The importance of using multiple outcome measures in infant research

Article (Accepted Version)

The importance of using multiple outcome measures in infant research

Vanessa LoBue[1], Lori B. Reider[1], Emily Kim[1], Jessica L. Burris[1], Denise S. Oleas[1], Kristin A. Buss[2], Koraly Pérez-Edgar[2], & Andy P. Field[3]

Rutgers University[1], Pennsylvania State University[2], University of Sussex[3]

Keywords: multiple measures; infancy; looking time; behavioral responses; threat detection

Corresponding Author: Vanessa LoBue, 101 Warren Street, Room 301, Newark, NJ 07102, vlobue@psychology.rutgers.edu, 973-353-3950, http://childstudycenter.rutgers.edu/

Abstract

Collecting data with infants is notoriously difficult. As a result, many of our studies consist of small samples, with only a single measure, in a single age group, at a single time point. With renewed calls for greater academic rigor in data collection practices, using multiple outcome measures in infant research is one way to increase rigor, and at the same time, enable us to more accurately interpret our data. Here, we illustrate the importance of using multiple measures in psychological research with examples from our own work on rapid threat detection, and from the broader infancy literature. First, we describe our initial studies using a single outcome measure, and how this strategy caused us to nearly miss a rich and complex story about attention biases for threat and their development. We demonstrate how using converging measures can help researchers make inferences about infant behavior, and how using additional measures allows us to more deeply examine the mechanisms that drive developmental change. Finally, we provide practical and statistical recommendations for how researchers can use multiple measures in future work.

Collecting data with infants is notoriously difficult. Infants are nonverbal, noncompliant, easily distracted, and difficult to predict. They can't talk, they can't follow directions, they get bored easily, they cry and fuss if they are tired or hungry, and the youngest ones can't really do that much at all. When there are limits to what our participants can do, we are left with a limited number of potential dependent measures on which to rely for our research. As a result, many studies with infant participants use only a single measure, in a single age group, at a single time point, to draw broad inferences about behavior and the mechanisms that underlie its development. In this paper, we argue that converging evidence from multiple outcome measures has benefits for building more robust models of the mechanisms that guide developmental change, and ultimately help us to more accurately characterize the richness of infant development. By using singular measures, we run the risk of making inferences that mischaracterize behavior, and attribute singular explanations to phenomena that might in reality be multi-faceted and deeply complex.

The recommendation that developmental researchers should implement designs with multiple outcome measures is not new (e.g., Aslin, 2007; Buss, 2011; LoBue & Adolph, 2019; Morris, Robinson, & Eisenberg, 2006). Indeed, Jerome Kagan and colleagues published a 2002 paper entitled "One measure, one meaning: Multiple measures, clearer meaning," calling for the addition of behavioral measures to research that, at the time, relied heavily on self-report alone (Kagan, Snidman, McManis, Wooward, & Hardway, 2002). We aim to expand on previous recommendations by providing detailed examples for when additional measures can be useful, and how to implement them in the lab, with a specific emphasis on infant research.

We begin with an analysis of how using multiple outcome measures helps us build richer theoretical models of development drawing upon examples from our own work and from the

broader infant literature. First, we draw from our own research on rapid attention to threat to describe how we nearly missed a rich and complex developmental story by relying on a single outcome—latency to look at, or detect, a target stimulus—for so long. We then describe how using additional measures added complexity to our once simple story, allowing us to make more accurate *inferences* about our attention measures, and draw conclusions about the *mechanisms* that drive the development of rapid threat detection. We describe examples from infancy to adulthood, as using multiple measures is important for research in all domains and across all ages. However, along the way, we emphasize why multiple measures are particularly important for work with preverbal infants, drawing from the existing literature. We close by discussing the statistical implications of this approach, and by providing some recommendations and strategies for how we can more commonly use multiple outcome measures in future infant work.

Threat Detection Across the Lifespan—One Simple Measure, One Simple Story

Researchers have been interested in humans' rapid responses to threat for decades. The ability to detect signals of threat in the environment quickly and efficiently has clear adaptive value in helping an individual escape from potential danger. Mistakes based on trial and error in this domain could be costly, possibly resulting in death, so several researchers have suggested that there would be an evolutionary advantage to rapidly detecting threatening stimuli early in development (Bolles, 1970; Boyer & Bergstrom, 2011). As a result, some believe that humans have an evolved fear module for the rapid detection of threatening stimuli (Öhman & Mineka, 2001). According to this view, evolutionarily recurrent threats—dangerous animals like snakes and spiders and threatening conspecifics—should be detected automatically, without the need for cognitive processing.

Support for this contention comes from several studies demonstrating that adults detect the presence of snakes, spiders, and threatening human faces (i.e., angry faces) more quickly than neutral stimuli. In the general adult paradigm, participants are presented with 2 × 2 and 3 × 3 matrices containing 4 or 9 photos from a single stimulus category, or matrices with 3 or 8 photos from a single stimulus category and one additional discrepant (target) image from a second category. Typically, researchers reported that adults detect (via button-press responses) discrepant snakes and spiders more quickly than flowers or mushrooms (Öhman, Flykt, & Esteves, 2001) and discrepant angry faces more quickly than happy or neutral faces (e.g., Öhman, Lundqvist, & Esteves, 2001).

These findings have been widely replicated (see LoBue, 2016, and LoBue & Rakison, 2013 for reviews), providing strong support for the evolutionary perspective on rapid threat detection. However, despite widespread claims of evolutionary origins for rapid threat detection, most of this work had been done only with adults. If humans evolved a predisposition to detect threatening stimuli rapidly, such a propensity should be evident much earlier in development. To investigate this question, we embarked on the first empirical studies of rapid threat detection in children. We modified the standard button press paradigm by presenting participants with 3 × 3 matrices on a touchscreen monitor. Only target-present matrices were included, so a participant's sole task was to find the single image and touch it on the screen as quickly as possible. This modification made the paradigm suitable for children as young as 3, and we found that children between the ages of 3 and 5 and adults detected snakes more quickly than flowers, frogs, and caterpillars (LoBue & DeLoache, 2008), and spiders faster than mushrooms and cockroaches (LoBue, 2010a). Children and adults also detected negative facial expressions—sad, fearful, and angry—more quickly than happy faces; further, they detected negative threat-relevant faces (i.e.,

angry, fearful) more quickly than negative non-threat-relevant faces (i.e., sad) (LoBue, 2009), consistent with the evolutionary perspective.

Others have replicated these findings with preschool-aged children and extended them to both color and black and white photos of the stimuli (Hayakawa, Kawai, & Masataka, 2011; LoBue & DeLoache, 2011; Masataka, Hawakawa, & Kawai, 2010). One group of researchers found stronger effects when they depicted snakes in an attack pose (Masataka, Hawakawa, & Kawai, 2010), and using this same touchscreen visual search paradigm, reported that Japanese monkeys detect snakes more quickly than flowers (Shibasaki & Kawai, 2009). Further, by simply presenting infants with two images side by side on a large screen—one threatening and one non-threatening—we found that even 8- to 14-month-olds looked more quickly to snakes than flowers, and more quickly to angry versus happy faces (LoBue & DeLoache, 2010).

Altogether, this body of work has been interpreted as providing support for the evolutionary perspective on threat detection, demonstrating that infants, children, and adults detect a variety of threatening stimuli—including non-social threats like snakes and spiders, and social threats like angry faces—faster than a wide variety of neutral stimuli. These findings were easily replicated, and consistent across ages and categories of threat.

However, other published data cast doubt on our original conclusions. First, several researchers demonstrated that adults detect a *variety* of threats, including stimuli like guns, knives, and syringes, more quickly than perceptually-matched neutral stimuli (Blanchette, 2006; Brosch & Sharma, 2005). Not enough time has passed for humans to evolve a predisposition to detect guns and knives quickly, so biased attention to these stimuli would have to be acquired through learning. We examined this question in our own lab by studying the detection of two modern threats that adults detect particularly quickly—syringes and knives—in 3-year-olds using

the touchscreen paradigm described above. Importantly, parents confirmed that all of the children we tested had negative experience with syringes through painful vaccinations, but none of them had negative experience with a knife. If learning could account for rapid threat detection, we predicted that children would detect syringes more quickly than perceptually matched controls (i.e. pens), but they would not show an attentional bias for knives (vs. spoons). The results confirmed this hypothesis, as the children detected syringes faster than pens, but did not detect knives faster than spoons (LoBue, 2010b). This work suggests that attention biases for *any* stimulus can be learned, and in fact, other research has shown that adults can develop an attention bias for any stimulus that has personal relevance to them, even images from the popular British TV show, *Dr. Who* (Purkis, Lester, & Field, 2011). These findings leave open the possibility that learning could account for many of the results we had found thus far.

Another seemingly inconsistent finding was that in some studies, low-level perceptual features of some commonly used threatening stimuli could produce rapid detection even when presented in a non-threatening context. For example, the "V" shape common of the brow of an angry face is enough to elicit rapid detection in adults (Larson, Aronoff, & Stearns, 2007). In contrast, other studies reported that the "V" shaped brow presented without a face-like context was not enough to elicit rapid detection (e.g., Schubö, Gendolla, Meinecke, & Abele, 2006; Tipples, Atkinson, & Young, 2002). In our own work, we found that 3-year-olds and adults indeed detected the "V" shape common of angry faces more quickly than an inverted "V" (LoBue & Larson, 2010). Likewise, 3-year-olds and adults detected snakes more quickly than other stimuli only if they were presented in a coiled position; they did not detect snakes more quickly than frogs when only their faces were shown, and they also detected other coiled

stimuli—like coiled hoses and wires—more quickly than non-coiled stimuli (LoBue & DeLoache, 2011).

So although a large body of research, including our own, suggests that infants, children, and adults all detect a variety of evolutionary threats more quickly than neutral control stimuli, a broader review of the literature reveals that attentional biases can also be learned, and might be driven by lower-order perceptual features other than threatening valence. Thus, while using a single paradigm (i.e., visual search) with a single outcome measure (i.e., latency to indicate that a target was detected) did provide us with important information about what kinds of stimuli are detected faster than others and were suggestive of some of the factors that might lead to rapid detection, this limited approach could only take us so far. Indeed, many of the studies from this literature were designed only to ask dichotomous questions—*Is rapid detection driven by threatening valence or by low-level stimulus features*? *Is rapid detection innate or learned?* These simple designs were bound to give us simple answers, and problematically, they could not test for the possibility that rapid attention might mean different things for different stimuli, and that the mechanisms underlying biased attention might be more complex than dichotomous questions will allow (LoBue, 2016). To go beyond these simple dichotomies, we had to start using multiple outcome measures within the same, and across different paradigms.

Converging Measures—Inferences and Complex Constructs

One question left open from previous research is whether rapid detection is driven by valence or by some low-level feature of the test stimuli. Researchers have long-assumed that rapid detection is akin to vigilance, and is thus related to the threat-relevance of the target stimuli. This is an *inference* that we are making about the meaning of latency to detect, or look

at, a target stimulus, and one that is common in both the threat detection literature and in the infancy literature more broadly.

One way to address problems of inference with singular dependent variables is to include a second, converging measure. Classically, a converging measure is defined as an alternative way of measuring the same construct to eliminate alternative hypotheses (e.g., Garner, Hake, & Ericksen, 1956). Converging measures can be useful for making inferences about infant behavior when used across different paradigms, and can be especially powerful when implemented within the same paradigm.

In our own work, we ran additional studies with both looking time and other behavioral measures to examine whether infants and children indeed perceive snakes, for example, as threatening. In two studies, instead of measuring rapid attention to snakes versus others animals, we measured total looking to and reaching for each animal. We reasoned that if infants perceive snakes as threatening, infants might avoid looking at snakes, and perhaps even avoid proximity to and contact with them. By using multiple measures within a single paradigm, we found that infants looked equally long at snakes versus other animals (DeLoache & LoBue, 2008; LoBue & DeLoache, 2010), and they were equally likely to reach and grasp for the animals, literally attempting to pick up live, slithering snakes from a screen (DeLoache & LoBue, 2008). We also explored infants' behavior towards four live animals—a snake, a spider, a hamster, and a fish—in an additional free play paradigm. We found that infants spent more time interacting with all four animals than with four attractive toys, and indeed, showed an avid *interest* in the snake and the spider that equaled their interest in the hamster and fish (LoBue, Bloom, Pickard, Sherman, Axford, & DeLoache, 2013). When taken together, these studies that examine several outcome measures (i.e., multiple infant behaviors), both within the same and across different paradigms

suggest that, for infants, animals like snakes might be attention grabbing, but not necessarily because they carry a threatening valence.

The advantage of combining infant looking behavior with other measures to address problems of inference extends well beyond our own work. Indeed, looking behavior (or "looking time"), including duration of infants' looks and latency to look, is one of the most common measures used in the infant literature. In fact, in the past three years (2016-2018), nearly half (47%) of the empirical articles published in *Infancy* were based on looking behavior. This measure is a perfect outcome for infancy research—it is easy to collect and does not involve language. However, while looking time paradigms were initially developed to investigate simple questions about sensory and perceptual development in infants, many researchers now use infants' looking time responses to investigate questions of higher level infant cognition, making inferences about infants' causal, physical, and numerical reasoning (Haith, 1998), or in our case, the affective impact of a stimulus (DeLoache & LoBue, 2009). In fact, based on looking time measures, researchers have concluded that 3½-month-olds have object permanence (Baillargeon, 1987), 10-month-olds evaluate social behaviors (Hamlin, Wynn, & Bloom, 2007) and expect resources to be distributed fairly and equally (Meristo, Strid, & Surian, 2016), and 15-month-olds make inferences about the mental states of others (Onishi & Baillargeon, 2005).

Relying only on looking behavior—or any one, single measure—to make inferences about cognitive and emotional processes has drawbacks. First, the same data can be interpreted in different ways. For example, longer and faster looking to some displays over others can be interpreted as indicative of high level processing directing infant behavior, whereas a different researcher with a different theoretical position might generate an explanation based on low-level processes. In a now classic study, Wynn (1992) presented infants with an item that was

subsequently hidden behind a screen, followed by a researcher who reached behind the screen to place a second item alongside the first. Infants looked longer when the screen was removed to reveal only one item (the impossible condition) when compared to two items (the possible condition). The original interpretation of this finding was that infants are innately endowed with the ability to do simple arithmetic. However, an equally plausible but lower-level explanation is that infants can track the location of a small number of objects (e.g., Uller et al., 1999; vanMarle, 2013). With only one looking time measure, these two possibilities cannot be empirically dissociated.

For the many infancy researchers that also rely on attention or looking time measures, one additional measure, especially used within the same paradigm, can likewise help with making inferences about the meaning of infants' looks. One way several researchers do this is by using an eye-tracker to supplement global looking time responses. Indeed, overall looking time to a stimulus does not allow us to differentiate between active attention and blank stares, and does not always give us enough information to make inferences about processes, like surprise or expectation, which are very common in the infancy literature (Aslin, 2007). However, by using an eye-tracker, we can address this problem by pairing duration of looking with additional measures, such as the number and sequence of looks, the duration of fixations, the number of times infants look away from a stimulus, and anticipatory looks.

Bremner and colleagues (2017), for example, used converging measures from an eye-tracker to differentiate between possible interpretations of Wynn (1992)'s research, described above. They reasoned that if babies' increased looking to the unexpected test display was driven by recognition of an incorrect numerical outcome, infants' looks should be directed at both the remaining object and at the location of the missing object, as together they constitute the

incorrect number. In contrast, if infants were simply tracking the location of the missing object, increased looking time should be accounted for by looks only to the location of the missing object. Using this converging measure, the authors replicated Wynn's original findings, and in addition, found that infants looked mostly to the location of the missing object, suggesting that infants' looking in this paradigm more likely reflects their ability to track the location of a small number of objects instead of any numerical competency. This example illustrates how the addition of one or more converging measures within standard looking time paradigms can help differentiate between various levels of interpretation when an infant looks longer at one display over another.

Importantly, although differences in looking time to different displays can reflect different cognitive processes, so can *similar* patterns of looking (Aslin, 2007). For example, while in some studies, infants' longer looking reflects a novelty preference, in others, the very same measure can reflect a familiarity preference. Although infants generally tend to show a pre-experimental preference for familiar stimuli, and post-familiarization/habituation preference for novel stimuli, this preference can vary based on the duration of the familiarization phase (Houston-Price & Nakaib, 2004). Likewise, in our own research, we found that infants, children, and adults quickly detect a number of stimuli—including snakes, spiders, angry faces, guns, knives, syringes, hoses, wires, and other simple shapes. Although responses to these stimuli appear to be similar, it is possible that they reflect different underlying processes. For example, several studies now suggest that rapid detection of snakes, at least early in development, is dependent on their low-level perceptual features, like their curvilinear shape (e.g., LoBue & DeLoache, 2011). In contrast, rapid detection of syringes is more likely the result of learning that syringes are often accompanied by an unpleasant prick (e.g., LoBue, 2010b). However, when

researchers rely on only a single outcome measure (i.e., rapid detection of the target stimulus), the processes underlying humans' responses to snakes versus syringes remains unclear.

In this way, using converging measures can help differentiate between the various processes that might be reflected by the same behavior. For example, although longer looking in violation of expectation (VoE) paradigms is often assumed to reflect surprise, researchers have argued that it could also reflect a familiarity preference for one of the displays (e.g., Cohen & Marks, 2002). Thus, when using VoE, adding a converging measure can help researchers make an inference about the underlying process driving longer looking to an unexpected event. Although emotional facial expressions are the obvious choice for a converging measure of surprise, several emotion researchers have pointed out that infants rarely show a stereotypical surprise face when their expectations are violated (Camras, et al., 2002; Scherer, Zentner, & Stern, 2004). There is evidence that infants do, however, demonstrate greater negative activity according to EEG/ERP measures (Berger, Tzur, & Postner, 2006), greater pupil dilation (Gredebäck & Melinder, 2010), and increased social referencing when expectations are violated (Dunn & Bremner, 2017). For example, Walden, Kim, McCoy, and Karass (2007) replicated Wynn's (1992) original results with the addition of a social referencing measure. Importantly, they demonstrated that infants not only look longer at the unexpected event, but they also initiate more looks towards their caregivers' faces during this event, suggesting that the event was indeed unexpected or ambiguous, and did not reflect a familiarity preference for one of the displays.

Similarly, anticipatory looks have also been useful in looking time paradigms that aim to measure infants' expectations. For example, researchers have reported that when watching an object being passed from one person to another, 12-month-old infants look longer at an upright

versus inverted display of a receiving hand. Alone, these data are ambiguous, but paired with the additional finding that infants also demonstrated more anticipatory looks to the receiving hand when it formed a "give-me" gesture, the data provide support for the conclusion that infants *expected* the object to be passed to the person motioning for it (Elsner, Bakker, Rohlfing, & Gredebäck, 2014). A similar study from the same lab reported that when observing one adult feeding another, 12-month-old infants show greater looking time to displays where the feeder brings the spoon to the other adult's hand versus their mouth, suggesting that infants already have expectations about feeding dialogues. The researchers again measured infants' anticipatory looks as a converging measure, and found that infants showed anticipatory fixations to the adult's mouth *before* the food reached it, providing stronger evidence to support their inference (Gredebäck & Melinder, 2010). Without the use of additional eye-tracking measures, such strong inferences could not have been drawn from these data.

Importantly, eye-tracking is not the only way to obtain converging measures for infant looking time data. Indeed, by 6 months of age, infants not only look longer at stimuli they prefer or differentiate, but they can also show their preferences by reaching. In one example, Hamlin, Wynn, and Bloom (2007) examined 6- to 10-month-old infants' preferences for prosocial versus antisocial behavior using a looking paradigm, where infants were familiarized to a live display of a character who tried to scale a hill, and was then either helped up the hill by a second character (i.e., the helper), or pushed down by a third character (i.e., the hinderer). In the test phase, the climber either approached the helper or the hinderer. Ten-month-olds looked longer when the climber approached the hinderer, suggesting that this event violated their expectations; the 6-month-olds did not. In addition to the looking time measure, the researchers also offered the helper and the hinderer characters to the infants and encouraged them to reach. They found that

in addition to looking longer when the climber approached the hinderer, 10-month-olds were more likely to reach for the helper, providing additional support that infants not only distinguished between the helper and the hinderer, but also *preferred* the prosocial character.

While reaching behavior converged with looking time responses in this study, not all researchers are so lucky. There are countless examples in the infancy literature where reaching and looking time measures do not converge, leaving researchers with confusing and seemingly contradictory results. In these cases, additional converging measures can be used to *disambiguate* the inconsistent results. For example, since the 1950's, there has been considerable debate about whether infants under 6 months of age have object permanence. Indeed, while classic Piagetian (1954) tasks demonstrate a failure to search for hidden objects in early infancy, others have used looking time measures to show that babies reason about hidden objects as young as 4 months of age (e.g., Baillargeon, 1987). Researchers have used converging measures using EEG to provide some insight into these discrepant findings. For example, Kaufman, Csibra, and Johnson (2003) used EEG to demonstrate that while infants looked longer when an object was expected to appear behind an occluder, but failed to appear, they also showed increased gamma-band activity, which in adult studies, has been associated with keeping an object in mind, providing converging evidence that infants can represent an absent object by 6 months of age.

Additional Measures—Mechanisms and Individual Differences

To summarize our argument thus far, converging measures can be useful in infancy research when making inferences about infants' behaviors, especially when responses are ambiguous, or can generate multiple levels of explanation. In our own research, the common assumption that rapid threat detection is driven by stimulus valence requires converging

measures that allow for inferences about the meaning of looks to a target stimulus. However, there are additional ambiguities in this line of work that cannot by resolved simply by convergence. For example, the question of whether rapid detection of threat is normative and evolutionarily predisposed, or whether it is learned over the course of development, also requires multiple measures (possibly across different experimental paradigms and multiple time points), but these additional measures should be designed to provide information about *mechanisms of change*. This can be done by measuring multiple responses that are theorized to be part of the same construct (as with convergence), or by measuring potential mediators and moderators of a desired effect. This is typically done within a single study.

As with questions of inference, eye-tracking is one way to move beyond single dependent variables, like latency to detect a target, and measure a number of factors related to rapid attention within a single trial. Using the classic adult button-press procedure with the addition of an eye-tracker, we examined the mechanisms that drive rapid threat detection by measuring multiple components of attention bias, including rapid first fixations to threatening versus non-threatening stimuli, latency to disengage from threatening versus non-threatening distractors, and latency to make a behavioral response once a target stimulus was detected.

First, we replicated previous findings, demonstrating that adults press a button more quickly to indicate the presence of discrepant threatening targets (snakes and spiders) versus non-threatening targets (flowers and mushrooms). Second, we found that the advantage for threat was two-fold: participants were faster to first fixate threatening versus non-threatening targets, *and* they were faster to indicate that they had detected a discrepant image by pressing a button after first fixating the target. These results—using multiple measures of attention bias—demonstrated that rapid detection of threat is driven both by an advantage in perception, or

bottom up processing, and an advantage in behavioral responding, or top down processing (LoBue, Matthews, Harvey, & Stark, 2014), altogether suggesting that multiple factors might drive biased attention.

Furthermore, Field (2006a) investigated whether an attention bias for neutral stimuli could be elicited after hearing emotionally valenced information using multiple outcome measures that mapped onto Lang's (1968) three response systems, namely, measures of subjective feelings, behavior, and physiological responses. He presented 7- and 9-year-old children with three novel animals, and provided positive information about one, threat information about the second, and no information about the third. He then presented the children with a dot-probe task in which pairs of the animal photos were presented very briefly, followed by the appearance of a neutral probe in place of one of the two photos. Children were asked to indicate where the probe appeared (on the left or right) as quickly as possible. Their fear beliefs about each of the animals were assessed before and after hearing the background information for each animal.

Responses to the probe were faster when they replaced a photo of the animal previously paired with threat versus positive information, consistent with an attention bias for the threat-information animal. Importantly, the change in children's fear beliefs about the novel animal after receiving threat information mediated the magnitude of the attention bias, suggesting that threat information increased fear beliefs, which in turn, induced an attention bias. The effects in this study were relatively weak. However, Field and colleagues ran several follow-ups using a similar procedure with added measures to account for some of the unexplained variance. Across several studies, they replicated their initial findings, showing that pairing a neutral animal with threat information induces a heightened attention bias for that animal. Further, he found that a

behaviorally inhibited temperament increased the likelihood of having an attention bias after receiving threat information, and was related to behavioral avoidance of a display containing a photo of the negative animal (Field, 2006b; Reynolds, Askew, & Field, 2018).

This example not only demonstrates how multiple measures can help uncover the mechanisms driving rapid threat detection—in this case, that increasing fear beliefs is the mechanism by which an attention bias is induced—but it also suggests that individual differences, or potential moderators, such as a behaviorally inhibited temperament, might help elucidate the mechanisms that underlie the development of rapid threat detection even further (also see Pérez-Edger et al., this issue). In fact, additional studies that add temperament measures to attention bias tasks suggest that some individuals might be more prone to attend to threat than others. For example, countless studies have linked attentional biases for social threats to clinical anxiety (for reviews see Bar-Haim, Lamy, Pergamin, Bakermans-Kranenburg, van IJzendoorn, 2007; Fu & Pérez-Edgar, 2019). In our own labs, we measured whether children who are temperamentally shy, and are thus *at-risk* for the development of social anxiety, may have a heightened attention bias for social threats. We found that while all 5-year-olds detect snakes more quickly than frogs and angry faces more quickly than happy faces consistent with previous research, children who were temperamentally shy showed a greater attention bias for angry faces, but not for snakes, when compared to non-shy participants (LoBue & Pérez-Edgar, 2014). Thus, by measuring attention biases for two different types of threat along with a measure of temperamental shyness, we found that temperament might play a role in augmenting attentional biases for social threats—but not non-social threats—over the course of development. More broadly, by including potential moderators, we are able to more readily learn something about the mechanisms that drive developmental change.

The inclusion of additional measures like the ones just described can be particularly helpful in uncovering the mechanisms that drive developmental change in infancy research. Soska and Johnson (2008), for example, first set out to examine 4- and 6-month-olds' 3D object completion by using a looking-time habituation paradigm. In the study, infants were habituated to the partial view of a 3-dimensional object, and were then tested with complete and incomplete versions of the same object. They found that while 4-month-olds looked equally long at the two displays, 6-month-olds looked reliably longer at the incomplete test display, suggesting that they could perceive the complete form during habituation. Although this study suggests that there is a transition between the ages of 4 and 6 months in infants' 3D object perception, it does not provide any information about what drives that developmental change, which is typical of infant looking-time work. However, in a follow-up study using the same habituation task with 4- to 7-month-olds, the looking time measure was combined with measures of infants' sitting experience and their manual behavior when exploring several novel toys. The researchers found that looking to the incomplete object in the test phase of the looking task was predicted by both self-sitting and manual exploratory skills, suggesting that the onset of independent sitting and subsequent experience of manually manipulating objects might facilitate 3D object perception (Soska, Adolph, & Johnson, 2010).

Measuring individual differences via parent report is another way for researchers to examine potential mechanisms of developmental change. In our own research described above, we often used parents' reports of infants' emotional responses to novel stimuli, and parents' self-reports of their own levels of anxiety and depression to investigate the factors that might drive the development of attention biases to threat. To cite another example from the infancy literature, Ziv and Sommerville (2017) asked parents to report on the naturalistic sharing behavior of their

9-month-old infants, and found that variability in sharing behavior predicted whether infants look longer at fair versus unfair distributions. Further, in the Gredebäck & Melinder (2010) study described above, in which 12-month-old infants showed anticipatory looks when watching one adult feed another adult, anticipatory looking was predicted by the infant's experience of being fed at home.

Altogether, the work reviewed in this section suggests that by incorporating additional measures, researchers can ask questions about the mechanisms that drive developmental change across infancy. When we began to incorporate multiple outcome measures in our own investigations of attention bias to threat, our once simple story became richer and more complex. We learned that multiple factors—both perceptual and cognitive—can drive attentional biases for threat, attention biases for previously neutral stimuli can develop based on specific experiences, and that there is a potentially complex developmental relation between attention biases to threat, temperament, and behavior. We could not reach these conclusions using only a single attention task. However, by incorporating multiple measures—including measures of attention bias, temperament, fear beliefs, physiology, and behavioral inhibition—a richer and more complex developmental story was able to unfold.

Methodological and Statistical Considerations

The reliance that infancy research has had on single outcome measures and small samples (Oakes, 2017; see also DeBolt, Rhemtulla, & Oakes, this issue) has statistical consequences. Small sample sizes create two major problems in the context of frequentist statistics. First, studies based on small samples are underpowered to detect effects that might be large enough to have theoretical substance. Second, when studies based on small samples yield significant

effects, these effects are more likely to be large, surprising and, therefore, publishable, but they are less likely to replicate because they will overestimate the population effect size—the so called 'winner's curse' (Young, Ioannidis, & Al-Ubaydli, 2008). Consequently, theoretical models will be built upon the shaky foundations of prized overestimated effects and discarded, but substantive, underestimated effects.

Measuring multiple outcomes may help in this regard. First, multiple outcomes can be used as indicators of latent psychological constructs and measurement theory tells us that, other things being equal, the reliability of composite measures becomes more reliable as the number of variables contributing to the composite increases (Cronbach, 1951). Second, if a common statistical model is fit to outcomes separately (rather than using them as indicators of a latent variable), consistency in the corresponding effect sizes from these analyses increases the likelihood that these effects reflect a substantive psychological phenomenon rather than being isolated, spurious results. Conversely, inconsistency in effect sizes across models fit to different outcome measures places an isolated significant effect from a particular outcome in an appropriate context.

However, measuring multiple outcome variables comes with risks and pragmatic considerations. One obvious risk is that *p*-hacking becomes more likely for two reasons. First, the potential set of analytic decisions that a researcher can make—the so-called "garden of forking paths" (Gelman & Loken, 2014; Simmons, Nelson, & Simonsohn, 2011)—escalates as a function of the increased complexity of the statistical models required to analyze multiple outcomes. The problem here is that the theories developed from the statistical models fit to the data may be different than those that would have been developed had the researcher made different analytic decisions. A particular risk is that, given the tendency to publish significant

results, having more data will lure researchers into making decisions that lead to significant results and not declare the lack of significance from a different set of analytic decisions (for example, adjusting for a confounding variable because doing so pushes the substantive effect of interest below the threshold for significance). Also, if a researcher has multiple outcomes to choose from, researchers can report identical statistical model's fit to different outcomes and may be tempted to cherry-pick the measures that yield consistent (significant) results and not report collecting measures that yield non-significant effects.

These temptations are not specific to infancy research—they are the research practices at the heart of the replication crisis in psychology (Open Science Collaboration, 2015). However, adopting a 'multiple outcomes' approach can sometimes increase these risks. Researchers can protect themselves from temptation by adopting scientific practices such as pre-registering studies and analysis plans, or submitting research as a registered report, which are effective in mitigating *p*-hacking, HARK-ing (i.e., hypothesizing after the results are known), and publication bias (Wicherts et al., 2016).

Another issue is that researchers need to give careful thought to what outcome measures they use and how they model them statistically. Broadly there are two strategies here: use multiple outcome measures as indicator variables of a latent variable (the structural equation modeling, SEM, approach), or fit the same statistical model to each outcome independently and look for convergence in the evidence. Other things being equal, the SEM approach has several advantages. As already mentioned, measurement theory tells us that using multiple measures should increase the reliability of the latent/composite variable. The SEM approach also allows flexibility in how the relations between outcome variables are conceptualized, and enables comparison of competing conceptualizations of the phenomenon being studied.

To give a concrete example, it is not a given that data from theoretically related measures should converge. As already noted, cognitive, behavioral, and physiological measures of fear are usually not synchronous. However, this fact is unsurprising. Attention-to-threat tasks are likely to tap low-level automatic processing whereas behavior tasks tap higher-level intentional processing. Attention might be drawn to a snake-like object, but once you have realized that it is not a snake you do not avoid it. Figure 1 shows two conceptualizations of emotion as a latent variable driving observable outcome measures. The right shows a single level model in which emotion is seen as a latent variable driving measured responses across three response systems (Lang's aforementioned cognition, behavior, and physiology). This conceptualization models synchrony between outcome measures because the latent psychological construct (emotion) has a direct link to the outcome measure. Statistically, this would be a useful framework for modeling outcomes expected to be synchronous based on theory. The model on the right (from Zinbarg, 1998) conceptualizes asynchrony between outcomes by assuming that different response systems are themselves lower-order latent variables. The psychological construct (emotion) now has an indirect link to outcome measures: it drives the response system, which in turn drives responses to relevant outcome measures. In this model, measures *within* a response system would be theoretically synchronous, but they would not need to be *across* response systems (e.g., M1 and M2 should correlate strongly but M1 and M3 need not).

In general, then, researchers need to think carefully about which outcome measures they expect to converge and which they do not and model the data accordingly. The SEM framework has this flexibility. Fit measures can be used to determine the best fitting model of the construct of interest. Comparing models in this way can be used both to inform theoretical models, and to test for measurement invariance over development. Latent variable models of psychological

constructs can also be modeled over time using latent growth models in which the intercept and rate of change of the construct over time are similarly modeled as latent variables (for an introduction see Newsom, 2015). This modeling framework is flexible enough to allow researchers to adjust for other variables theoretically related to the construct of interest whether they are time variant or invariant.

The main drawback of the SEM approach is that it requires large samples to get stable parameter estimates and models may not converge in small samples. In addition, as we have highlighted already, infancy research often involves small samples. One solution is for labs studying similar phenomena to join forces, agree upon the most pertinent theoretical questions, and pool resources to study them. Projects such as the Many Labs projects led by the Center for Open Science (https://cos.io/), the ManyBabies initiative (https://manybabies.github.io/), the Psychological Science Accelerator (https://psysciacc.org/), and The Collaborative Replications and Education Project (https://osf.io/wfc6u/wiki/home/) have demonstrated that it is possible to generate vast datasets through mass collaboration. Moving towards a more collective and less-individualistic model of infancy research will, in the long run, benefit the discipline.

Nevertheless, pioneering work on small samples is also vital. Latent variable models are unlikely to be useful in small samples. Instead, researchers will be reliant on fitting the same statistical models multiple times to different outcome measures and evaluating the consistency of evidence across those measures. How should they approach this task? Null hypothesis significance testing is probably the least suited tool. First, as mentioned above, significance tests will be underpowered in small samples making consistency across outcome measures unlikely. The application of an arbitrary threshold for significance, such as .05, in these models will not

help. Second, using multiple significance tests will inflate the familywise Type I error rate, and correcting this rate for the number tests will further reduce the statistical power of each test.

In this case, one could adopt an estimation approach. For each outcome measure, a model is fit and an effect size estimated. In most designs, the model's fit will be some variant of a general or generalized linear model (for a basic introduction to the general linear model see Field, 2016), and an 'estimation approach' might be as simple as extracting the model parameters (which are unstandardized effect sizes). Measures of uncertainty around these effects should also be used to place the effects in context. For example, Bayesian HPD intervals can be used to determine the range of plausible population values of the effect and frequentist 95% confidence intervals can be similarly used under the limiting assumption that the interval is one of the 95% that captures the population value. Researchers should evaluate whether these effects are substantively important, and also whether they are consistent across measures. Comparing the overlap in HPD intervals may be particularly useful in this respect. The choice of Bayesian or frequentist estimation is up to the researcher, but Bayesian estimation allows researchers to build in prior beliefs about the likely size of effects based on past research and theoretical expectations, which may be desirable (for an accessible introduction to Bayesian estimation see McElreath, 2016).

In situations where pure estimation is undesirable and hypothesis testing is essential to address the substantive research question, Bayes Factors can be used to quantify how the researcher's beliefs about an alternative hypothesis should shift relative to the null given the data (see Dienes, 2014, for an introduction). Unlike null hypothesis significance testing, Bayes factors do not need to be adjusted when they are used multiple times. Converging evidence across measures would be indicated by Bayes factors of a similar magnitude in the same direction.

Bayes factors also enable researchers to draw conclusions about the plausibility of the null, which significance tests cannot, and this may be important for identifying hypotheses as implausible.

Conclusions: Using Multiple Measures in Future Infant Research

Here, we propose that using multiple outcome measures in our study designs can help us make sense of infant data. It is important to note that the common approach of testing theories using a series of single-measure experiments can still help us understand infant behavior, and has been used to generate important developmental data for decades. Further, there are cases where converging measures are not necessary, such as when researchers seek to answer questions about behavior that can be measured directly, without the need for inference. However, we hope that the work we described here from our own labs and from the labs of others demonstrates that relying only on a single dependent measure, even across a wide variety of studies and age groups, can limit the kinds of questions we ask and the subsequent conclusions we can draw from the data. Collecting converging evidence from multiple outcome measures in the infancy literature also has benefits for allowing researchers to make more accurate inferences when infant behaviors can be attributed to more than one process, are susceptible to multiple levels of explanation, reflect complex constructs, or when primary measures fail to converge. Finally, multiple measures can also be useful in building more robust models of the mechanisms that guide developmental change, and ultimately help us to more accurately characterize the richness of infant development.

As mentioned above, nearly half (47%) of the empirical articles published in *Infancy* in the past three years were based on looking behaviors; thus we used several examples from infant

looking time data to demonstrate that additional eye-tracking and behavioral measures can be used to provide measures of convergence and to answer questions about mechanisms of change. But besides eye trackers, additional measures using ERPs or physiology can also provide important stand-alone or converging measures of infant attention. Indeed, a small percentage of the research we surveyed from *Infancy* (~7%) reported using physiological and neural (e.g., cortisol, DNA, ERP) measures as their main dependent variables, or as converging measures for studies focused on the development of infant attention. For example, Reynolds and Richards (2017) found that physiological measures like heart rate and ERP responses (the Nc component) can act as converging measures of attentional control; in other words, like looking time, by 6 months of age, heart rate and Nc activity change as infants are shown repeated trials of the same stimulus. Furthermore, by using all three measures, these researchers are modeling how neural, physiological, and behavioral systems interact to impact infant attention (e.g., Reynolds & Richards, 2017; Xie, Mallin, & Richards, 2018).

Although eye-tracking, physiology, and EEG/ERP are all viable candidates for additional measures in infancy research, not everyone has the funding to purchase the expensive equipment or the methodological expertise required to implement designs using these methodologies. However, it is important to note that multiple outcome measures can be extracted solely from video recordings of infant behavior. Again, other behavioral measures, such as reaches, vocalizations, or head-turns for example, were almost as common as looking time measures in our survey of recently published *Infancy* research (46%). This is not surprising, as behavior captured on video is inexpensive, and can be used—and reused—to collect a large number of dependent variables. For example, Karasik, Tamis-LeMonda, and Adolph (2011) videotaped 50 infants in their homes for 1 hour at two time points—at 11 months and 13 months. The study was

initially aimed at examining infants' object use during the transition from crawling to walking, so they measured the amount of time infants handled objects, where the object was located, whether infants had to crawl or walk to obtain the object, instances of carrying objects, and when the infants shared objects with their mothers. Using a combination of these variables, the researchers discovered that the onset of walking was associated with new forms of object behavior, including traveling long distances to obtain objects, carrying objects for the first time, and approaching mothers to share objects in joint attention. Further, infants who engaged in these behaviors as crawlers were more likely to walk at 13 months, suggesting that the act of carrying objects might promote locomotor development.

Importantly, the researchers later coded for additional behaviors, this time scoring how often the infants fell while carrying objects. They reported that infants were less likely to fall when carrying than when not carrying objects, providing additional support that carrying objects might help locomotor development (Karasik, Adolph, Tamis-LeMonda, & Zuckerman, 2012). Finally, in one last pass, they scored mothers' verbal responses to their infants' actions, and found that in response to walkers' bids for sharing objects, mothers responded with more action directives (Karasik, Tamis-LeMonda, & Adolph, 2014). When combined, these data provide evidence for a complex developmental cascade, in which infants' carrying behavior affects their locomotor development, which in turn affects their ability to share objects with their mothers, which then shapes mothers' behavior. Not only did these researchers code for a large number of behavioral responses, but they also accomplished it by simply recording infant's behavior on video, without the use of eye-trackers, brain caps, heartbeats, or even a lab space.

In conclusion, this review suggests that using multiple outcome measures in infant research can help provide converging evidence for our inferences, and expand our theories about

the mechanisms of developmental change. Eye-tracking, ERP's, and physiological measures can be useful in providing converging evidence from which to make inferences about infant data. However, in addition to or in the absence of these technologies, video recording infant behavior or asking parents to provide additional information can also get us closer to making reliable and accurate conclusions about infant development. Finally, it is important to note that while collecting and analyzing infant data is time consuming and labor intensive for researchers, bringing an infant to a developmental laboratory is time consuming and labor intensive for *parents*. Make it worth their while and at the same time, produce richer data—instead of collecting only a single measure, in a single age group, at a single time point, use multiple measures, and in the process, create the means by which to make better inferences and learn more about the mechanisms that underlie developmental change.

References

Aslin, R. N. (2007). What's in a look?. *Developmental Science*, *10*(1), 48-53.

Baillargeon, R. (1987). Object permanence in 3½-and 4½-month-old infants. *Developmental Psychology*, *23*(5), 655-664.

Bar-Haim, Y., Lamy D., Pergamin, L., Bakermans-Kranenburg, M.J., van IJzendoorn, M.H. (2007). Threat-related attentional bias in anxious and nonanxious individuals: a meta-analytic study. *Psychological Bulletin. 133*(1),1–24

Berger, A., Tzur, G., & Posner, M. I. (2006). Infant brains detect arithmetic errors. *Proceedings of the National Academy of Sciences*, *103*(33), 12649-12653.

Blanchette, I. (2006). Snakes, spiders, guns, and syringes: How specific are evolutionary constraints on the detection of threatening stimuli? *The Quarterly Journal of Experimental Psychology*, *59*(8), 1484-1504.

Bolles, R. C. (1970). Species-specific defense reactions and avoidance learning. *Psychological Review*, *77*(1), 32-48.

Boyer & Bergstrom (2011). Threat-detection in child development: An evolutionary perspective. *Neuroscience and Biobehavioral Reviews, 35*, 1034–1041.

Bremner, J. G., Slater, A. M., Hayes, R. A., Mason, U. C., Murphy, C., Spring, J., ... & Johnson, S. P. (2017). Young infants' visual fixation patterns in addition and subtraction tasks support an object tracking account. *Journal of Experimental Child Psychology*, *162*, 199-208.

Brosch, T., & Sharma, D. (2005). The role of fear-relevant stimuli in visual search: A comparison of phylogenetic and ontogenetic stimuli. *Emotion*, *5*(3), 360-364.

Buss, K. A. (2011). Which fearful toddlers should we worry about? Context, fear regulation, and anxiety risk. *Developmental Psychobiology, 47*(3), 804-819.

Camras, L. A., Meng, Z., Ujiie, T., Dharamsi, S., Miyake, K., & Oster, H.et al. (2002). Observing emotion in infants: Facial expression, body behavior, and rater judgments of responses to an expectancy-violation event. *Emotion, 2*(2), 179–192.

Cohen, L. B., & Marks, K. S. (2002). How infants process addition and subtraction events. *Developmental Science, 5,* 186–212.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

DeBolt, M. C., Rhemtulla, M., & Oakes, L. M. (2019, under review). Robust data and power in infant looking time research: Number of infants and number of trials. *Infancy*.

DeLoache, J. S., & LoBue, V. (2009). The narrow fellow in the grass: Human infants associate snakes and fear. *Developmental Science*, *12*(1), 201-207.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00781

Dunn, K., & Bremner, J. G. (2017). Investigating looking and social looking measures as an index of infant violation of expectation. *Developmental Science*, *20*(6), e12452.

Elsner, C., Bakker, M., Rohlfing, K., & Gredebäck, G. (2014). Infants' online perception of give-and-take interactions. *Journal of Experimental Child Psychology*, *126*, 280-294.

Field, A. P. (2006a). Watch out for the beast: Fear information and attentional bias in children. *Journal of Clinical Child and Adolescent Psychology, 35*(3), 431-439.

Field, A. P. (2006b). The behavioral inhibition system and the verbal information pathway to children's fears. *Journal of abnormal psychology, 115*(4), 742-752.

Field, A. P. (2016). *An adventure in statistics: the reality enigma*. London: Sage.

Fu, X. & Pérez-Edgar, K. (2019). Threat-related attention bias in socioemotional development: A critical review and methodological considerations. *Developmental Review, 51,* 31-57.

Garner, W. R., Hake, W. K., & Ericksen, C. W. (1956). Operationism and the concept of perception. *Psychological Review, 63*(3), 149-159.

Gelman, A., & Loken, E. (2014). The Statistical Crisis in Science. *American Scientist, 102*(6), 460. https://doi.org/10.1511/2014.111.460

Goldsmith, H. H., & Rothbart, M. K. (1999). *The Laboratory Temperament Assessment Battery (Locomotor Version, Edition 3.1)*. Madison, WI: University of Wisconsin-Madison.

Gredebäck, G., & Melinder, A. (2010). Infants' understanding of everyday social interactions: A dual process account. *Cognition, 114,* 197-206.

Haith, M. M. (1998). Who put the cog in infant cognition? Is rich interpretation too costly? *Infant Behavior and Development, 21*(2), 167-179.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature, 450*(22), 557-559.

Hayakawa, S., Kawai, N., & Masataka, N. (2011). The influence of color on snake detection in visual search in human children. *Scientific Reports, 1,* 1-4.

Houston-Price, C., & Nakai, S. (2004). Distinguishing novelty and familiarity effects in infant preference procedures. *Infant and Child Development: An International Journal of Research and Practice, 13,* 341-348.

James, W. (1890). *The Principles of Psychology (Vol. 1).* Read Books Ltd.

Kagan, J., Snidman, N., McManis, M., Woodward, S., & Hardway, C. (2002). One measure, one meaning: Multiple measures, clearer meaning. *Development and Psychopathology, 14*(3), 463-475.

Karasik, L. B., Adolph, K. E., Tamis-LeMonda, C. S., & Zuckerman, A. L. (2012). Carry on: Spontaneous object carrying in 13-month-old crawling and walking infants. *Developmental Psychology*, *48*(2), 389-397.

Karasik, L. B., Tamis-LeMonda, C. S., & Adolph, K. E. (2011). Transition from crawling to walking and infants' actions with objects and people. *Child Development, 82*(4), 1199-1209.

Karasik, L. B., Tamis-LeMonda, C. S., & Adolph, K. E. (2014). Crawling and walking infants elicit different verbal responses from mothers. *Developmental Science*, *17*(3), 388-395.

Kaufman, J., Csibra, G., & Johnson, M. H. (2003). Representing occluded objects in the human infant brain. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*(suppl_2), S140-S143.

Lang, P. J. (1968). Fear reduction and fear behavior: Problems in treating a construct. In J. M. Schlien (Ed.), *Research in Psychotherapy* (Vol. 3, pp. 90-103). Washington, D. C.: American Psychological Association.

Larson, C. L., Aronoff, J., & Stearns, J. J. (2007). The shape of threat: Simple geometric forms evoke rapid and sustained capture of attention. *Emotion, 7*(3), 526-534.

LoBue, V. (2009). More than just a face in the crowd: Detection of emotional facial expressions in young children and adults. *Developmental Science, 12*(2), 305-313.

LoBue, V. (2010a). And along came a spider: Superior detection of spiders in children and adults. *Journal of Experimental Child Psychology, 107*(1), 59-66.

LoBue, V. (2010b). What's so scary about needles and knives? Examining the role of experience in threat detection. *Cognition and Emotion, 24*(1), 80-87.

LoBue, V. (2016). When is a face no longer a face? A problematic dichotomy in visual detection research. *Emotion Review, 8*(3), 250-257.

LoBue, V. & Adolph, K. E. (2019). Fear in infancy: Lessons from snakes, spiders, heights, and strangers. *Developmental Psychology, 55,* 1889-1907.

LoBue, V., Bloom Pickard, M., Sherman, K., Axford, C., & DeLoache, J. S. (2013). Young children's interest in live animals. *British Journal of Developmental Psychology, 31*(1), 57-69.

LoBue, V. & DeLoache, J. S. (2008). Detecting the snake in the grass: Attention to fear relevant stimuli by adults and young children. *Psychological Science, 19*(3), 284-289.

LoBue, V., & DeLoache, J. S. (2010). Superior detection of threat-relevant stimuli in infancy. *Developmental Science, 13*(1), 221-228.

LoBue, V., & DeLoache, J. S. (2011). What so special about slithering serpents? Children and adults rapidly detect snakes based on their simple features. *Visual Cognition, 19*(1), 129-143.

LoBue, V., & Larson, C. L. (2010). What makes angry faces look so…angry? Examining visual attention to the shape of threat in children and adults. *Visual Cognition, 18*(8), 1165-1178.

LoBue, V., Matthews, K., Harvey, T., & Stark, S. L. (2014). What accounts for the rapid detection of threat? Evidence for an advantage in perceptual and behavioral responding from eye movements. *Emotion, 14*(4), 816-823.

LoBue, V., & Pérez-Edgar, K. (2014). Sensitivity to social and non-social threats in temperamentally shy children at-risk for anxiety. *Developmental Science, 17*(2), 239-247.

LoBue, V., & Rakison, D. (2013). What we fear most: A developmental advantage for threat-relevant stimuli. *Developmental Review, 33*(4), 285-303.

Masataka, N, Hayakawa, S., & Kawai, N. (2010). Human young children as well as adults demonstrate 'superior' rapid snake detection when typical striking posture is displayed by the snake. *PLoS ONE, 5*(11), pp. e15122.

McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press.

Meristo, M., Strid, K., & Surian, L. (2016). Preverbal infants' ability to encode the outcome of distributive actions. *Infancy, 21*(3), 353-372.

Morris, A. S., Robinson, L. R., & Eisenberg, N. (2006). Applying a multimethod perspective to the study of developmental psychology. In M. Eid & E. Diener (Eds.), *Handbook of Multimethod Measurement in Psychology* (pp. 371-384). Washington, DC: American Psychological Association.

Newsom, J. T. (2015). *Longitudinal Structural Equation Modeling: A Comprehensive Introduction*. New York, NY: Routledge.

Oakes, L. M. (2017). Sample size, statistical power, and false conclusions in infant looking-time research. *Infancy*, *22*(4), 436-469.

Öhman, A., Flykt, A., & Esteves, F. (2001). Emotion drives attention: detecting the snake in the grass. *Journal of Experimental Psychology: General*, *130*(3), 466-478.

Öhman, A., Lundqvist, D., & Esteves, F. (2001). The face in the crowd revisited: a threat advantage with schematic stimuli. *Journal of Personality and Social Psychology, 80*(3), 381-396.

Öhman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review, 108*(3), 483-522.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*(5719), 255-258.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716.

Piaget, J. 1954 The construction of reality in the child. New York: Basic Books.

Purkis, H. M., Lester, K. J., & Field, A. P. (2011). But what about the Empress of Racnoss? The allocation of attention to spiders and Doctor Who in a visual search task is predicted by fear and expertise. *Emotion, 11*(6), 1484-1488.

Reynolds, G., Askew, C., & Field, A. P. (2018). Behavioral inhibition and the associative learning of fear. In K. Pérez-Edgar & N. A. Fox (Eds.) *Behavioral Inhibition: Integrating Theory, Research, and Clinical Perspectives*, 237-261. Cham, Switzerland: Springer.

Reynolds, G. D., & Richards, J. E. (2017). Infant visual attention and stimulus repetition effects on object recognition. *Child Development, 90*(4), 1027-1042.

Scherer, K. R., Zentner, M. R., & Stern, D. (2004). Beyond surprise: the puzzle of infants' expressive reactions to expectancy violation. *Emotion, 4*(4), 389-402.

Schubö, A., Gendolla, G. H., Meinecke, C., & Abele, A. E. (2006). Detecting emotional faces and features in a visual search paradigm: Are faces special? *Emotion, 6*(2), 246-256.

Shibasaki, M., & Kawai, N. (2009). Rapid detection of snakes by Japanese Monkeys (*Macaca fuscata*): An evolutionarily predisposed visual system. *Journal of Comparative Psychology, 123*(2), 131-135.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Soska, K. C., Adolph, K. E., & Johnson, S. P. (2010). Systems in development: motor skill acquisition facilitates three-dimensional object completion. *Developmental Psychology, 46*(1), 129-138.

Soska, K. C., Johnson, S. P. (2008). Development of three-dimensional object completion in infancy. *Child Development, 79*(5), 1230–1236.

Tipples, J., Atkinson, A. P., & Young, A. W. (2002). The eyebrow frown: A salient social signal. *Emotion*, *2*(3), 288-296.

Uller, C., Carey, S., Huntley-Fenner, G., & Klatt, L. (1999). What representation might underlie infant numerical knowledge? *Cognitive Development, 14*, 1-36.

vanMarle, K. (2013). Infants use different mechanisms to make small and large number ordinal judgements. *Journal of Experimental Child Psychology, 114*, 102-110.

Walden, T., Kim, G., McCoy, C., & Karrass, J. (2007). Do you believe in magic? Infants' social looking during violations of expectations. *Developmental Science, 10*(5), 654-663.

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. (2016). Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01832

Wynn, K. (1992). Addition and subtraction by human infants. *Nature*, *358*(6389), 749-750.

Young, N. S., Ioannidis, J. P. A., & Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLOS Medicine*, *5*(10), e201. https://doi.org/10.1371/journal.pmed.0050201

Zinbarg, R. E. (1998). Concordance and synchrony in measures of anxiety and panic reconsidered: A hierarchical model of anxiety and panic. *Behavior Therapy*, *29*(2), 301–323.

Ziv, T., & Sommerville, J. A. (2017). Developmental differences in infants' fairness expectations from 6 to 15 months of age. *Child Development*, *88*(6), 1930-1951.

Xie, W., Mallin, B. M., & Richards, J. E. (2018). Development of infant sustained attention and its relation to EEG oscillations: an EEG and cortical source analysis study. *Developmental Science*, *21*(3), e12562.

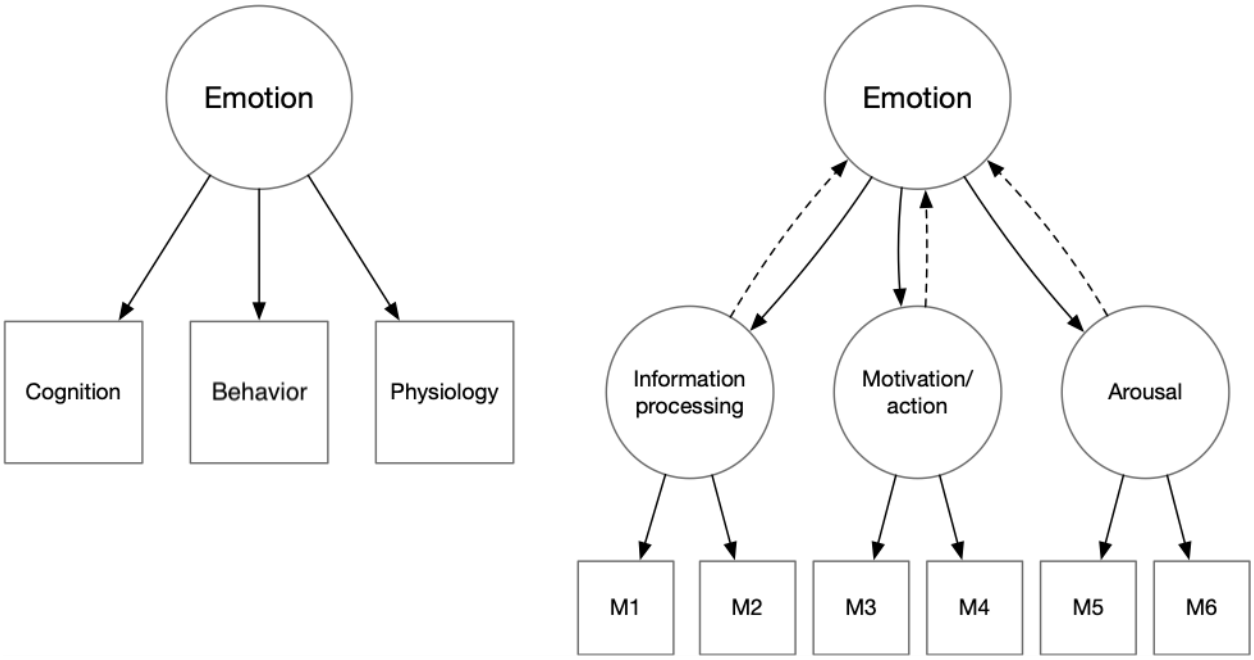

Figure 1. A one-level (left) and two-level (right) conceptualization of the relationship between a psychological construct (in this case emotion) and outcome measures

Acknowledgements

This research was supported by National Institute of Mental Health Grant R01-MH109692 to Koraly Pérez-Edgar, Kristin Buss, and Vanessa LoBue, and by a James McDonnell Foundation Scholar Award for Understanding Human Cognition to Vanessa LoBue. The authors declare no conflicts of interest with regard to the funding sources for this study.