we also consider their general measurement models. The chapter ends with a discussion of pattern classification, an analytic technique that is touted as the way to identify "fingerprints," "signatures," or "biomarkers" of various emotion categories. We discuss how such claims are tainted by the lure of essentialism, and how the solutions from pattern classification are variable across studies and do not match those of unsupervized clustering, which is a data-driven approach to discovering structure in the data when no emotion category labels are applied to guide the analysis. Hopefully, this will give you a firmer footing on which to make informed measurement decisions.

Our goal is not to convince you that essentialism is detrimental. As you might have surmised, we believe it is. Essentialism has been shown to interfere with scientific thinking, particularly when it comes to evolution and natural selection (Gelman & Rhodes, 2012) and emotion (Barrett, 2017a). Over a century ago, William James (1890/1998) pleaded for psychology to abandon essentialism: "The trouble with the emotions in psychology is that they are regarded too much as … psychic entities, like the old immutable species in natural history" (p. 449). Essentialism is so powerful, in fact, that is has, ironically, led the field to completely misinterpret James's ideas as essentialist in nature (Gendron & Barrett, 2009). Our goal in this chapter is more modest, however. It is to make you more aware of your own assumptions, as well as their consequences of your measurement choices.

2.2 Typologies or populations? That is the question

The organizational framework we offer begins with the observation that any instance of emotion can be described (and measured) as a variety of features: *physical* features (such as patterns of expressive facial movements, vocal acoustics, autonomic nervous system changes, and neural activity), *affective* features that capture what the instance feels like (e.g., how pleasant or unpleasant the episode feels, how arousing it feels; Barrett & Bliss-Moreau, 2009; Russell & Barrett, 1999), *appraisal* features that refer to how the situation is experienced (e.g., whether the situation is experienced as novel or familiar, as conducive to one's immediate goals or not, and so on; Barrett, Mesquita et al., 2007; Clore & Ortony, 2008, 2013; Scherer, Mortillaro, & Mehu, 2017) and *functional* or *goal-based* features that refer to the goals that a person is attempting to meet (e.g., to avoid a predator, to get closer to someone, to win a competition, etc.; e.g., Adolphs, 2017; Lazarus, 1966/1991).

An emotion category, then, is a grouping of emotional episodes that share a feature or set of features in common. In common English parlance, people refer to "an emotion" as if anger, happiness, or any emotion word refers to a group of events that are highly similar in their features on most occasions. But an emotion word refers to a category of instances that can and do vary from one another in their physical and mental features, including their functional or goal-based features. Few scientists who study emotion take the view that every instance of an emotion category, such as anger, is identical to every other instance across situations, people, and cultures. Nonetheless, there is a considerable scientific debate about the extent of the within-category variation, the specific features that vary, the causes of the within-category variation, and implications of this variation for the nature of emotion. The similarities and differences in the various basic, appraisal, constructionist, and other theoretical approaches to emotion (including functionalist approaches) can be understood in terms of the sources and magnitudes of this variation.

Many debates about the nature of emotion boil down to disagreements about the nature of the similarities shared by instances of the same emotion category and the degree of variation in the relevant features, as well as potential similarities and differences in features across emotion categories. These debates regarding the source and magnitude of variation in the instances of an emotion category are summarized by two dimensions, presented in Fig. 2.1, and offer some guidance for consumers of emotion research who are focused on the practical issue of whether emotion categories can be diagnosed or

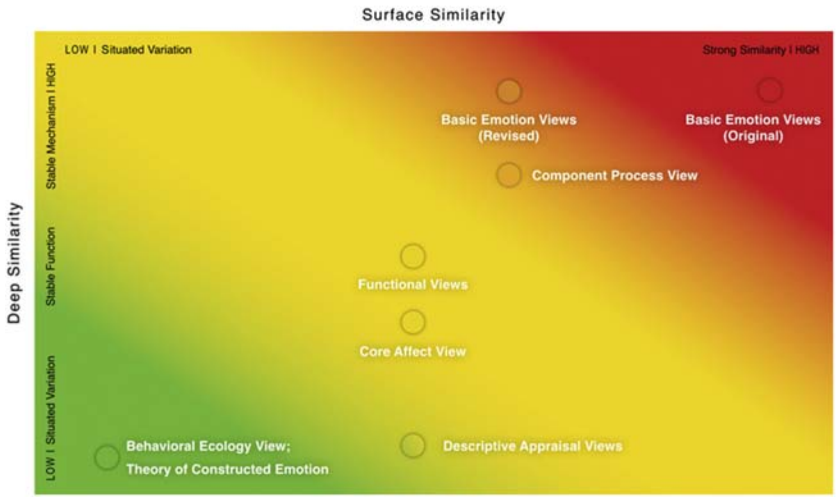

FIG. 2.1 Explanatory frameworks in emotion science. Surface similarity: hypotheses about the degree to which instances of an emotion category vary in their observable features. Deep similarity: hypotheses about the similarities in the mechanisms that cause instances of the same emotion category (e.g., the neural circuits or assemblies that cause instances of the same emotion category). The colors represent the type of emotion categories proposed: ad hoc, abstract categories (green zone); prototype or theory-based categories (yellow zone); classical or natural-kind categories (red zone). Examples of these theoretical hypotheses are: Ekman, 1972 (Basic Emotion Views, Original); Cowen & Keltner, 2017 (Basic Emotion Views, Revised); Scherer et al., 2017 (Component Process View); Adolphs, 2017 (Functional Views); Russell, 2003 (Core Affect View); Clore & Ortony, 2013 (Descriptive Appraisal Views); Fridlund, 2017 (Behavioral Ecology View); Barrett, 2017a (Theory of Constructed Emotion). *Reprinted with permission from Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements.* Psychol. Sci. Publ. Interest, 20*(1), 1–68.*

distinguished by physical measurements with sufficient regularity and distinctiveness so that it is possible to read emotion in a person's body, face, voice or brain. The two dimensions, together, characterize the degree to which a theoretical view on the nature of emotion categories assumes essentialism.

In fact, Fig. 2.1 summarizes various theoretical hypotheses about the nature of emotion categories according to the degree of essentialism they embody: the horizontal axis arrays the degree to which a theoretical view hypothesizes that instances of an emotion category vary in their observable physical features (surface similarities, akin to a Platonic essence). The vertical axis arrays the degree to which theoretical views hypothesize that instances of a category share the same (or similar) causal mechanisms (deep similarities, akin to a Lockean essence, such as dedicated neural circuits (Tracy & Randles, 2011), a hypothetical affect program (Ekman & Cordaro, 2011), or a set of emotion-specific computations (Bach & Dayan, 2017). The colors within the body of

Fig. 2.1 represent the type of emotion categories proposed by a particular theoretical view, according to established proposals about concepts and categories in cognitive science. An emotion concept is a mental representation of an emotion category. Theoretical hypotheses about the proposed degree of variation in surface and deep features of an emotion category, therefore, strongly relate to the proposed nature of emotion categories and their relation to emotion concepts.

The red zone captures theoretical views on emotion which hypothesize that emotion categories are classical or natural-kind categories (also called Aristotelian categories). A classical category is a grouping of instances that share observable, perceptual features (a classical emotion category, for example, will be formed by instances who share the same facial expression, bodily changes, and so on). The instances are thought to be caused by a shared mechanism, such as dedicated neural circuits (Tracy & Randles, 2011), a hypothetical affect program (Ekman & Cordaro, 2011), or a set of emotion-specific computations (Bach & Dayan, 2017). The associated concept is thought to be a single representation consisting of a dictionary definition of necessary and sufficient features. Accordingly, a classical view of emotion categories assumes that each category can be diagnosed by a single coordinated pattern of physical features (e.g., a facial expression, a physiological change, a pattern of neural activity, etc.) that is a valid cue to the presence of that category in any and all circumstances, barring error. Within-category feature variation is usually explained, post hoc, by hypothesizing phenomena that are independent of the emotional instance itself, such as display rules, cultural dialects, regulation strategies, or stochastic variation (Ekman & Cordaro, 2011; Elfenbein, 2013, 2017; Matsumoto, 1990; Matsumoto et al., 2008; Tracy & Randles, 2011). Basic emotion views (e.g., Ekman, 1972; Tracy & Randles, 2011) and the discrete emotion approach (e.g., Izard, 2007, 2010; Izard, Woodburn & Finlon, 2010), in their original form, described emotion categories as classical categories. Typically, there is more scientific emphasis on finding fingerprints than on identifying a single causal mechanism, because fingerprints should always be measurable, whereas essences can be hidden.

The idea that most categories are classical in structure dominated science and philosophy from antiquity but was replaced in the 1970s by the hypothesis of prototype categories, prompted by observations that the instances of a category vary from one another in their features. For example, consider all the things you do when you are angry: you might tremble, freeze, scream, withdraw, attack, cry, and even laugh or joke. Some of which are more frequent or more typical (meaning that the instance has a majority of the features of a category). To deal with this variation, scientists proposed that emotion categories each have a most typical or frequent instance (a prototype) which possesses a common set of features, and other category instances are graded in their similarity to the prototype. Or the prototype might be a belief that describes the most typical category instance (e.g., Clore & Ortony, 1991).

The prototype is the category's concept. Theoretical views on emotion which hypothesize that emotion categories are prototype categories can be found in the yellow zone.

Interestingly, at one time or another, versions of basic emotion (e.g., Cowen & Keltner, 2017), appraisal (e.g., Scherer, 1999), functional (e.g., Adolphs, 2017) and constructionist (Russell, 1991, 2003) views have all proposed that emotion categories are structured as prototype categories. Some prototype views hypothesize much more variation in the physical features around the prototype of each category (e.g., Adolphs, 2017; Russell, 1991, 2003) than do others (e.g., Cowen & Keltner, 2017) and functional views propose that the instances of each category shares a universal (and species-general) functional similarity across situations (e.g., fear is the desire to escape from a predator; Adolphs, 2017; Adolphs & Anderson, 2018; Campos, Mumme, Kermoian, & Campos, 1994). Prototype proposals also differ in their hypotheses about the underlying causes of emotional instances within the same category. In Adolphs & Anderson's functional view, for example, the functional feature that unites the instances of an emotion category can be similar across individuals and species but can vary in the neural mechanisms that produce the instances. In the revised basic-emotion case, within-category feature variation around the prototype is usually explained, post hoc, by hypothesizing the same sorts of processes as for the classical view. In such cases, the surface, physical features are assumed to be valid cues in many, but not all, circumstances, to the presence of the category. That is, each instance of an emotion category is hypothesized to share enough of a characteristic pattern in its physical features that is consistently present and recognizably different from the patterns found in other emotion categories (for specific quotations, see Cordaro et al., 2018; Ekman, 1992, p. 550; Ekman & Cordaro, 2011, p. 364; Levenson, 2011, p. 379; Scarantino & Griffiths, 2009, pp. 448−449).

When instances are grouped together into a category based on the similarity of their inferred, functional features, or goals, rather than similar in their physical features, they are referred to as *abstract* categories. Beginning in the early 1980s, the psychologist Larry Barsalou observed that abstract categories are formed in a *situation-specific* (i.e., ad hoc) way based on the function that the category serves in a particular situation (Barsalou, 1983, 1985, 2008; Barsalou, Simmons, Barbey, & Wilson, 2003). For example, in playful situations, a person might construct the category 'things that fly' with balls, Frisbees, kites, and darts; in situations that require travel, the same category might include an airplane, hot air balloon, and helicopter. In a park, the category will contain birds, bats, bees, and squirrels. The concept for an ad hoc abstract category is the most representative instance (i.e., the prototype) that best describes the function of a category *in a given situation*. The prototype is situated and therefore changes with context; it need not exist in nature at all − it could just be the ideal instance that satisfies the function of a category in that situation. Theoretical views on emotion which hypothesize that emotion categories are ad hoc, abstract goal-based categories (also called *conceptual categories*) can be found in the green zone. These views hypothesize that, in a given instance, similarity in both surface and deep features is conditioned on the situation, such that both the surface and deep features of an emotion category will show considerable variation from situation to situation within a person, as well as across people within the same culture and also across cultures. Context is not considered a process that moderates or modifies deep and surface features that are universal, but rather is hypothesized to be intrinsic and meaningful (i.e., functional) for the variation that is observed (and measured).

Furthermore, these views hypothesize that instances within the same emotion category (e.g., happiness), instances across emotion categories (e.g., happiness vs. fear), and even instances of nonemotion categories such as thoughts, beliefs, perceptions, and so on, all emerge from the dynamic interactions of more fundamental *common* or *domain general processes* within the nervous system (e.g., Barrett, 2013, 2017a, b; Barrett & Satpute, 2013, 2019; Lindquist & Barrett, 2012). No emotion category (or, in fact, any mental category) is assumed to have a Lockean essence. No emotion category is presumed to be any more biologically basic than any other. The validity of these views depends on specifying the shared causal mechanisms, not on finding a single mechanism for each emotion category, or a stable pattern of observable consequences stemming from those mechanisms.

A prominent approach to emotion categories as conceptual categories can be found in *construction approaches* to emotion (e.g., Gendron & Barrett, 2009; Lindquist & Barrett, 2012). Constructionist approaches come in three flavors: social construction (how emotions are influenced by social roles and values), psychological construction (how emotions emerge from more basic psychological processes related to making meaning of affective feelings), and neuroconstruction (experience wires a brain for emotion during brain development). The theory of constructed emotion (Barrett, 2017a,b; formerly the Conceptual Act Theory; Barrett, Wilson-Mendenhall, & Barsalou, 2015; Barrett, 2014; Quigley & Barrett, 2014) integrates these three varieties of construction, along with the rational constructionism of developmental theory (Barrett, 2017b; Fedyk & Xu, 2018; Xu, 2016) to propose that instances of an emotion category are intrinsically constructed in a context-dependent way that has been learned and wired into the brain in a particular culture.

All the views described thus far are consistent with some version of evolutionary theory and derive inspiration from Charles Darwin (albeit from different books, making very different assumptions about the nature of biological categories; for discussion, see Barrett 2017a, b). Classical and prototype hypotheses about emotion categories are consistent with classical typological thinking, or the idea that types of something can be classified according to the common features shared by the instances of the category. Typological thinking about emotion categories is very much on display in Charles Darwin's *The Expression of the Emotions In Man and Animals*

(Darwin, 1872/1965). This is consistent with the evolutionary view which is termed the "modern evolutionary synthesis" in which genes are the only way for emotion-related information and capacities to be transmitted across generations. The hypothesis that emotion categories are conceptual categories, in contrast, is rooted in population thinking, or the idea that a biological category is populated with context-dependent, variable instances; any single representation of the entire category is assumed to be an abstraction, a stereotypic summary of the category. The proposal that emotion categories are goal-based, conceptual categories, derives from Darwin's use of population thinking in *On the Origin of Species* (1859/1964; see Mayr, 2004). The prototype of any category is thought to be context-dependent, and represents the ideal instance that best suits the function or goal of that category in a specific situation, whether or not it actually exists in nature (e.g., Barsalou, 1983; Voorspoels, Vanpaemel, & Storms, 2011). Correspondingly, the hypothesis is that the similarity of instances within an emotion category is not fixed or static but instead varies from situation to situation because the similarity of its instances is based on the goal that the instances serve in a particular situation at a particular moment in time. This is consistent with what has been termed the "extended evolutionary synthesis" in which culture, combined with the mechanisms of brain development, allow for the transmission of emotion-related knowledge and capacities across generations.

2.2.1 Measurement implications

When viewed through the framework presented in Fig. 2.1, the implications for measurement are straightforward. As one variant of a typological approach, classical or Aristotelian views of emotion categories hypothesize *strong reliability and specificity* in the physical features that can be measured in the instances of an emotion category, such as a pattern of facial movements (as a universal emotional expression) or a pattern of neural activity. Therefore, it should be possible to objectively read an emotional state in a person or non-human animal by measuring physical features alone, and this pattern should not vary much by individual or context.

As another variant of a typological approach, prototype views hypothesize that an emotion category is populated by instances that share some graded, family resemblance in their features. Of all the features that might describe a category, each instance might contain only a sample of features (resulting in more within category variation and more between-category similarity than is true for Aristotelian categories). The assumption is that there is a single set of features — a single prototype — that best describes each category across all contexts (i.e., there is a single probability structure for the features of the category across situations and contexts because the prototype does not change with context). As a consequence, *moderate reliability and specificity* is expected in the physical features that can be measured in the instances of an emotion category — as much reliability and specificity as the variation in the emotion category will allow. Therefore, it should be possible to objectively read an emotional state in a person (or potentially in a non-human animal) by measuring physical features alone; this pattern may vary some by individual or context, but there may be sufficient reliability and specificity in some measurement situations.

As an example of a populations-based approach, the goal-based view of emotion concepts and categories (i.e., emotion categories as conceptual categories) hypothesizes that, there may be strong *reliability and specificity for an instance of a given emotion category within a given person in a specific situation*, but there will often be insufficient reliability and specificity in the measurements of physical features to diagnose the presence of an emotion category when context and individual variation is ignored. The instances of a given emotion category are thought to share a common set of features within a specific situation, but these features (including the goal or function of the category) can change from situation to situation (Barrett, 2006a, 2012, 2013, 2017a,b; Barrett et al., 2015; LeBois, Wilson-Mendenhall, Simmons, Barrett, & Barsalou, 2018; Wilson-Mendenhall, Barrett, Simmons, & Barsalou, 2011). The hypothesis is that emotion categories, like all abstract categories, do not have conceptual cores (Barrett et al., 2015; Wilson-Mendenhall et al., 2011), meaning that emotion concepts are constructed on the spot, as needed (i.e., they are ad hoc concepts). It is still possible to measure people's understandings of the stereotype pattern for each emotion category, but that is not equivalent to measuring an instance of emotion in an individual within a specific situation in order to predict their behavior.

In the next section, we unpack these generalities and discuss in more detail the measurement considerations for typological versus population-based views of emotion, keeping in mind the different assumptions about the existence of emotion essences.

2.3 Measurement implications of the typological view of emotion categories

There is a particular measurement theory implied in the typological view of emotion categories. If emotion categories are distinct kinds, then the best way to measure emotional instances is using perceiver- independent tools, such as facial electromyography, measures of autonomic nervous system activation, or brain imaging. All measures (except perhaps, self-reports of subjective experience) should be correlated and therefore interchangeable with one another, because all signals have a common cause (the emotion essence; for a discussion, see Barrett, 2000, 2006a, 2006b, 2011a, 2011b; Coan, 2010; Russell, 2003). This idea is embedded in classical measurement theory and depicted using the notation of probabilistic graphical models in Fig. 2.2. From an information theory standpoint, individual measures do not carry unique

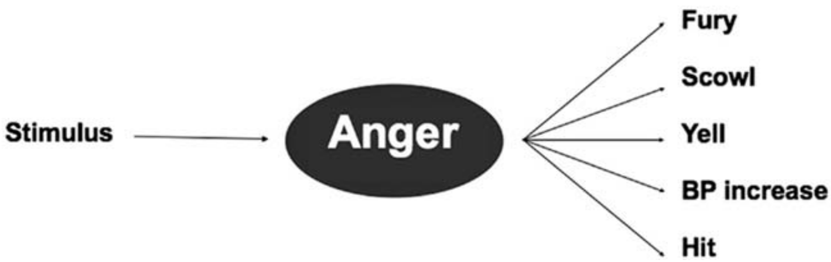

FIG. 2.2 The measurement model for the typological view of emotion categories. *BP*, blood pressure. This is a causal model and a measurement model. The causal mechanism (the emotion essence) is depicted in the *black oval*. The resulting emotion category fingerprint is on the right. In structural equation modeling, which is a type of probabilistic graphical model, the oval is latent (i.e., not measurable) but its existence is established by correlations between the output measurements that make up the fingerprint (scowling, yelling, etc.). Various emotion essences have been proposed, including an affect program (e.g., Ekman), appraisals (e.g., Roseman, Scherer), brain circuits (Tracy & Randles, 2011), a dynamical brain system (Lewis, 2005), and even a pattern of brain activity (Kragel & LaBar, 2015).

information about the emotion cateogry. Self-reports, on their own, are considered a fallible way of measuring emotional instances, because according to this view, people may not be conscious of their emotional state. So it is necessary to scientifically test the accuracy of self-reports by comparing them with the perceiver- independent measurements before they can be used (and they should be correlated when accurate). When self-reports do not correlate with more objective measures (as is often the case), then self-reports are assumed to be inaccurate.

In the typological view, certain emotion categories are universal in humans (meaning that measurement tools should reveal the same underlying causes and fingerprints for each emotion category across cultures) and are homologous in other animals (meaning that studying emotions in nonhuman animals will sufficiently reveal the secrets of emotions in humans). Therefore, classical conditioning with an aversive cue (like an electric shock) becomes interpretable as "fear learning" (for a discussion see Barrett, 2012; LeDoux, 2014). Reward processing in rodents and monkeys should be the same as in humans because both are located in the so-called "limbic system," which is supposed to be evolutionarily primitive in the brain and identical in all mammals (or at least in primates) (MacLean, 1949, 1990; Panksepp, 1998); only humans are supposed to have a well-developed neocortex necessary for cognition and therefore for sophisticated emotion regulation. One implication of this formulation is that emotion words and concepts (which are part of cognition) are separate from the instances of emotion themselves (Izard, 1993, 2011); self-reports of emotional experience are assumed to involve consciously accessing an emotional state (with more or less accuracy) compared to and described with adjectives on a questionnaire (such that the experience of emotion is presumed to be independent of the emotional state itself).

2.3.1 Drawbacks to a typological measurement model

The major stumbling block in using a typological view to guide measurement of emotional instances is that it does not fit the majority of data that have been collected to study it (e.g., Barrett, 2006a, 2006b, 2017a, b; Barrett, Lindquist et al., 2007). Some studies do support the typological view, but these exist in a context of many more studies that call it into question. Studies that are inconsistent with a typological interpretation are rarely discussed in scientific reviews that are written in support of the typological view, which ultimately confuses newcomers to the field who, because they are reading broadly in an effort to familiarize themselves, quickly realize that there is substantially more to digest and integrate than those selectively written reviews provide.

Broader summaries of the literature do exist, however, and they paint a very different picture of the empirical landscape — a landscape full of variation. Consider all the things you do when you are angry: you might tremble, freeze, scream, withdraw, attack, cry, and even laugh or joke. The physiological changes in your body will be tied to the metabolic demands that support your actions in a given situation (e.g., cardiac output increases when you are about to run, but not when you freeze and are vigilant for more information to resolve uncertainty or ambiguity; Obrist, 1981). Sometimes you might feel pleasant but other times unpleasant (Harmon-Jones et al., 2009). Recent meta-analyses and reviews indicate that instances of 'anger,' like the instances of other emotion categories, vary considerably in physical features, and those features are often similar across instances of different emotion categories. This includes associated physiological changes (Siegel et al., 2018),[c] expressive facial movements (Barrett et al., 2019), and neural correlates, whether measured at the level of individual neurons (Clark-Polner, Wager, Satpute, & Barrett, 2016; Guillory & Bujarski, 2014; Minxha et al., 2017; Quiroga, Reddy, Kreiman, Koch, & Fried, 2005; Viskontas, Quiroga & Fried, 2009), as activity in specific brain regions (Lindquist, Wager, Kober, Bliss-Moreau, & Barrett, 2012), networks (Satpute & Lindquist, 2019; Touroutoglou et al., 2015) or as distributed patterns of activity (Clark-Polner et al., 2016; Clark-Polner, Johnson, & Barrett, 2017).[d] Instances of an emotion

c. Even studies using the same methods, stimuli, and sampling from the same population of participants display such variation (e.g., compare findings from Kragel & LaBar, 2013, with Stephens et al., 2010).

d. For example, amygdala neurons respond too slowly to be the brain essence of fear (usually responding about 250 ms after the image is shown). If you consider that it takes about another 500+ ms to mobilize a physical movement, then this is too slow to sound a fear "alarm" when a fearful object or event occurs. Cells in the medial temporal lobe (including the amygdala) appear to act as a memory cache for important things (e.g., photos of friends, family, famous people, the patients themselves, landscapes, directions; some cells don't respond to anything for a few days, and then begin to respond when the experimenters walk into the room); at some other point, the cells might adopt and code for something entirely different that becomes important (Cerf, personal communication, 7/30/15).

category can vary in their affective features (e.g., some instances of fear can feel pleasant, and some instances of happiness can feel unpleasant; Wilson-Mendenhall et al., 2013, 2015). Instances of different emotion categories can also be similar in a range of features, which is not surprising: sometimes you might smile when you are sad, cry when you are afraid, or scream when you are happy. There are also variations across cultures (e.g., Gendron et al., 2020, 2018; Hoemann et al., 2019). Even non-human animals show variable, situated behaviors across instances of the same emotion category (Barrett, 2012). [For a discussion of some reasons why experiments on nonhuman animals do not reveal emotion essences, see Barrett (2012), Barrett & Finlay (2018) and LeDoux (2014, 2015)].

When considered in the broader context of such findings, evidence that appears to support the typological view often has alternative explanations, such as contextual and methodological considerations (e.g., in the literature on facial expressions, see Hoemann et al., 2019; see Barrett, 2011a, 2011b; Barrett et al., 2019; Gendron, Mesquita, & Barrett, 2013; Gendron et al., 2018 for discussions). For example, a recent series of studies has attempted to identify a typology of emotion categories by measuring subjective experience when watching video clips (27 emotion categories identified; Cowen & Keltner, 2017) or listening to music (13 emotion categories identified; Cowen, Fang, Sauter, & Keltner, 2020), in human vocalizations (24 emotion categories; Cowen, Elfenbein, Laukka, Keltner, 2018), and when observing facial and bodily expressions (28 emotion categories identified; Cowen & Keltner, 2020a,b) as well as observing depictions of facial expressions in ancient art (5 emotion categories identified; Cowen & Keltner, 2020a,b). A major critique of this research is beyond the scope of the present chapter, but two key observations are relevant: First, when looking across the findings in this series of studies, they do not conceptually replicate one another by identifying the same number of emotion categories, despite the researchers selecting stimuli with specific category labels in mind (e.g., "The videos were gathered by querying search engines and content aggregation websites with contextual phrases targeting 34 emotion categories" pg. 2, Cowen & Keltner, 2017). The number of categories varied across testing contexts, such as measurement modalities and stimuli. Second, the studies were not designed in such a way as to observe variation if it present (as we discuss below).

Perhaps more importantly, the evidence that is strongly in support of a typological view of emotion categories, when incorporated into the larger empirical landscape, can be re-interpreted as being part of the larger picture of situated variation that occurs in people's emotional lives. This is best explained by example. Participants from large-scale urban cultures tend to scowl about 30% of the time when angry (which is above chance levels; as discussed in Barrett et al., 2019). This means that 70% of the time, they move their faces in other meaningful ways. And people scowl when they are not angry (when confused or concentrating or when they have gas). This evidence, which can be interpreted as low reliability and specificity for a scowl as the facial expression of anger, does not mean that people never scowl in anger, or that they move their faces in random ways during anger. Instead, it implies that scowling is merely one expression of anger that might occur. For example, when the goal of being angry is to overcome an obstacle, it may be more useful to scowl during some instances of anger, smile or laugh, or even stoically widen one's eyes, depending on the temporospatial context. This variation is thought to be a meaningful part of an emotional expression because facial movements are functionally tied to the immediate context, which includes a person's internal context (e.g., the person's metabolic condition, the past experiences that come to mind) and outward context (e.g. whether a person is at work, at school, or at home, who else is present and the broader cultural conditions), both of which vary in dynamic ways over time.

The facial expressions that are assumed to be specific to each emotion category — the Western stereotypes – were not discovered by observing how people move their faces during emotional episodes but were actually stipulated, first by Darwin (1872/1965), and then later by Tomkins and McCarter (1964) (for a brief history, see Barrett et al., 2019, supplementary online materials (SOM), Box 4; Widen & Russell, 2013). Furthermore, engineers have shown that different combinations of action units can produce similar-looking expressions, violating a basic assumption of the typological view (Tian, Kanade, & Cohn, 2001).[e] The empirical evidence, when considered in

e. To create facial movements that are visible to the naked eye, facial muscles contract, moving skin into folds and wrinkles on an underlying skeletal structure. These are the movements people observe in one another, that are captured in photographs and videos, and that express instances of emotion. Individual differences in facial anatomy, as well as in the brain's control of facial muscles, cause variation in the details of how facial movements are executed at the muscular level and how they look to the naked eye. People vary in the underlying bone structure of the face and details of the skin, the structure and strength of their facial muscles (Pessa et al., 1998), the dynamics of facial muscle movements, and consequently they vary in how their facial movements look to a human observer (e.g., Farahvash et al., 2010; Shim et al., 2008; Shimada & Gasser, 1989). In addition, some people have strong asymmetries for one side of the face or the other, and some people lack certain smaller muscles altogether (Waller, 2008). In fact, if you inserted your exact facial muscles into a different face (someone with a different bone structure or someone much older or younger than you, or whose face is thinner or fatter), the resulting muscle movements would look different than they do on your face. And even when facial movements look the same to the naked eye, there may be differences in their execution under the skin. As human perceivers, we see stable facial behaviors (i.e., a frown) when in reality, under the skin, there is more variation that meets the eye. A facial behavior, like frowning, or scowling, is, in fact, a category of variable instances. When you watch a frown unfold in the same person on two different occasions, the exact muscle contractions that curl the upper lip and turn down the corners of the mouth can subtly vary from one instance to another. What to the naked eye looks like the same frown in two different people can result from different patterns of underlying muscle contractions. Put simply: something as seemingly simple as a single facial movement is best understood as a conceptual category, resulting from a variable set of more basic, variable physical changes.

its totality, suggests that variation is the norm when it comes to emotion (Barrett, 2017b).

Here is the takeaway point: if you adopt a typological view to guide experimental design and emotion measurement, you are accepting a set of assumptions that are thus far unvalidated. Much of the confusion in the science of emotion could be avoided if scientists properly evaluated their assumptions (e.g., if they discovered, rather than stipulated, how people move their faces and bodies, how vocal acoustics change, and how autonomic reactivity fluctuates during naturalistic instances of an emotion category in a range of contexts; for examples of two recent studies taking this approach, see Hoemann et al., 2020; Azari et al., 2020). In the meantime, it is crucial to be cautious about studying and measuring instances of emotion as if emotion categories have essences. We find it ironic that most studies in the science of emotion are designed to induce and observe only the most stereotypical instances of emotion, yet those studies routinely produce evidence of substantial variation in facial movements, autonomic patterns, and brain activity, well beyond what would be expected by error. So imagine what the science of emotion would be like if our starting assumption was that variation is the norm, and we attempted to measure and capture that variation, both within a person across contexts as well as across individuals both within and across cultures. And, in fact, an increasing number of reports are doing just that (Ceulemans, Kuppens, & Van Mechelen, 2012; Hortensius, Schutter, & Harmon-Jones, 2011; Kuppens, Van Mechelen, & Rijmen, 2008; Kuppens, Van Mechelen, Smits, De Boeck, & Ceulemans, 2007; Nezlek, Vansteelandt, Van Mechelen, & Kuppens, 2008; Stemmler, Aue, & Wacker, 2007), but more are desperately needed.

Those who adhere to the typological view have a standard approach to solving the ever-widening chasm between their assumptions and the accumulating scientific evidence. The response to variation is usually to create more fine-grained typologies, in an attempt to bring nature under control and make it easier to identify emotion essences. Maybe there are "primary" and "secondary" emotion categories? Maybe there are "basic" and "non-basic" emotion categories? Maybe there are "non-social" and "social" emotion categories? Scientists sometimes try other typological divisions, like distinguishing emotion "elicitation" from "emotion regulation," and when that does not do the trick, they distinguish "steps" and "sequences" in their proposals about emotion elicitation. They distinguish "affect" from "value," and then cleave "value" into different types of "valuation." And so on. But there is another approach to emotion categories, one that a priori predicts considerable variation in emotional life and accounts for all the empirical evidence (both the evidence that supports typological views and that which does not). This is an essence-free view that considers typological thinking as a lot of misplaced creativity and effort.

2.4 A populations view of emotion categories and its measurement implications

A populations view of emotion categories *assumes* that variability is the norm rather than a nuisance to be explained after the fact. It draws on Darwin's population thinking, in which a biological category, such as a species, consists of a population of highly variable individuals (Darwin, 1859/1964; for a discussion, see Barrett 2017a,b, Barrett, 2020; Siegel et al., 2018). A population of variable instances, as a category, can be represented as a distribution, with a mean (the average instance of that category) and variance (the spread of the distribution around the mean). In a typological view, the mean of the distribution is real, and any variation around the mean is considered to be due to error or other influences. A populations view posits the opposite: the mean is a statistical summary that may not exist in nature, and the variation within a category is real.

According to a populations view, a biological category does not represent a physical type with a pattern of features that are reliable and specific across all or even most category instances. For example, an instance of happiness can be pleasant and arousing (e.g., you are finishing a challenging task with no errors and hear applause), pleasant and quiescent (e.g., you feel comfortable and rested after a good night's sleep), and even unpleasant (e.g., you want to call your friend to share your recent success but he is unreachable) (Wilson-Mendenhall et al., 2013). The actions you make in happiness will depend on the situation (e.g., you might laugh, smile, cry, jump, sigh, shout, slam your fist against a table, and so on)—whatever has been most functional for you in past, similar situations. And your cardiovascular response will be similarly variable across instances of happiness, because those responses support action (Obrist 1981; Obrist, Webb, Sutterer, & Howard, 1970); therefore, there is no one-to-one correspondence between a behavior, a physiological pattern, and an emotion word (e.g., Lindquist, Siegel, Quigley, & Barrett, 2013).

Thus, in the populations approach, an emotion category does not have an essence. Without essences, it is not meaningful to ask what is, and what is not, an instance of emotion. Nor do you "have" emotions or "recognize" them. Your brain constructs instances of emotion as experiences of yourself in the world or as perceptions of others —they emerge from complex dynamics within your nervous system, which is constantly in dynamic interaction with the surrounding context, often including other creatures who each have a nervous system. These constructed instances of emotion may or may not include experiences of emotion, which require an explicit awareness of being in a particular emotional state (Adolphs, Mlodinow & Barrett, 2019). In a sense, you can think about instances of emotion as tools, born of the social reality humans create, to influence and regulate your actions, your experience of the world and one another's nervous systems (Barrett, 2012, 2017a, b).

Because constructionist theories are much less intuitive, and because they are highly heterogeneous, it usually helps to focus in more closely on one theory to more fully understand its implications for the measurement of emotion. Here, we will focus on the theory of constructed emotion.

2.4.1 The theory of constructed emotion

According to the theory of constructed emotion, an instance of an emotion category (and any other mental category) emerges as the brain makes meaning of incoming sensory inputs from the body and the world. Every waking moment of your life, your brain is taking in constantly changing, noisy sensory information from the world and transforming it into sights, sounds, smells, and so on. From your brain's perspective, your body is another domain that is sending constantly changing, noisy, incomplete sensory inputs from your blood rushing, your muscles stretching, your lungs expanding, and so on; the autonomic nervous system, endocrine system, and immune system all create sensory changes within what scientists call the internal milieu of your body, and you brain makes sense of these as affective feelings that belong to physical symptoms, emotions, thoughts, perceptions, and so on. How does your brain make sensations meaningful? By categorizing them. This means using past experience, organized as concepts, to explain what caused the sensations and what to do about them (i.e., how to act). Here is a succinct summary of the theory:

> *In every waking moment, your brain uses past experience that function as concepts to guide action and give sensations meaning. In this manner, your brain models your body in the world. When the concepts involved are emotion concepts, your brain constructs instances of emotion.*

To demonstrate categorization using concepts from past experience, have a look at Fig. 2.3.

If you are like most people who have never seen the image in Fig. 2.3 before, then you are in a state of "experiential blindness." Your brain cannot categorize the visual input—it cannot make sense of it—so all you see are black and white blobs. To cure your experiential blindness, please turn to the appendix (Fig. A.1), and then return to this page.

After viewing the appendix (Fig. A.1), most people now see an object in Fig. 2.3. So what does this exercise demonstrate? Your brain added information, stored from your (very recent) past, to make sense of the incoming sensory input (visual) to *construct* your experience of the object in Fig. 2.3. This example is instructive in several ways.

First, the construction process is ongoing, obligatory, and automatic; notice that you had no sense of agency or effort in the construction process. No matter how hard you try, you cannot introspect about how your brain accomplished this feat of making incoming sensations from Fig. 2.3 into a

FIG. 2.3 An example of categorizing to construct an experience.

meaningful visual experience. Also, it is virtually impossible to "unsee" the object—to deconstruct the experience by the sheer force of will.

To the best of our current knowledge, here is what went on in your brain. Neurons in certain parts of your brain changed the firing of neurons in your visual cortex to construct your experience of lines that actually aren't present on the page, linking the blobs into the shape of a cow whose image isn't physically there on the page. Scientists call this "simulation" (Barsalou, 2008). Simulation is when the neurons in some parts of your brain changed the firing of sensory neurons in other parts of the brain so that you can, for example, see lines and other visual features, without sensory input. Simulation can be visual, as in this example, but it also involves your other senses. If you've ever had a song stuck in your head, or put food into your mouth, expecting to taste one thing but then experienced the shock of tasting something entirely different, then you have experienced simulation in other sensory modalities. Memories, daydreams, mind wandering—these are also examples of simulation. In the science of emotion, we measure this kind of simulation all the time without realizing it. We hook people up to blood pressure monitors, electrocardiograms, and so on, show them evocative images, and then measure changes in autonomic nervous system activity, and even though people are sitting perfectly still this works, not because the images "trigger" reactions, but because people are simulating the action that they would make in that situation, as well as the interoceptive sensations (from the core of the body) that they would receive. Whenever you ask respondents to report on an experience that has happened in the past, this is also tapping simulation. Simulation during brain scanning produces activity in somatosensory and motor cortices when subjects are completely still, in primary visual cortex when eyes are closed, and even in primary interoceptive cortex (for sensing changes in the core of the body) when there is no threat or reward immediately present (Wilson-Mendenhall et al., 2013, 2019). Understanding the brain dynamics of simulation—how people apply knowledge wired into the brain to create experiences and perceptions in a particular context will reveal how the brain constructs experiences and perceptions of emotion.

A full explanation of simulation is beyond the scope of this chapter, so a brief summary will have to suffice (for an accessible discussion, see Barrett, 2017b; for a more academic discussion, see Atzil et al., 2018; Barrett, 2017a, b; Barrett & Finlay, 2018; Gendron, Mesquita, & Barrett, 2020; Wilson-Mendenhall et al., 2019). The first thing you must realize is that your brain is not merely responding to stimuli in the world. Your simulations function like predictions that continuously *anticipate*, rather than react to, sensory inputs from the world. Your brain is wired to be a generative model of your world, by using past experience to actively create simulations that best fit the situation you are in. The second insight is that predictions, as simulations, are then corrected by sensory input from the world; so, information from the world is feedback on how good the simulations are. This includes not only the neurons for vision, audition, touch, taste, and smell, but also for interoception, because from the brain's perspective, the body is part of the brain's world (since the body holds the brain); it's also likely true for affect, which is the low dimensional experience of interoceptive sensations. Your brain is constantly generating predictions of upcoming sensations and then adjusting these predictions (more or less) by computing error signals that track the difference between the predicted sensations and those that are incoming from the sensory world. And the brain is not only making sensory predictions—it is also making motor predictions; it is anticipating the motor changes that will be required in a moment from now by changing the firing of motor neurons before they are needed. In fact, your brain generates visceromotor predictions (to control your autonomic nervous system, your neuroendocrine system, and your immune system) and voluntary motor predictions first, and then anticipates the sensory consequences of those visceromotor/motor predictions (i.e., predicted motor changes produce sensory predictions) so that, in a sense, sensation follows (and is dependent on) action (Barrett, 2017a; Barrett & Simmons, 2015; Chanes & Barrett, 2016; Clark, 2013; Friston, 2010; Hohwy, 2013; Hutchinson & Barrett, 2019; Kleckner et al., 2017; Seth & Friston, 2016).

When your brain creates a prediction from past experience, it does not issue one neural pattern, but an entire population of potential predictions, each one having some probability (computed with Bayesian priors) of being the best fit to the current circumstances (Barrett, 2017a). This population of neural patterns is, for all intents and purposes, being treated by your brain as similar for some purpose—to make meaning of and dealing with the impending sensory array. Another insight of the theory of constructed emotion, then, is that this population of predictions is a *concept*, constructed as you need it, on the fly (what Barsalou and colleagues call an "ad hoc" concept; Barsalou, 1983, 2003; Barsalou et al., 2003). Certain predictions will provide a better fit to the incoming sensory input, and these become your perception and guide your action (i.e., they categorize your sensory inputs). So, constructing meaning by correctly anticipating (predicting and adjusting to) incoming sensations is what we mean when we say that the brain is using emotion concepts to categorize sensations to construct an instance of emotion. Sensations are conceptualized (i.e., categorized) so that they are (1) actionable in a situated way and therefore (2) meaningful, based on past experience. The sensory array in need of prediction and action contains both interoceptive inputs from the body representing the allostatic changes in the body's various systems (the internal world) and exteroceptive inputs representing sensory changes in the outside world. When past instances of emotion (e.g., happiness) are used to categorize the predicted sensory array and guide action, then an instance of that emotion category (happiness) is experienced or perceived. An emotional instance is constructed the way that all other perceptions are constructed, using the same neural systems (and correspondingly, the same domain-general psychological processes) (Barrett & Satpute, 2013; Lindquist & Barrett, 2012; Satpute & Lindquist, 2019). This is why the neuroscientist Edelman (1998) referred to experience as "the remembered present."

To see how this might work, let's do a thought experiment. For example, in the past, you might have experienced the comfort of dozing on a hammock on a sunny day, the comfort of hugging a long-lost friend, the comfort of eating a piece of chocolate cake, the comfort of a warm bath, the comfort of flying on an airplane, and the comfort of reclining on a sofa in front of a fire. Each instance of comfort is not identical to every other, and when the brain creates an ad hoc concept for comfort to predict incoming sensory inputs, it constructs simulations (as potential actions and perceptions) of those instances that are most similar to the current situation (each prediction having some probability of being correct, given past experience). So the brain simulates an on-line concept of comfort, not in absolute terms, but with reference to your particular goal in the moment (e.g., to relax and minimize stress, to feel close to others through shared comfort, and so on). This means that an emotion category word like "comfort" or "happiness" has a specific meaning, but its specific meaning can change from one instance to the next (Barrett, 2017a; LeBois et al., 2018).

As we briefly noted earlier, the theory of constructed emotion is an evolutionary theory, but it does not hypothesize that emotion categories are universal. Unlike typological views, which take their inspiration from the more essentialist *The expression of the emotions in man and animals* (Darwin, 1872/1965), the theory of constructed emotion uses conceptual innovations found in Darwin's *On the origin of species* (Barrett, 2013). In fact, Darwin is credited with vanquishing essentialism in biology in *Origin* (Mayr, 2004), so it is ironic that he went on to write a highly essentialized treatment of emotion categories slightly more than a decade later (for a hypothesis of why Darwin did this, see

Barrett, 2017a; also see Fridlund, 1992). The theory of constructed emotion's use of population thinking comes from Darwin's *Origin* (e.g., a species is not a natural kind category with one, fixed Platonic form serving as its essence, where variation around this form is error; instead, a species is a conceptual category populated with unique individuals who have differing degrees of fit to the environment). The theory of constructed emotion also incorporates Darwin's focus on holism, or the need to study a part in the context of the whole that influences it. Intrinsic to holism is the importance of studying a phenomenon in context, rather than attempting to find general, context-free laws (as is typical in certain forms of reductionism; Mayr, 2004).

The theory also incorporates several other concepts from biology, the most important being that there is more than one cause to produce the same phenomenon, called degeneracy (Edelman & Gally, 2001; Marder & Taylor, 2011; Tononi, Sporns, & Edelman, 1999). Degeneracy is a property of virtually every level of analysis in biological systems, from the systems inside cells to the entire organism. For example, different proteins can catalyze the same reaction of enzymes (Edelman & Gally, 2001; Tononi et al. 1999), different antibodies can bind to the same antigen (Edelman & Gally, 2001), different genotypes can produce the same phenotype (Edelman & Gally, 2001; Tononi et al. 1999), different neurons can give rise to the same intrinsic network (Marder & Taylor, 2011; Tononi, Edelman, & Sporns, 1998; Tononi et al., 1999), and different patterns of network interaction can give rise to the same behavior (Price & Friston, 2002). Degeneracy refers to the capacity for structurally dissimilar systems or processes to give rise to identical outcomes (Edelman & Gally, 2001), such as many different facial configurations, autonomic configurations, or brain states mapping to the same emotion category.

2.4.2 Measurement implications

The measurement model implied by the theory of constructed emotion is difficult to draw, because it involves tracking a high-dimensional brain and body state dynamically over time in an iterative way. A poor approximation is depicted in Fig. 2.4, although strictly speaking, this figure has limitations for modeling the theory of constructed emotion (for a discussion, see Barrett, 2011b).

In this measurement model, assessments of facial movements, autonomic reactivity, and vocal acoustics will not necessarily be correlated, and therefore carry unique information about an instance of emotion (i.e., they are not necessarily interchangeable from an information theory standpoint). (This measurement model is useful, because in reality, measures from different modalities are rarely correlated with one another.) One measure cannot stand in for another, so that *optimal* measurement of instances of emotion requires a multimodal approach. The experience of an emotional instance can change from one instance to the next (LeBois et al., 2018), and therefore cannot be

FIG. 2.4 A measurement model for theory of constructed emotion. This is a measurement model, but not a causal model, of emotion (i.e., this figure does not depict the mechanisms that cause instances of emotion, but only how measures might configure to assess an instance of emotion) (for a fuller explanation, see Barrett, 2011a, 2011b). For a causal model, see Barrett (2017a) or Hutchinson & Barrett (2019).

stipulated in advance; instead, the relationships between measures must be inductively discovered within a person across contexts (for a discussion, see Barrett 2020; Siegel et al., 2018; for examples of this approach, see Hoemann et al., 2020; Azari et al., 2020). From this standpoint, it will never be possible to properly measure instances of emotion by merely measuring the face, or skin conductance, or any single or set of physical measures alone.

In the theory of constructed emotion, without an idiographic, multimodal measurement approach (described in the next section), it is perhaps best to measure instances of emotion via self-report, because there is no "objective" way of determining when someone is, or is not, in a particular emotional state (Barrett, 2006b). Furthermore, to date, we have no ability to model emergence mathematically when it comes to emotion (where the product of dynamically interacting systems has properties that the systems themselves do not). Self-reports, on their own, have limitations, of course, because they only capture some instances of emotion (those of which the respondent is aware); in this view, as in the typological view, people may not be conscious of the emotion category they have constructed (although for an entirely different set of theoretical hypotheses). But when self-reports do not correlate with more objective measures, the self-reports are *not* necessarily assumed to be inaccurate (more on this below). Moreover, because emotion concepts are integral to the construction of emotional experiences and perceptions, words and other symbols that prime emotion concepts (i.e., that launch predictions and simulations) will influence what is experienced and felt. As a consequence, self-reports of emotional experience are influenced by the words that we give respondents to communicate their experiences or perceptions. It is possible to change a person's feeling merely by the type of measurement instrument you give them. And respondents will use the measure you give them to report what they want to tell you, which may not necessarily match what you are asking (e.g., if a respondent feels excited, but you ask if he/she is happy, the respondent will use the item to tell you how excited he/she is).

In the theory of constructed emotion, emotion categories are not universal (meaning studying all aspects of emotion, including emotion concepts, is crucial across cultures; Gendron et al., 2018, 2020; Hoemann et al., 2019). Nor are emotion categories assumed to be homologous in other animals (meaning that studying emotion in nonhuman animals will *not* reveal all the secrets of emotions in humans, although such study is undoubtably crucial to learn about some ingredients of emotion; Barrett & Finlay, 2018). Humans and other animals are hypothesized to share some species-general core systems, and some species-specific core systems. Therefore, classical conditioning with an aversive cue (like an electric shock) is not "fear learning" but "threat learning" (for a discussion see Barrett, 2012; LeDoux, 2014). Reward processing in rodents and monkeys might be the same as in humans in some ways, but it might also be different because the brain did not evolve according to a phylogenetic scale like sedimentary rock; the cortex did not evolve on top of preserved subcortical regions like icing an already-baked cake; as brains grow, they expand and reorganize (Barrett et al., 2007; Striedter, 2005).

Another implication is that questions about "recognition accuracy" are not scientific, because emotion categories have no perceiver-dependent essences with which to compare to a perceiver-based judgment (whether a perception of someone else's emotion or a self-report of experience). Instead, what we actually measure is consensus (Do you and I agree on the category of emotion you are feeling? Do you and I agree on the category of emotion that some other person is feeling? Do our perceptions agree with the cultural norms for emotion categories in this specific situation?). The reliance on consensus (or agreement) is not a bug—it is a feature that reflects the status of emotion categories as social (not biological) kinds (Barrett, 2012).

Emotion categories are not assumed to be perceiver-independent phenomena, waiting to be discovered in nature by a human mind. They are instances that are created within a human mind, in concert with other human minds. Emotion categories depend on the human mind for existence—they are perceiver-dependent phenomena. Thus, emotion categories are made, not found. They are perceived, not detected. And measuring instances of emotion *requires measuring human experience and perception*, as well as "objective" measures, such as facial muscle movements, cardiovascular reactivity, and so on. Measuring an instance of emotion means capturing when and how mere physical changes are categorized so as to serve the psychological functions of an emotional instance (as opposed to other times when the same physical changes are not understood as an instance of emotion and therefore serve some other psychological function; Barrett, 2012; MacCormack & Lindquist, 2019).

2.4.3 Drawbacks to using the construction approach's measurement model

Complexity and cost, both in terms of time and money, are the primary drawbacks to using the measurement model of the theory of constructed emotion. Context-sensitive measurement of emotional instances requires deeply idiographic sampling, with a large number of unique, variable instances across many different situations, times of day, and so forth. Research should sample many domains of features at high dimensionality, including measuring a combination of biological (e.g., monitoring of autonomic physiology changes, eye-tracking to assess gaze and attention, brain imaging, etc.) and psychological features (e.g., measuring a broad set of mental features that might describe instances of emotion, such as appraisal features or functional features) that sample both a person's internal context (e.g., someone's autonomic physiology data) and external context (e.g., where a person is, broader cultural conditions, etc.). An idiographic, multimodal sampling approach will allow for the potential discovery of reliable emotion categories that characterize a person's emotional life over time and contexts as well the extent to which these categories generalize across individuals (either due to similarities in context, learning history, cultural background, etc.). For example, Hoemann et al. (2020) used a physiologically-triggered experience sampling procedure to idiographically sample participants' states and experiences throughout daily life and observed variation in both the number and nature of physiological patterns within a person. Some of these patterns replicated across participants. And most importantly, all showed a many-to-many mapping between the physiological categories and the emotion labels that participants provided.

In this approach, the more observable aspects of emotional instances (facial movements, autonomic responses, etc.) cannot be used to stand in for or validate a person's own experience, at least with the methods that we have currently available. Validation requires measuring the underlying mechanisms that create an instance of emotion, and optimally this requires developing a more suitable epistemological approach (i.e., following individuals in context over time) rather than anchoring and adjusting way from an approach that was developed to evaluate typological views.

In the meantime, lack of correspondence between verbal reports and behavior does not necessarily indicate that the verbal reports are invalid. In fact, because instances of emotion are thought to emerge from more basic processes, the instances will have features that are not reflected in measurements of the individual processes themselves. Thus, for now, verbal report, even with all of its failings, may be the best means of assessing the experience of emotional instances in a quick and efficient way. If you want to know whether people are experiencing an instance of emotion, you have to ask them (e.g., Barrett, 2006a). This may sound easy, but there are a few landmines here as well. You can't assume that a feeling of happiness in one context is similar to the feeling of happiness in another (LeBois et al., 2018; Wilson-Mendenhall et al., 2011, 2013). You also can't necessarily assume that two people mean the same thing by happiness, as people differ in emotional granularity; for some people, the word "happiness" refers to a specific feeling state, whereas for others, it refers to a general, pleasant feeling (e.g., Barrett, 2004; Barrett & Bliss-Moreau, 2009). It is also not possible to assume that an emotion word means the same thing in different cultures (e.g., Barrett, 2017a; Jackson et al., 2019; Pavlenko, 2014). In fact, a construction mindset helps explain why certain emotion categories exist in some cultures, but not in others, and that what counts as an emotion category in some cultures is not an emotion category in others (for a discussion, see Barrett, 2017a; Pavlenko, 2014; Russell, 1991; Mesquita, Boiger, & De Leersnyder, 2016). The implication is that it is always a good idea to include a measure of emotion concepts whenever measuring the experience of an emotional instance.

2.5 A cautionary note: beware of lurking essentialism

Whether you rely on the assumption of emotion essences or not, it is important to be vigilant for the use of an essentialist mindset where it does not belong, lest you misinterpret your own (or someone else's) findings. A good example of such misinterpretation can be found in the increasingly frequent use of pattern classification approaches to distinguish the instances of one emotion category from another. Pattern classification techniques refer to a family of statistical methods designed to categorize data by learning from existing categories or grouping variables to make predictions about and assign membership to new instances. These techniques are being used with measures of autonomic physiology, facial movements, and changes in neural response measured as blood oxygenation level-dependent (BOLD) signal within the brain (that is divided up into three-dimensional cubes called voxels) (Horikawa, Cowen, Keltner, & Kamitani, 2020; Kassam, Markey, Cherkassky, Loewenstein, & Just, 2013; Kragel & Labar, 2013, 2015; Kragel, Knodt, Hariri, & LaBar, 2016; Park, Jang, Chung, & Kim, 2013; Rainville, Bechara, Naqvi, & Damasio, 2006; Saarimäki et al., 2016, 2018; Stephens, Christie, & Friedman, 2010; Wager et al., 2015; Yuen et al., 2012). Researchers train a classifier using some set of measurements for known instances of emotion categories, and then use the classifier to diagnose new instances of those categories using similar measurements (see Kragel, Koban, Barrett, & Wager, 2018 for a discussion). For example, Fig. 2.5 presents the multivoxel patterns that successfully classified five categories of emotion above chance in our recent meta-analytic paper (Wager et al., 2015). We trained classifiers on brain maps from existing studies of anger, sadness, fear, disgust, and happiness, and then used the classifiers to diagnose the emotion category being represented in brain maps from new studies.

The lure of essentialism leads scientists to claim that the patterns are something like the neural essence for each emotion category. For example, several recent studies using pattern classification have claimed to find 'biomarkers' 'fingerprints,' or 'signatures,' in the brain for certain emotion categories (Kassam et al., 2013; Kragel & LaBar, 2015; Saarimaki et al., 2016, 2018), where a biomarker (or 'fingerprint' or 'signature', which are synonyms) is a measurable indicator of some category, such that its presence in an

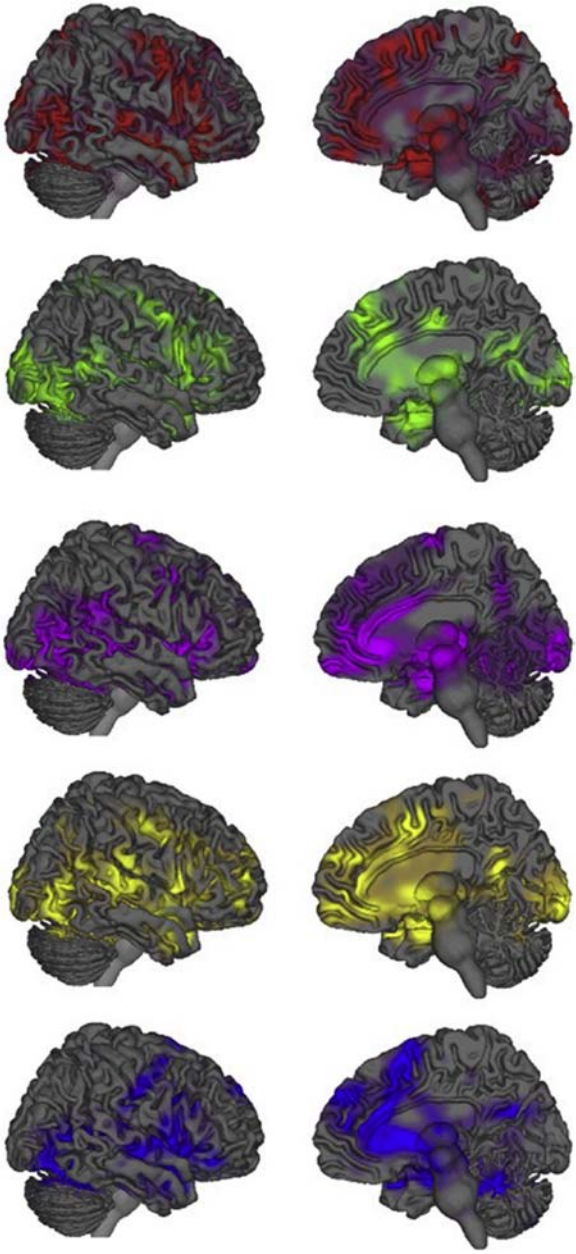

FIG. 2.5 Intensity maps for each of the five emotion categories examined by Wager et al. (2015). Classification rates were Anger (red) = 43%, Disgust (green) = 76%, Fear (pink) = 86%, Happiness (yellow) = 58%, and Sadness (blue) = 65%.

instance indicates that the instance belongs to a particular category (Strimbu & Tavel, 2010).[f] In medicine, a biomarker is a measurable substance that is present in all members of a category. Biomarkers must be sensitive and unique to a specific category to work properly. These recent pattern classification studies interpret their findings as support for the typological view of emotion categories, when, in fact, they actually found evidence in support of the theory of constructed emotion (Barrett, 2017b; Clark-Polner et al., 2016, 2017).

Patterns that successfully distinguish one emotion category from another are not emotion essences. They are not biomarkers, fingerprints, or signatures of an emotion category. The implication in using terms like "biomarker," "fingerprint," and "signature" is that the pattern for an emotion category is it's brain state—the elements of the pattern (e.g., the pattern of voxels) are assumed to be both unique to a single emotion category and unchanging across its instances. Yet, a pattern derived from pattern classification techniques does not appear in *every instance* (or in fact, in *any instance*) of a category, even when the pattern can classify a category's instances with 100% accuracy. We have demonstrated this with simulations (Clark-Polner et al., 2017). Patterns should be understood from the standpoint of population thinking—the pattern is an abstract, statistical summary of a category's instances; the pattern does not (and need not) exist in nature to work well. Although as a group, the instances of any emotion category can be diagnosed with a pattern, the pattern itself is an abstraction. Similarly, the average middle-class US family has 3.13 children, but this is an abstract representation, because no family actually has 3.13 children. To assume that a pattern is the fingerprint or biomarker for an emotion category is to mistake a statistical summary for the norm, when it is in fact a statistical abstraction that may not exist in nature. What exists instead is real and meaningful variation around this summary. The take-away point is that successful pattern classification using *any* type of measures provides evidence that emotion categories, similar to biological categories, are conceptual categories populated by unique and highly variable instances that do not share any necessary features.

Consistent with a populations approach, a close look at recent multivoxel pattern analysis findings (e.g., Kassam et al., 2013; Kragel & Labar, 2015; Saarimaki et al., 2016, 2018; Wager et al., 2015) reveals that the patterns distinguishing one emotion category from another are not consistent across studies (for a discussion, see Azari et al., 2020). Methodological considerations, such as small sample sizes, variable preprocessing workflows, variable classification algorithms, and so on, might contribute to the variation in these findings across studies. However, the same inconsistency is seen even when two studies classified patterns of autonomic physiology using the same stimuli,

f. Kragel and LaBar (2015) are careful to point out "it is unlikely that the patterns we identified perfectly capture the essence of an emotion, but are better characterized as an amalgamation of the components which make emotions unique" (p. 1446).

the same experimental method, and sample participants from the same population (Kragel & LaBar, 2013; Stephens et al., 2010).

An additional explanation for the inconsistency across findings is that the different patterns reflect real variation in the data. If this explanation is correct, then the 'fingerprints' or 'biomarkers' for different emotion categories that have been reported in pattern classification studies do not reflect objective categories that fail to replicate, but are better understood as statistical abstractions for a given study that may not exist in the real world but that summarize real variation that does exist in the real world. In fact, we recently compared findings from supervised pattern classification (where trials are labeled a priori as belonging to specific emotion categories, such as fear, happiness, etc.) to those from unsupervized clustering, in which no a priori labels are applied to the data and we identified reliable clusters of data that did not correspond to the labels used in supervised classification (Azari et al., 2020). Instead, the clusters contained a mixture of data corresponding to different emotion categories (e.g., example participant 1 had two clusters of brain data, each containing a mixture of trials labeled as fear, happiness, and sadness). These findings suggest that emotion category labels refer to populations of highly variable, context-specific instances, rather than biological kinds that are stable across individuals. Once again, these findings provide evidence that variation is the norm.

2.6 Conclusions

Essentialism is not necessarily a bad thing. Utilizing it unknowingly is, however. The goal of this chapter is to allow you to view the current science of emotion as a continuation of the long-standing debate over whether mental categories are carved into nature by essences, or whether they are more flexible groupings of highly variable and situated instances, created from more basic mechanisms (for a discussion, see Lindquist et al., 2012). Some scientists consider essentialism a useful strategy for scientific inquiry because they believe that it mirrors the structure of the real world (i.e., they believe the world is full of natural kind categories) (e.g., Bloom, 2000; Kornblith, 1993; Pinker, 1997). Others, however, believe that essentialism is a particularly poor strategy for scientific inquiry (Lewontin, 2000). You can make your decision. Just do it explicitly, and with an appreciation of the consequences.

In the new "emotion economy," many businesses are investing millions of dollars and tremendous person-hours developing "emotion-aware" technology that they have been led to believe will be able to "read" instances of emotion from perceiver-independent measurements of the face, body, and behavior (i.e., they are developing technology and algorithms to measure emotional instances solely by tracking how facial muscles move, how autonomic signals in the body change, how electricals signals across the scalp change, and so on; for a discussion and review, see Barrett et al., 2019). Companies have unwittingly embraced essentialism, ignoring years of disconfirming evidence. But if the scientific literature is to be trusted, and one includes all well-designed and well-executed studies (not just those that support's a particular set of deeply held beliefs), then our best conclusion is that universal, coordinated suites of physiology, behavior, and so on (i.e., a Platonic essence) issuing from a dedicated mechanism (i.e., a Lockean essence) don't exist for any emotion category that has been studied. The unfortunate consequence is that so much effort and investment will be wasted. Even worse, those who consume the science of emotion might end up concluding that emotions are not an important part of the equation in predicting behavior, all because they unwittingly used a set of unwarranted guiding assumptions to begin with.

If you are going to use a typological view to guide your measurement of emotional instances, you should do so with caution, realizing that the bulk of the scientific evidence does not yet support it. This leaves you vulnerable to investing a lot of time and money in an enterprise that might seem deeply intuitive, but that might deliver very little in return. Alternatively, for those of us who relinquish essentialism as a guiding assumption for the nature and measurement of emotion, we should refrain from feeling smug or superior. Beware the lesson of Charles Darwin, who became famous in one field (biology) by vanquishing essentialism, while becoming famous (psychology) in another by relying on it (see Barrett, 2017b).

It is worth pointing out that the history of science can be read as a long, slow march away from essentialist thinking, discovering that universal laws are actually contextual (e.g., in physics, with the discovery of relativity theory and then quantum mechanics) and discovering that variation is meaningful and is not error (e.g., in biology, with Darwin's *On the origin of species*, and then again a century later with the study of epigenetics and genomics). Construction approaches have also emerged in psychology, precisely when it is discovered that instances grouped together as the same phenomenon do not share sufficient organizational coherence to be explained by a common mechanism (e.g., recognizing variation is not always error and is, in fact, meaningful, such that instances of the same category are caused by different mechanisms). More recently, it has been discovered that instances which have been designated as different phenomena (e.g., "stress," "emotion," and "memory") in fact arise from the same mechanisms, revealing shared regularities across these instances that had been thus far ignored (i.e., recognizing similarities across categories).

Essentialism is a habit of the human mind that is difficult to vanquish, particularly when it comes to thinking about emotion categories. The very enterprise of measuring instances of emotion tempts us with essentialism. As William James put it, "Whenever we have made a word … to denote a certain group of phenomena, we are prone to suppose a substantive entity existing beyond the phenomena, of which the word shall be the name" (James, 1890, p. 195). Essentialism is also difficult to vanquish because when people group instances together and treat them as similar, they are often unaware of their own goals in the process. As a consequence, they come to mistakenly believe that the similarity they perceive is real in nature, and the variation is error in a more universal sense, rather than merely not useful for meeting a particular goal in some particular context (but may very well be useful for some other goal in some other context).

Categorization is necessary: we can't live or do science in a world where everything is different from everything else, where learning in one situation does not allow us to predict in the next. But it's possible to use categories, and measure them, without reifying them with essences.

Appendix A

The last several thousand years of scholarly writing on the nature of emotion can be understood as an ongoing debate between essentialism and construction (of one sort or another). Although a comprehensive history of this literature is beyond the scope of this chapter, a few touch points are instructive for the uninitiated. In the Western scholarly tradition, essentialist approaches to emotion include Plato and Aristotle in Ancient Greece; Descartes (1649/1989) during the Enlightenment, with a theory that foreshadowed Carl Lange (1885/1922) and Damasio and Carvalho (2013); Gall, of phrenology fame (Zola-Morgan, 1995); Irons (1897a, 1897b) and Dewey (1895) who were the first modern classical appraisal theorists; Darwin (who vanquished essentialism in biology with *On the origin of species* but then went on to write a highly essentialized book on emotions a decade later); Carl Lange (1885/1922), who crafted a modern basic emotion theory which Dewey (1895) then tattooed on to William James's (1890) constructionist theory (mangling James's theory to

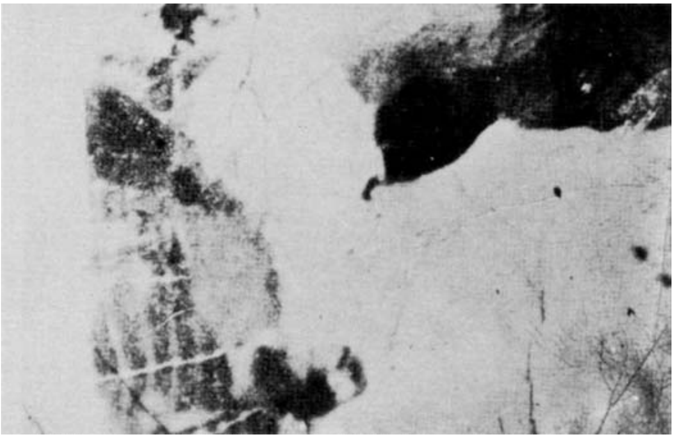

FIG. A.1 An example of categorizing to construct and experience.

create the James–Lange theory and creating a misunderstanding that has survived to this day[g]; McDougall (1923) with a theory very similar to Panksepp's basic emotion theory; Panksepp, 1998); Allport (1922, 1924) who invented the facial feedback hypothesis; Cannon (1927) who localized emotion to a specific brain region; Papez (1937) who expanded the region into a circuit; and MacLean (1949, 1990) who created an elixir of Plato's tripartite mind (rational thoughts, passions (which today we would call emotions), and appetites like hunger and sex drive), Aristotle's phylogenetic scale (scala naturae) idea, and Darwin's ideas from *The descent of man* (1871), wrapped them in neuroanatomy to create the triune brain concept which is still popular today (Panksepp, 1998) (for a review, see Gendron & Barrett, 2009). In the traditional Buddhist view (the Abhidharma school) dating back to the second century BCE, a mind is created from a set of universal, discrete 82 psychophysical elements called *dharmas*, 52 of which are mental; several of the mental dharmas bear a striking resemblance to certain emotion categories that are considered to be biologically basic (anger, pride, joy, shame, regret, jealousy).

Historically, construction is more varied and difficult to summarize, but key ideas can be found in the writings of Heraclitus in Ancient Greece, Ibn al Haythan (in the middle ages), Locke and Kant during the Enlightenment, Spencer (1855), James (1890,1894), and Wundt (1897/1998) during the 19th century (all of whom offered arguments against faculty psychology), as well as Duffy (1934a, 1934b, 1941), Dunlap (1932), Hunt (1941), and Harlow and Stanger (1932, 1933) in the first half of the 20th century (all of whom observed that emotion categories had no essences, and therefore must be constructed as a person makes sense of autonomic changes, although no specific mechanisms or processes were offered). In the past, constructionist ideas on the nature of emotion were often nascent, embedded in broader critiques of typological view. More recently, a new generation of psychological construction theories

g. Notice that William James was a constructionist. He wrote, for example, that "Surely there is no definite affection of 'anger' in an 'entitative' sense" (1894/1994, p. 206), believing instead that each instance of emotion had its own associated physical state. James believed that believing in emotion essences was the psychologist's fallacy. The "trouble with emotions in psychology" he wrote, is that they are regarded too much as absolutely individual things. … But if we regard them as products of more general causes (as "species" are now regarded as products of heredity and variation), the mere distinguishing and cataloging becomes of subsidiary importance (James, 1890/1998, p. 449). Instead, James prescribed an entirely constructionist approach for the study of psychology, including emotion. "A science of the relations of mind and brain" James wrote, "must show how the elementary ingredients of the former correspond to the elementary functions of the latter" (1890/1998, p. 28). Although it is beyond the scope of this chapter, it is fascinating how James's constructionist theory of emotion was transformed (by ignoring the construction- ist features) to integrate it with Lange's highly essentialized vasomotor theory of emotion was strongly essentialistic (each emotion had its own bodily essence), creating the James-Lange theory (first named by Dewey, 1895). So the James–Lange theory is based on a profound misunderstanding of William James.

have emerged, articulating a more detailed and nuanced scientific agenda for the study of emotion (Barrett & Russell, 2015a, 2015b; also see; Lane & Schwartz, 1987; LeDoux, 2012, 2015; Olsson & Ochsner, 2008; Roy, Shohamy & Wager, 2012; Seth, 2013). Buddhist philosophy also has a more constructionist version of the mind. A range of influential thinkers associated with the *Sautrāntika, Madhyamika*, and *Yogācāra* schools (e.g., third century CE through the seventh century CE), the most well-known of which was named Dharmakirti in the seventh century CE who suggested that the dharmas are not basic elements of the human mind, but are themselves creations of that mind, emerging as a function of human concepts (Dreyfus & Thompson, 2007).

References

Adolphs, R. (2017). How should neuroscience study emotions? By distinguishing emotion states, concepts, and experiences. *Soc. Cognit. Affect Neurosci., 12*, 24−31.

Adolphs, R., & Anderson, D. J. (2018). *The neuroscience of emotion: A new synthesis*. Princeton University Press.

Adolphs, R., Mlodinow, L., & Barrett, L. F. (2019). What is an emotion? *Curr. Biol., 29*, R1−R5.

Allport, F. H. (1922). A physiological-genetic theory of feeling and emotion. *Psychol. Rev., 29*, 132−139.

Allport, F. H. (1924). *Social psychology*. New York: Houghton Mifflin.

Arnold, M. B. (1960a). *Emotion and personality: Vol. 1. Psychological aspects*. New York: Columbia University Press.

Arnold, M. B. (1960b). *Emotion and personality: Vol. 2. Physiological aspects*. New York: Columbia University Press.

Atzil, S., Gao, W., Fradkin, I., & Barrett, L. F. (2018). Growing a social brain. *Nat. Hum. Behav., 2*, 624−636. https://doi.org/10.1038/s41562-018-0384-6.

Azari, B., Westlin, C., Satpute, A. B., Hutchinson, J. B., Kragel, P. A., & Barrett, L. F. (2020). Comparing supervised and unsupervised approaches to emotion categorization in the human brain, body and subjective experience. *Scientific Reports, 10*, 20284. https://doi.org/10.1038/s41598-020-77117-8.

Bach, D. R., & Dayan, P. (2017). Algorithms for survival: A comparative perspective on emotions. *Nat. Rev. Neurosci., 18*, 311−319.

Barrett, L. F. (2000). Modeling emotion as an emergent phenomenon: A causal indicator analysis. In *Paper presented at the annual meeting of the society for personality and social psychology, Nashville, TN*. Retrieved from: http://www.affective-science.org/publications.shtml .

Barrett, L. F. (2004). Feelings or words? Understanding the content in self-report ratings of emotional experience. *J. Pers. Soc. Psychol., 87*, 266−281.

Barrett, L. F. (2006a). Emotions as natural kinds? Perspect.Psychol.Sci.,1 , 28−58.

Barrett, L. F. (2006b). Solving the emotion paradox: Categorization and the experience of emotion. *Pers. Soc. Psychol. Rev., 10*(1), 20−46.

Barrett, L. F. (2011a). Was Darwin wrong about emotional expressions? *Curr. Dir. Psychol. Sci., 20*, 400−406.

Barrett, L. F. (2011b). Bridging token identity theory and supervenience theory through psychological construction. *Psychol. Inq., 22*, 115−127.

Barrett, L. F. (2012). Emotions are real. *Emotion, 12*, 413–429.

Barrett, L. F. (2013). Psychological construction: The Darwinian approach to the science of emotion. *Emotion Rev., 5*(4), 379–389.

Barrett, L. F. (2014). The conceptual act theory: A précis. *Emotion Rev., 6*, 292–297.

Barrett, L. F. (2017a). The theory of constructed emotion: An active inference account of interoception and categorization. *Soc. Cognit. Affect Neurosci., 12*(1), 1–23.

Barrett, L. F. (2017b). *How emotions are made: The secret life of the brain*. New York: Houghton Mifflin Harcourt.

Barrett, L. F. (2018). Seeing fear: It's all in the eyes? *Trends Neurosci., 41*, 559–563.

Barrett, L. F. (2020). *John P. McGovern award lecture in the behavioral sciences: Variation is the norm: Darwin's population thinking and the science of emotion*. American Association for the Advancement of Science. https://www.youtube.com/watch?v=FUohKL5WWi8.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychol. Sci. Publ. Interest, 20*(1), 1–68.

Barrett, L. F., & Bliss-Moreau, E. (2009). Affect as a psychological primitive. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 41, pp. 167–218). Burlington, VA: Academic Press.

Barrett, L. F., & Finlay, B. L. (2018). Concepts, goals and the control of survival-related behaviors. *Curr. Opin. Behav. Sci., 24*, 172–179.

Barrett, L. F., Lindquist, K., Bliss-Moreau, E., Duncan, S., Gendron, M., Mize, J., et al. (2007). Of mice and men: Natural kinds of emotion in the mammalian brain? *Perspect. Psychol. Sci., 2*, 297–312.

Barrett, L. F., Mesquita, B., Ochsner, K. N., & Gross, J. J. (2007). The experience of emotion. *Annu. Rev. Psychol., 58*, 373–403.

Barrett, L. F., Ochsner, K. N., & Gross, J. J. (2007). On the automaticity of emotion. In J. Bargh (Ed.), *Social psychology and the unconscious: The automaticity of higher mental processes* (pp. 173–218). New York: Psychology Press.

Barrett, L. F., & Russell, J. A. (1999). The structure of current affect - controversies and emerging consensus. *Curr. Dir. Psychol. Sci., 8*, 10–14.

Barrett, L. F., & Russell, J. A. (2015a). An introduction to psychological construction. In L. F. Barrett, & J. A. Russell (Eds.), *The psychological construction of emotion* (pp. 1–17). New York: Guilford.

Barrett, L. F., & Russell, J. A. (Eds.). (2015b). *The psychological construction of emotion*. New York: Guilford.

Barrett, L. F., & Satpute, A. (2013). Large-scale brain networks in affective and social neuroscience: Towards an integrative architecture of the human brain. *Curr. Opin. Neurobiol., 23*, 361–372.

Barrett, L. F., & Satpute, A. B. (2019). Historical pitfalls and new directions in the neuroscience of emotion. *Neurosci. Lett., 693*, 9–18. https://doi.org/10.1016/j.neulet.2017.07.045.

Barrett, L. F., & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nat. Rev. Neurosci., 16*, 419–429.

Barrett, L. F., Wilson-Mendenhall, C. D., & Barsalou, L. W. (2015). The conceptual act theory: A road map. In L. F. Barrett, & J. A. Russell (Eds.), *The psychological construction of emotion* (pp. 83–110). New York: Guilford.

Barsalou, L. W. (1983). Ad hoc categories. *Mem. Cognit., 11*, 211–227.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *J. Exp. Psychol. Learn. Mem. Cogn., 11*(4), 629–654.

Barsalou, L. W. (2003). Situated simulation in the human conceptual system. *Lang. Cognit. Process., 18*, 513–562.

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol., 59*, 617–645.

Barsalou, L. W., Simmons, W. K., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends Cognit. Sci., 7*, 84–91.

Bloom, P. (2000). *How children learn the meanings of words: Learning, development and conceptual change*. Cambridge, MA: MIT Press.

Brooks, J. A., Chikazoe, J., Sadato, N., & Freeman, J. B. (2019). The neural representation of facial-emotion categories reflects conceptual structure. *Proc. Natl. Acad. Sci. U.S.A., 116*(32), 15861–15870.

Brooks, J. A., Shablack, H., Gendron, M., Satpute, A. B., Parrish, M. H., & Lindquist, K. A. (2017). The role of language in the experience and perception of emotion: A neuroimaging meta-analysis. *Soc. Cognit. Affect Neurosci., 12*(2), 169–183.

Campos, J. J., Mumme, D. L., Kermoian, R., & Campos, R. G. (1994). A functionalist perspective on the nature of emotion. *Monogr. Soc. Res. Child Dev., 59*(2–3), 284–303.

Cannon, W. B. (1927). The James-Lange theory of emotions: A critical examination and an alternative theory. *Am. J. Psychol., 39*, 106–124.

Ceulemans, E., Kuppens, P., & Mechelen, I. V. (2012). Capturing the structure of distinct types of individual differences in the situation-specific experience of emotions: The case of anger. *Eur. J. Pers., 26*(5), 484–495.

Chanes, L., & Barrett, L. F. (2016). Redefining the role of limbic areas in cortical processing. *Trends Cognit. Sci., 20*, 96–106.

Clark, A. (2013). Whatever next? Predictive brains, situated agents and the future of cognitive science. *Behav. Brain Sci., 36*, 181–253.

Clark-Polner, E., Johnson, T., & Barrett, L. F. (2017). Multivoxel pattern analysis does not provide evidence to support the existence of basic emotions. *Cerebr. Cortex, 27*(3), 1944–1948.

Clark-Polner, E., Wager, T. D., Satpute, A. B., & Barrett, L. F. (2016). Neural fingerprinting: Meta-analysis, variation, and the search for brain-based essences in the science of emotion. In L. F. Barrett, M. Lewis, & J. M. Haviland-Jones (Eds.), *The handbook of emotion* (4th ed., pp. 146–165). New York: Guilford.

Clore, G. L., & Ortony, A. (1991). What more is there to emotion concepts than prototypes? *J. Pers. Soc. Psychol., 60*, 48–50.

Clore, G. L., & Ortony, A. (2000). Cognition in emotion: Never, sometimes, or always? In R. D. Lane, & L. Nadel (Eds.), *The cognitive neuroscience of emotion* (pp. 24–61). New York: Oxford University Press.

Clore, G. L., & Ortony, A. (2008). Appraisal theories: How cognition shapes affect into emotion. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *Handbook of emotions* (3rd ed., pp. 628–642). New York: Guilford Press.

Clore, G. L., & Ortony, A. (2013). Psychological construction in the OCC model of emotion. *Emotion Rev., 5*(4), 335–343.

Coan, J. A. (2010). Emergent ghosts of the emotion machine. *Emotion Rev., 2*, 274–285.

Cordaro, D. T., Sun, R., Keltner, D., Kamble, S., Huddar, N., & McNeil, G. (2018). Universals and cultural variations in 22 emotional expressions across five cultures. *Emotion, 18*(1), 75–93.

Cowen, A. S., Elfenbein, H. A., Laukka, P., & Keltner, D. (2018). Mapping 24 emotions conveyed by brief human vocalization. *Am. Psychol., 74*(6), 698–712.

Cowen, A. S., Fang, X., Sauter, D., & Keltner, D. (2020). What music makes us feel: At least 13 dimensions organize subjective experiences associated with music across different cultures. *Proc. Natl. Acad. Sci. U.S.A., 117*(4), 1924–1934.

Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proc. Natl. Acad. Sci. U.S.A., 114*(38), E7900−E7909.

Cowen, A. S., & Keltner, D. (2020a). Universal facial expressions uncovered in art of the ancient Americas: A computational approach. *Sci. Adv., 6*(34), eabb1005.

Cowen, A. S., & Keltner, D. (2020b). What the face displays: Mapping 28 emotions conveyed by naturalistic expression. *Am. Psychol., 75*(3), 349−364.

Damasio, A., & Carvalho, G. B. (2013). The nature of feelings: Evolutionary and neurobiological origins. *Nat. Rev. Neurosci., 14*, 143−152.

Darwin, C. (1964). *On the origin of species*. Harvard University Press (Original work published in 1859).

Darwin, C. (1965). *The expression of the emotions in man and animals*. Chicago, IL: University of Chicago Press (Original work published 1872).

Dewey, J. (1895). The theory of emotion. *Psychol. Rev., 2*(1), 13.

Dreyfus, G., & Thompson, E. (2007). Asian perspectives: Indian theories of mind. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *Cambridge handbook of consciousness* (pp. 88−114). Cambridge: Cambridge University Press.

Duffy, E. (1934a). Is emotion a mere term of convenience? *Psychol. Rev., 41*, 103−104.

Duffy, E. (1934b). Emotion: An example of the need for reorientation in psychology. *Psychol. Rev., 41*, 184−198.

Duffy, E. (1941). An explanation of "emotional" phenomena without use of the concept "emotion". *General J. Psychol., 25*, 283−293.

Dunlap, K. (1932). Are emotions teleological constructs? *Am. J. Psychol., 44*, 572−576.

Edelman, G. M. (1998). *The remembered present: A biological theory of consciousness*. New York: Basic Books.

Edelman, G. M., & Gally, J. A. (2001). Degeneracy and complexity in biological systems. *Proc. Natl. Acad. Sci. U.S.A., 98*(24), 13763−13768.

Ekman, P. (1972). Universal and cultural differences in facial expressions of emotions. In K. Cole (Ed.), *Nebraska symposium on motivation* (pp. 207−283). Lincoln, NE: University of Nebraska Press, 1971.

Ekman, P. (1992). An argument for basic emotions. *Cognit. Emot., 6*, 169−200.

Ekman, P., & Cordaro, D. (2011). What is meant by calling emotions basic. *Emotion Rev., 3*, 364−370.

Elfenbein, H. A. (2013). Nonverbal dialects and accents in facial expressions of emotion. *Emotion Rev., 5*, 90−96.

Elfenbein, H. A. (2017). Emotional dialects in the language of emotion. In J.-M. Fernández-Dols, & J. A. Russell (Eds.), *The science of facial expression* (pp. 479−496). New York, NY: Oxford University Press.

Farahvash, M. R., Abianeh, S. H., Farahvash, B., Farahvash, Y., Yagoobi, A., & Nazparvar, B. (2010). Anatomic variations of midfacial muscles and nasolabial crease: A survey on 52 hemifacial dissections in fresh Persian cadavers. *Aesthetic Surg. J., 30*(1), 17−21.

Fedyk, M., & Xu, F. (2018). The epistemology of rational constructivism. *Rev. Philos. Psychol., 9*, 343−362.

Fridlund, A. J. (1992). Darwin's anti-darwinism and the expression of the emotions in man and animals. In K. T. Strongman (Ed.), *International review of emotion* (Vol. 2, pp. 117−137). New York: Wiley.

Fridlund, A. J. (2017). The behavioral ecology view of facial displays, 25 years later. In J.-M. Fernández-Dols, & J. A. Russell (Eds.), *The science of facial expression* (pp. 77−92). Oxford University Press.

Frijda, N. H. (1986). *The emotions*. New York: Cambridge University Press.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci., 11*, 127−138.

Gelman, S. A., & Rhodes, M. (2012). Two-thousand years of stasis. How psychological essentialism impedes evolutionary understanding. In K. R. Rosengren, S. Brem, E. M. Evans, & G. Sinartra (Eds.), *Evolution challenges integrating research and practice in teaching and learning about evolution* (pp. 3−21). New York: Oxford University Press.

Gendron, M., & Barrett, L. F. (2009). Reconstructing the past: A century of ideas about emotion in psychology. *Emotion Rev., 1*, 316−339.

Gendron, M., Crivelli, C., & Barrett, L. F. (2018). Universality reconsidered: Diversity in making meaning of facial expressions. *Curr. Dir. Psychol. Sci., 27*(4), 211−219.

Gendron, M., Hoemann, K., Crittenden, A. N., Mangola, S. M., Ruark, G. A., & Barrett, L. F. (2020). Emotion perception in Hadza hunter-gatherers. *Sci. Rep., 10*(1), 3867.

Gendron, M., Mesquita, B., & Barrett, L. F. (2020). The brain as a cultural artifact. In L. J. Kirmayer, C. M. Worthman, S. Kitayama, R. Lemelson, & C. A. Cummings (Eds.), *Culture, Mind, and Brain: Emerging Concepts, Models, and Applications* (pp. 188−222). New York: Cambridge University Press.

Gendron, M., Mesquita, B., & Barrett, L. F. (2013). Emotion perception. In D. Reisberg (Ed.), *Oxford handbook of cognitive psychology* (pp. 539−556). New York: Oxford University Press.

Greenwald, A. G., & Ronis, D. L. (1981). On the conceptual disconfirmation of theories. *Pers. Soc. Psychol. Bull., 7*(1), 131−137.

Gross, J. J. (2015). Emotion regulation: Current status and future prospects. *Psychol. Inq., 26*, 1−26.

Gross, J. J., & Barrett, L. F. (2011). Emotion generation and emotion regulation: One or two depends on your point of view. *Emotion Rev., 3*(1), 8−16.

Guillory, S. A., & Bujarski, K. A. (2014). Exploring emotion using invasive methods: Review of 60 years of human intracranial electrophysiology. *Soc. Cognit. Affect Neurosci., 9*, 1880−1889.

Harlow, H. F., & Stagner, R. (1932). Psychology of feelings and emotions. I. Theory of feelings. *Psychol. Rev., 39*(6), 570−589.

Harlow, H. F., & Stagner, R. (1933). Psychology of feelings and emotions. II. Theory of emotions. *Psychol. Rev., 40*, 184−195.

Harmon-Jones, E., Harmon-Jones, C., Abramson, L., & Peterson, C. K. (2009). PANAS positive activation is associated with anger. *Emotion, 9*, 183−196.

Hoemann, K., Crittenden, A. N., Msafiri, S., Liu, Q., Li, C., Roberson, D., et al. (2019). Context facilitates performance on a classic cross-cultural emotion perception task. *Emotion, 19*(7), 1292−1313.

Hoemann, K., Khan, Z., Feldman, M., Nielson, C., Devlin, M., Dy, J., et al. (2020). Context-aware experience sampling reveals the scale of variation in affective experience. *Sci. Rep., 10*, 12459.

Hohwy, J. (2013). *The predictive mind.* Oxford: Oxford University Press.

Horikawa, T., Cowen, A. S., Keltner, D., & Kamitani, Y. (2020). The neural representation of visually evoked emotion is high-dimensional, categorical, and distributed across transmodal brain regions. *iScience, 23*(5), 101060.

Hortensius, R., Schutter, D. J., & Harmon-Jones, E. (2011). When anger leads to aggression: Induction of relative left frontal cortical activity with transcranial direct current stimulation increases the anger−aggression relationship. *Soc. Cognit. Affect Neurosci., 7*(3), 342−347.

Hunt, W. A. (1941). Recent developments in the field of emotion. *Psychol. Bull., 38*(5), 249−276.

Hutchinson, J. B., & Barrett, L. F. (2019). The power of predictions: An emerging paradigm for psychological research. *Curr. Dir. Psychol. Sci., 28*, 280−291.

Irons, D. (1897a). The nature of emotion. *Phil. Rev., 6*, 242–256.

Irons, D. (1897b). The primary emotions. *Phil. Rev., 6*, 626–645.

Izard, C. E. (1993). Four systems for emotion activation: Cognitive and noncognitive processes. *Psychol. Rev., 100*, 68–90.

Izard, C. E. (2007). Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspect. Psychol. Sci., 2*(3), 260–280.

Izard, C. E. (2010). The many meanings/aspects of emotion: Definitions, functions, activation, and regulation. *Emotion Rev., 2*(4), 363–370.

Izard, C. E. (2011). Forms and functions of emotions: Matters of emotion–cognition interactions. *Emotion Rev., 3*, 371–378.

Izard, C. E., Woodburn, E. M., & Finlon, K. J. (2010). Extending emotion science to the study of discrete emotions in infants. *Emotion Rev., 2*(2), 134–136.

Jackson, J. C., Watts, J., Henry, T. R., List, J., Forkel, R., Mucha, P. J., et al. (2019). Emotion semantics show both cultural variation and universal structure. *Science, 366*(6472), 1517–1522.

James, W. (1890). *The principles of psychology* (Vol. 1). New York: Holt.

James, W. (1894). The physical basis of emotion. *Psychol. Rev., 1*, 516–529.

James, W. (1998). *The principals of psychology.* Bristol: Thoemmes Press (Original work published 1890).

Kassam, K. S., Markey, A. R., Cherkassky, V. L., Loewenstein, G., & Just, M. A. (2013). Identifying emotions on the basis of neural activation. *PLoS One, 8*, e66032.

Kleckner, I. R., Zhang, J., Touroutoglou, A., Chanes, L., Xia, C., Simmons, W. K., et al. (2017). Evidence for a large-scale brain system supporting allostasis and interoception in humans. *Nat. Human Behav., 1*, 0069.

Kornblith, H. (1993). *Inductive inference and its natural ground: An essay in naturalistic epistemology.* Cambridge, MA: The MIT Press.

Kragel, P. A., Knodt, A. R., Hariri, A.,R., & LaBar, K. S. (2016). Decoding spontaneous emotional states in the human brain. *PLoS Biol., 14*(9), e2000106.

Kragel, P. A., Koban, L., Barrett, L. F., & Wager, T. D. (2018). Representation, pattern information, and brain signatures: From neurons to neuroimaging. *Neuron, 99*, 257–273.

Kragel, P. A., & LaBar, K. S. (2013). Multivariate pattern classification reveals autonomic and experiential representations of discrete emotions. *Emotion, 13*(4), 681–690.

Kragel, P. A., & LaBar, K. S. (2015). Multivariate neural biomarkers of emotional states are categorically distinct. *Soc. Cognit. Affect Neurosci., 10*(11), 1437–1448.

Kuppens, P., Van Mechelen, I., & Rijmen, F. (2008). Towards disentangling sources of individual differences in appraisal and anger. *J. Pers., 76*, 1–32.

Kuppens, P., Van Mechelen, I., Smits, D. J. M., De Boeck, P., & Ceulemans, E. (2007). Individual differences in patterns of appraisal and anger experience. *Cognit. Emot., 21*, 689–713.

Lane, R., & Schwartz, G. E. (1987). Levels of emotional awareness: A cognitive-developmental theory and its application to psychopathology. *Am. J. Psychiatr., 144*, 133–143.

Lange, C. G. (1922). The emotions: A psychophysiological study (I. A. Haupt, trans.). In Dunlap (Ed.), *The emotions* (pp. 33–90). Baltimore, MD: Williams & Wilkins Company (Original work published 1885).

Lazarus, R. S. (1966). *Psychological stress and the coping process*. New York: McGraw.

Lazarus, R. S. (1991). *Emotion and adaptation*. New York: Oxford University Press.

LeBois, L. A. M., Wilson-Mendenhall, C. D., Simmons, W. K., Barrett, L. F., & Barsalou, L. W. (2018). Learning situated emotions. *Neuropsychologia, 39*(8), 1764–1801.

LeDoux, J. (2012). Rethinking the emotional brain. *Neuron, 73*, 653–676.

LeDoux, J. (2014). Coming to terms with fear. *Proc. Natl. Acad. Sci. U.S.A., 111*(8), 2871−2878.

LeDoux, J. (2015). *Anxious: Using the brain to understand and treat fear and anxiety.* New York: Viking.

Levenson, R. W. (2011). Basic emotion questions. *Emotion Rev., 3*(4), 379−386.

Lewis, M. D. (2005). Bridging emotion theory and neurobiology through dynamic systems modeling (target article). *Behav. Brain Sci., 28*, 169−194.

Lewontin, R. C. (2000). *The triple helix: Gene, organism, and environment*. Cambridge, MA: Harvard University Press.

Lindquist, K. A., & Barrett, L. F. (2012). A functional architecture of the human brain: Insights from emotion. *Trends Cognit. Sci., 16*, 533−540.

Lindquist, K. A., Siegel, E. H., Quigley, K., & Barrett, L. F. (2013). The hundred years emotion war: Are emotions natural kinds or psychological constructions? Comment on Lench, Flores, & Bench (2011). *Psychol. Bull., 139*, 255−263.

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Barrett, L. F. (2012). The brain basis of emotion: A meta-analytic review. *Behav. Brain Sci., 35*, 121−143.

MacCormack, J. K., & Lindquist, K. A. (2019). Feeling hangry? When hunger is conceptualized as emotion. *Emotion, 19*(2), 301−319.

MacLean, P. D. (1949). Psychosomatic disease and the "visceral brain": Recent developments bearing on the Papez theory of emotion. *Psychosom. Med., 11*, 338−353.

MacLean, P. D. (1990). *The triune brain in evolution: Role in paleocerebral functions*. New York: Plenum Press.

Marder, E., & Taylor, A. L. (2011). Multiple models to capture the variability in biological neurons and networks. *Nat. Neurosci., 14*(2), 133−138.

Matsumoto, D. (1990). Cultural similarities and differences in display rules. *Motiv. Emot., 14*, 195−214.

Matsumoto, D., Keltner, D., Shiota, M., O'Sullivan, M., & Frank, M. (2008). Facial expressions of emotions. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *Handbook of emotions* (3rd ed., pp. 211−234). New York, NY: Macmillan.

Mayr, E. (2004). *What makes biology unique?* New York, NY: Cambridge University Press.

McDougall, W. (1923). *An outline of psychology.* London: Methuen.

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou, & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179−195). New York: Cambridge University Press.

Mesquita, B., Boiger, M., & De Leersnyder, J. (2016). The cultural construction of emotions. *Curr. Opin. Psychol., 8*, 31−36.

Minxha, J., Mosher, C., Morrow, J. K., Mamelak, A. N., Adolphs, R., Gothard, K. M., et al. (2017). Fixations gate species-specific responses to free viewing of faces in the human and macaque amygdala. *Cell Rep., 18*(4), 878−891.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Nezlek, J. B., Vansteelandt, K., Van Mechelen, I., & Kuppens, P. (2008). Appraisal-emotion relationships in daily life. *Emotion, 8*(1), 145−150.

Obrist, P. A. (1981). *Cardiovascular psychophysiology.* New York, NY: Plenum Press.

Obrist, P. A., Webb, R. A., Sutterer, J. R., & Howard, J. L. (1970). The cardiac-somatic relationship: Some reformulations. *Psychophysiology, 6*(5), 569−587.

Olsson, A., & Ochsner, K. N. (2008). The role of social cognition in emotion. *Trends Cognit. Sci., 12*, 65−71.

Ortony, A., & Clore, G. (2015). Can an appraisal model be compatible with psychological constructionism? In L. F. Barrett, & J. A. Russell (Eds.), *The psychological construction of emotion* (pp. 305–333). New York: Guilford.

Ortony, A., & Turner, T. J. (1990). What's basic about basic emotions? *Psychol. Rev., 97*(3), 315–331.

Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions*. New York: Oxford University Press.

Papez, J. W. (1937). A proposed mechanism of emotion. *Arch. Neurol. Psychiatry, 38*, 725–743.

Park, B. J., Jang, E. H., Chung, M. A., & Kim, S. H. (2013). Design of prototype-based emotion recognizer using physiological signals. *ETRI J., 35*(5), 869–879.

Pavlenko, A. (2014). *The bilingual mind: And what it tells us about language and thought*. Cambridge: Cambridge University Press.

Pessa, J. E., Zadoo, V. P., Adrian, E. K., Yuan, C. H., Aydelotte, J., & Garza, J. R. (1998). Variability of the midfacial muscles: Analysis of 50 hemifacial cadaver dissections. *Plast. Reconstr. Surg., 102*, 1888–1893.

Pinker, S. (1997). *How the mind works*. New York, NY: W. W. Norton & Company.

Price, C. J., & Friston, K. J. (2002). Degeneracy and cognitive anatomy. *Trends Cognit. Sci., 6*(10), 416–421.

Quigley, K. S., & Barrett, L. F. (2014). Is there consistency and specificity of autonomic changes during emotional episodes? Guidance from the conceptual act theory and psychophysiology. *Biol. Psychol., 98*, 82–94. https://doi.org/10.1016/j.biopsycho.2013.12.013

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature, 435*(7045), 1102–1107.

Rainville, P., Bechara, A., Naqvi, N., & Damasio, A. R. (2006). Basic emotions are associated with distinct patterns of cardiorespiratory activity. *Int. J. Psychophysiol., 61*(1), 5–18.

Roseman, I. J. (2011). Emotional behaviors, emotivational goals, emotion strategies: Multiple levels of organization integrate variable and consistent responses. *Emotion Rev., 3*(4), 434–443.

Roy, M., Shohamy, D., & Wager, T. D. (2012). Ventromedial prefrontal-subcortical systems and the generation of affective meaning. *Trends Cognit. Sci., 16*(3), 147–156.

Russell, J. A. (1980). A circumplex model of affect. *J. Pers. Soc. Psychol., 39*(6), 1161–1178.

Russell, J. A. (1991). Culture and the categorization of emotions. *Psychol. Bull., 110*(3), 426–450.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychol. Rev., 110*, 145–172.

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *J. Pers. Soc. Psychol., 76*(5), 805–819.

Saarimäki, H., Ejtehadian, L. F., Glerean, E., Jääskeläinen, I. P., Vuilleumier, P., Sams, M., et al. (2018). Distributed affective space represents multiple emotion categories across the human brain. *Soc. Cognit. Affect Neurosci., 13*(5), 471–482.

Saarimäki, H., Gotsopoulos, A., Jääskeläinen, I. P., Lampinen, J., Vuilleumier, P., Hari, R., et al. (2016). Discrete neural signatures of basic emotions. *Cerebr. Cortex, 26*(6), 2563–2573.

Sander, D., Grandjean, D., & Scherer, K. R. (2005). A systems approach to appraisal mechanisms in emotion. *Neural Networks, 18*(4), 317–352.

Satpute, A. B., & Lindquist, K. A. (2019). The default mode network's role in discrete emotion. *Trends Cognit. Sci., 23*(10), 851–864.

Scarantino, A., & Griffiths, P. (2009). Emotions in the wild: The situated perspective on emotion. In P. Robbins, & M. Aydede (Eds.), *Cambridge handbook of situated cognition* (pp. 437–453). Cambridge University Press.

Scherer, K. R. (1999). Appraisal theory. In T. Dalgeish, & M. J. Power (Eds.), *Handbook of cognition and emotion* (pp. 637–663). Wiley.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Soc. Sci. Inf., 44*(4), 695–729.

Scherer, K. R. (2009). Emotions are emergent processes: They require a dynamic computational architecture. *Philos. Trans. R. Soc. Lond. B Biol. Sci., 364*(1535), 3459–3474.

Scherer, K. R., Mortillaro, M., & Mehu, M. (2017). Facial expression is driven by appraisal and generates appraisal inference. In J.-M. Fernández-Dols, & J. A. Russell (Eds.), *The science of facial expression* (pp. 353–373). Oxford University Press.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends Cognit. Sci., 17*, 565–573.

Seth, A. K., & Friston, K. J. (2016). Active interoceptive inference and the emotional brain. *Phil. Trans. Roy. Soc. Lond. B Biol. Sci., 371*(1708), 20160007.

Shimada, K., & Gasser, R. F. (1989). Variations in the facial muscles at the angle of the mouth. *Clin. Anat., 2*, 129–134.

Shim, K.-S., Hu, K.-S., Kwak, H.-H., Koh, K.-S., Fontaine, C., & Kim, H.-J. (2008). An anatomical study of the insertion of the zygomaticus major muscle in humans focused on the muscle arrangement at the corner of the mouth. *Plast. Reconstr. Surg., 121*, 466–473.

Siegel, E. H., Sands, M. K., Condon, P., Chang, Y., Dy, J., Quigley, K. S., et al. (2018). Emotion fingerprints or emotion populations? A meta-analytic investigation of autonomic features of emotion categories. *Psychol. Bull., 144*(4), 343–393.

Spencer, H. (1855). *Principals of psychology.* London: Longman, Brown, Green & Longmans.

Stemmler, G., Aue, T., & Wacker, J. (2007). Anger and fear: Separable effects of emo tion and motivational direction on somatovisceral responses. *Int. J. Psychophysiol., 66*(2), 141–153.

Stephens, C. L., Christie, I. C., & Friedman, B. H. (2010). Autonomic specificity of basic emotions: Evidence from pattern classification and cluster analysis. *Biol. Psychol., 84*(3), 463–473.

Striedter, G. F. (2005). *Principles of brain evolution*. Sutherland, MA: Sinauer Associates.

Strimbu, K., & Tavel, J. A. (2010). What are biomarkers? *Curr. Opin. HIV AIDS, 5*(6), 463.

Tian, Y.-I., Kanade, T., & Cohn, J. F. (2001). Recognizing action units for facial expression analysis. *Trans. Pattern Anal. Mach. Intell., 23*, 1–19.

Tomkins, S. S., & McCarter, R. (1964). What and where are the primary affects? Some evidence for a theory. *Percept. Mot. Skills, 18*, 119–158.

Tononi, G., Edelman, G. M., & Sporns, O. (1998). Complexity and coherency: Integrating information in the brain. *Trends Cognit. Sci., 2*(12), 474–484.

Tononi, G., Sporns, O., & Edelman, G. M. (1999). Measures of degeneracy and redundancy in biological networks. *Proc. Natl. Acad. Sci. U.S.A., 96*(6), 3257–3262.

Touroutoglou, A., Lindquist, K. A., Dickerson, B. C., & Barrett, L. F. (2015). Intrinsic connectivity in the human brain does not reveal networks for "basic" emotions. *Soc. Cognit. Affect Neurosci., 10*(9), 1257–1265. https://doi.org/10.1093/scan/nsv013/

Tracy, J. L., & Randles, D. (2011). Four models of basic emotions: A review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt. *Emotion Rev., 3*, 397.

Viskontas, I. V., Quiroga, R. Q., & Fried, I. (2009). Human medial temporal lobe neurons respond preferentially to personally relevant images. *Proc. Natl. Acad. Sci. U.S.A., 106*, 21329–21334.

Voorspoels, W., Vanpaemel, W., & Storms, G. (2011). A formal ideal-based account of typicality. *Psychon. Bull. Rev., 18*, 1006–1014.

Wager, T. D., Kang, J., Johnson, T. D., Nichols, T. E., Satpute, A. B., & Barrett, L. F. (2015). A Bayesian model of category-specific emotional brain responses. *PLoS Comput. Biol., 11*(4), e1004066. https://doi.org/10.1371/.journal.pcbi.1004066

Waller, B. M. (2008). Selection for universal facial emotion. *Emotion, 8*, 435−439.

Widen, S. C., & Russell, J. A. (2013). Children's recognition of disgust in others. *Psychol. Bull., 139*(2), 271−299.

Wilson-Mendenhall, C. D., Barrett, L. F., & Barsalou, L. W. (2013). Neural evidence that human emotions share core affective properties. *Psychol. Sci., 24*(6), 947−956.

Wilson-Mendenhall, C. D., Barrett, L. F., & Barsalou, L. W. (2015). Variety in emotional life: Within-category typicality of emotional experiences is associated with neural activity in large-scale brain networks. *Soc. Cognit. Affect Neurosci., 10*, 62−71.

Wilson-Mendenhall, C. D., Barrett, L. F., Simmons, W. K., & Barsalou, L. W. (2011). Grounding emotion in situated conceptualization. *Neuropsychologia, 49*(5), 1105−1127.

Wilson-Mendenhall, C. D., Henriques, A., Barsalou, L. W., & Barrett, L. F. (2019). Primary interoceptive cortex activity during simulated experiences of the body. *J. Cognit. Neurosci., 31*, 221−235.

Wundt, W. (1897). In C. H. Judd (Ed.), *Outlines of psychology.* Leipzig: Wilhelm Engelmann. Trans.

Wundt, W. (1998). Trans. In C. H. Judd (Ed.), *Outlines of psychology.* Bristol, UK: Thoemmes Press (Original work published 1897).

Xu, F. (2016). Preliminary thoughts on a rational constructivist approach to cognitive development: Primitives, symbols, learning, and thinking. In D. Barner, & A. S. Baron (Eds.), *Core knowledge and concept change*. New York, NY: Oxford University Press.

Yuen, K. S. L., Johnston, S. J., De Martino, F., Sorger, B., Formisano, E., Linden, D. E. J., et al. (2012). Pattern classification predicts individuals responses to affective stimuli. *Transl. Neurosci., 3*, 278−287.

Zola-Morgan, S. (1995). Localization of brain function: The legacy of Franz Joseph Gall (1758−1828). *Annu. Rev. Neurosci., 18*(1), 359−383.