Developmental Psychology
2017, Vol. 53, No. 8, 1405–1417

0012-1649/17/$12.00 http://dx.doi.org/10.1037/dev0000330

Babies and Math: A Meta-Analysis of Infants' Simple Arithmetic Competence

Joan Christodoulou
University of California, Los Angeles

Andrew Lac
University of Colorado, Colorado Springs

David S. Moore
Pitzer College and Claremont Graduate University

Wynn's (1992) seminal research reported that infants looked longer at stimuli representing "incorrect" versus "correct" solutions of basic addition and subtraction problems and concluded that infants have innate arithmetical abilities. Since then, infancy researchers have attempted to replicate this effect, yielding mixed findings. The present meta-analysis aimed to systematically compile and synthesize all of the primary replications and extensions of Wynn (1992) that have been conducted to date. The synthesis included 12 studies consisting of 26 independent samples and 550 unique infants. The summary effect, computed using a random-effects model, was statistically significant, $d = +0.34$, $p < .001$, suggesting that the phenomenon Wynn originally reported is reliable. Five different tests of publication bias yielded mixed results, suggesting that while a moderate level of publication bias is probable, the summary effect would be positive even after accounting for this issue. Out of the 10 metamoderators tested, none were found to be significant, but most of the moderator subgroups were significantly different from a null effect. Although this meta-analysis provides support for Wynn's original findings, further research is warranted to understand the underlying mechanisms responsible for infants' visual preferences for "mathematically incorrect" test stimuli.

Keywords: infant cognition, arithmetical abilities, addition, subtraction

Over the past 30 years, many studies on numerical cognition have investigated the development of basic number understanding in infancy. Classic (e.g., Starkey & Cooper, 1980) and more recent (Brannon, Abbott, & Lutz, 2004; Cordes & Brannon, 2009; Xu & Arriaga, 2007) evidence for numerical processing in the first months of life has been provided largely by measuring infant looking times using habituation and familiarization procedures (for review, see Cantrell & Smith, 2013). Research undertaken with these methodological paradigms has suggested that infants can discriminate and track differences between small number sets (e.g., Brannon et al., 2004; Cordes & Brannon, 2009; Lipton & Spelke, 2004; Starkey & Cooper, 1980; Xu & Arriaga, 2007). Some research has even suggested that infants can transfer numerical information from the auditory to the visual modality (Starkey, Spelke, & Gelman, 1983, 1990; but see also Moore, Benenson, Reznick, Peterson, & Kagan, 1987, for contradictory evidence). In 1992, Wynn extended this line of investigation by conducting the first numerical transformation study asking if infants can perform simple mathematical calculations.

The Original Study and Subsequent Studies

In the original experiments Wynn (1992) conducted, 5-month-old infants were exposed initially to two pretrials (a presentation of one item and then a presentation of two items, or vice versa) to record baseline looking times for one and two items. These presentations served as initial measures to assess if infants looked at displays of different numerosities for different amounts of time. Although Wynn did not identify these as "familiarization trials," these pretrials allowed infants to become at least marginally familiar with the stimuli. Overall, no significant differences were found in the looking times during these pretrials.

Next, each infant was randomly assigned to one of two conditions: addition (1 + 1) or subtraction (2 − 1). In the addition condition, infants initially saw a Mickey Mouse doll on a puppet theater-like stage. A screen was then raised that obstructed their view of the doll, after which a second doll was added to the stage. Although the second doll was placed behind the raised screen, it was added through a side door so the infant could see the addition of the doll (1 + 1) without seeing the "solution," that is, the total number of dolls occupying the stage after the addition. In the subtraction condition, infants initially saw two Mickey Mouse dolls on the stage. Then a screen was raised to obstruct their view of both dolls. This event was followed by a hand that was seen removing one of the dolls (2 − 1) from behind the screen through the side door. The ultimate "solution"

This article was published Online First June 5, 2017.

Joan Christodoulou, Department of Psychiatry & Biobehavioral Sciences, Semel Institute, University of California, Los Angeles; Andrew Lac, Department of Psychology, University of Colorado, Colorado Springs; David S. Moore, Psychology Field Group, Pitzer College and Claremont Graduate University.

Correspondence concerning this article should be addressed to Joan Christodoulou, Department of Psychiatry & Biobehavioral Sciences, Semel Institute, University of California, Los Angeles, 10920 Wilshire Boulevard, Suite 350, Los Angeles, CA 90024. E-mail: jchristodoulou@mednet.ucla.edu

in both conditions was revealed in a test display when the screen was lowered to show either one doll (representing the "incorrect" solution for the addition condition and the "correct" solution for the subtraction condition), or two dolls (representing the "correct" solution for the addition condition and the "incorrect" solution for the subtraction condition). Infants saw this series of events six times, with the test display alternating between one and two dolls across the six trials (see Figure 1).

Wynn (1992) reported that infants looked longer at the "incorrect" (or "unexpected") numerical solution than at the "correct" numerical solution (i.e., infants preferred looking at one doll after the addition events and at two dolls after the subtraction events). Wynn also reported an extension of her original study, in which the 1 + 1 condition was followed by test displays of either two or three Mickey Mouse dolls. Once again, infants preferred looking at the "incorrect" test display (i.e., three dolls rather than two).

Based on the view that infants look longer at unexpected than expected events (Fagan, 1990) and previous evidence of infants' sensitivity to small numerical changes (Starkey & Cooper, 1980), Wynn (1992) concluded that "infants can calculate the solutions of simple arithmetical operations on a small number of items" and that her results indicated "that 5-month-old infants possess true numerical concepts and that humans are innately endowed with arithmetical abilities" (p.749). By referring to these abilities as "innate," Wynn was arguing that humans are born with the ability to manipulate numerical information according to mathematical rules. This interpretation has important implications for the development of quantitative thinking, as it suggests that some mathematical concepts develop independently of postnatal experience.

The impact of Wynn's results on the field of developmental psychology, and researchers' persistent interest in this phenomenon as it relates to advances in understanding the development of numerical cognition, is evident in the number of related studies that followed Wynn's experiments (e.g., Cohen & Marks, 2002; McCrink & Wynn, 2004; Moore & Cocas, 2006) as well as in ongoing discussions about her findings in numerous publications, such as the recent edition of the *Oxford Handbook of Numerical Cognition* (McCrink & Birdsall, 2015), recent introduction to

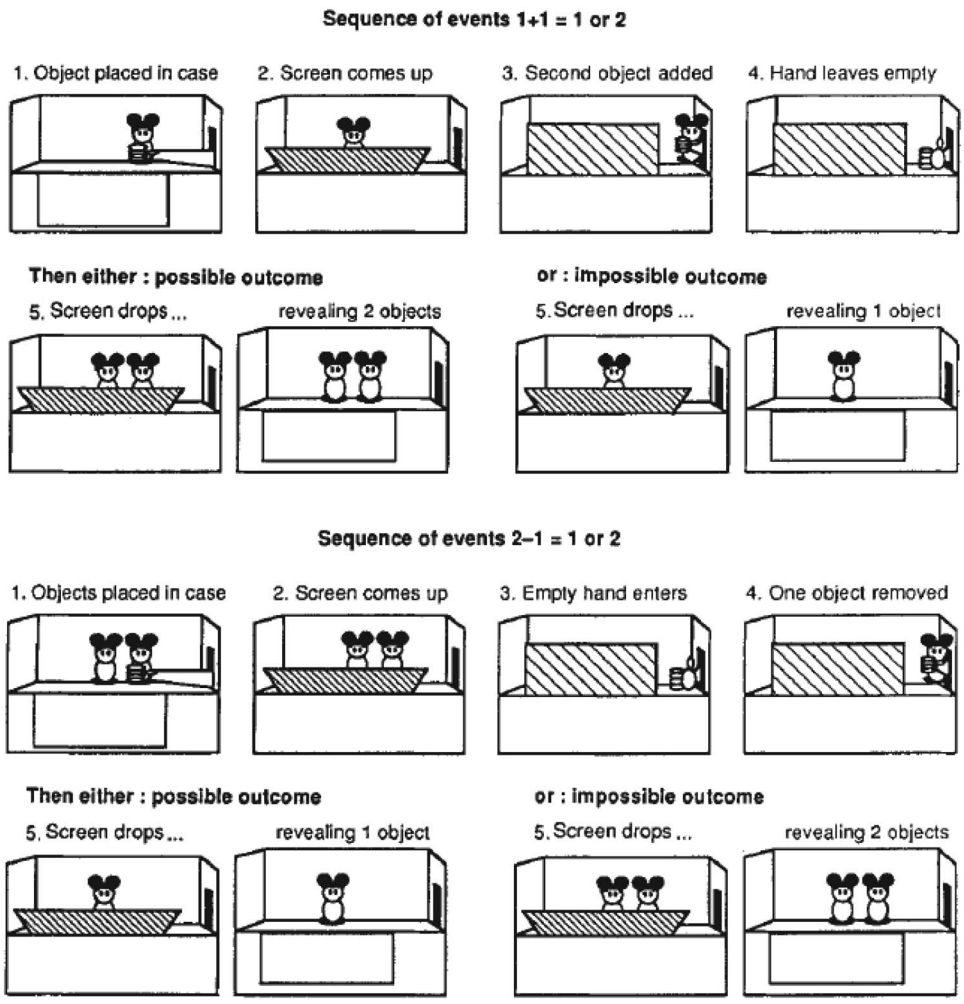

Figure 1. Illustration of Wynn's (1992) puppet stage design. Reprinted with permission from "Addition and Subtraction by Human Infants," by K. Wynn, 1992, *Nature*, *358*, p. 749. Copyright 1992 by Nature Publishing Group.

psychology textbooks (e.g., Myers & Dewall, 2016), and recent textbooks on child development (e.g., Lightfoot, Cole, & Cole, 2012). The current meta-analysis compiled all of the subsequent replications and extensions of Wynn's experiment to systematically evaluate the replicability of Wynn's phenomenon.

Since the original 1992 investigation, many researchers have attempted to replicate and/or extend this work; some of this research was designed to test alternative explanations for the original findings. While some of these studies successfully replicated Wynn's effect (Berger, Tzur, & Posner, 2006; Cohen & Marks, 2002; Kobayashi, Hiraki, Mugitani, & Hasegawa, 2004; McCrink & Wynn, 2004, 2009; Slater, Bremner, Johnson, & Hayes, 2010), several studies reported mixed findings (Clearfield & Westfahl, 2006; Koechlin, Dehaene, & Mehler, 1997; Moore & Cocas, 2006; Uller, Carey, Huntley-Fenner, & Klatt, 1999; Wakeley, Rivera, & Langer, 2000).

The Value and Limitations of Meta-Analyses

Meta-analyses are useful in such cases of high variability in the literature. This type of quantitative statistical analysis is usually possible only after enough time has passed to allow for the completion of a reasonable number of attempted replications. Now that Wynn's original report is 25 years old, it is finally possible to pool data from separate-but-similar experiments that have been conducted by independent researchers, to test the replicability of the findings. This method provides greater statistical power than can be found in a single study and allows for a broader investigation of patterns and relationships involving 2 or more variables (Hedges & Olkin, 1985).

Furthermore, with greater sample sizes, a meta-analysis considers multiple moderators across studies. Specifically, interesting subgroups of studies are evaluated to assess whether effect sizes differ between these groups. In meta-analyses, moderator effects reveal the relationship between the moderator variable and the effect size, disclosing how moderators affect the direction or strength of the effect (Hedges & Olkin, 1985).

Cantrell and Smith (2013) recently conducted a qualitative review on the extant literature on studies of infant numerical abilities over the past 30 years and reported inconsistencies therein. Their review considered the methods used to measure infant numerical abilities—including head turn procedures, violation of expectation procedures, and habituation procedures—and assessed the limitations of the subsequent interpretations drawn from these approaches. Cantrell and Smith also considered ways to improve experimental control in future studies, in an effort to reduce the inconsistencies reported in the literature. The current meta-analysis complements Cantrell and Smith's investigation by systematically and quantitatively pooling and analyzing all of the data generated in the attempted replications that followed Wynn's (1992) investigation. These experiments specifically aimed to test infant numerical abilities using violation-of-expectation paradigms to assess addition and subtraction abilities.

It is important to understand that while a meta-analysis can assess the extent to which the effect Wynn reported has proven replicable, it cannot address questions about the proper *interpretation* of these sorts of results. Wynn's experiment, like all of the studies that followed it, utilized violation-of-expectation methods that relied on the duration of infant looking at the "correct" versus "incorrect" number of items in a test display. Looking time procedures are useful in assessing infants' visual behaviors, but researchers must be cautious in their subsequent interpretations, for the following reason.

Although research has shown that looking time can provide insight into infants' perceptual and cognitive processes (e.g., Fagan, 1970; Kavšek, 2013; Moore & Johnson, 2008, 2011), interpretations of looking time patterns found in violation-of-expectation studies must consider a variety of factors previously identified as being able to influence looking: familiarity and novelty effects, stimulus intensity effects, as well as an infant's age and other individual differences (Schöner & Thelen, 2006). Thus, before interpreting a visual preference for an "incorrect" number of items as being a reflection of an infant's arithmetical abilities, researchers should consider simple alternative interpretations based on the multiple factors known to influence looking. Although individual studies included in the current meta-analysis attempted to explore such alternative interpretations, this meta-analysis should not be expected to resolve these interpretational questions, because meta-analyses cannot establish *why* infants may look at the "incorrect" more than the "correct" number of items in a display. Instead, given the variability in the results of Wynn-style studies, the purpose of this meta-analysis is to consider the *replicability* of Wynn's original effect; this must be investigated independently of addressing any questions of interpretation.

Because Wynn's original effect was unexpected in the developmental science community, numerous researchers have been inspired to replicate and extend this research by exploring alternative explanations for her phenomenon, including explanations invoking possible object identity and location expectancies, preferences for certain characteristics of the mathematical operation used, and familiarity preferences. These replications and extensions highlighted several interesting characteristics of this research and eventually motivated the selection of moderators examined in this meta-analysis.

Experimental Variations: Stimulus Placement and Physical Properties

Some extensions of the Wynn design tested alternative explanations concerning object placement and the physical properties of the objects in the displays; these studies reported mixed results (Koechlin et al., 1997; Simon, Hespos, & Rochat, 1995; Uller et al., 1999). Koechlin and colleagues extended Wynn's (1992) phenomenon by investigating possible object location expectancies as an alternative explanation. Their research was designed to exclude the possibility that infants' understandings of objects' behaviors in space might permit location-based inferences that could underlie their performances in Wynn's experiment. That is, these researchers considered the prospect that Wynn's infants behaved as they did not because they are mathematically competent, but because they are able to infer the trajectory of a moving hidden object, and thereby attribute different locations to objects that occupied distinct locations previously (Baillargeon, 1994; Baillargeon & DeVos, 1991; Hespos & Baillargeon, 2001). Koechlin and colleagues used a rotating tray to ensure that the location of the object was not predictable and reported mixed effects for the subtraction and addition conditions; specifically, a significant effect was found only for the subtraction conditions. These results are not sufficient

to support a mathematical interpretation, as infants may have just preferred to look at more items (2) versus fewer items (1). Previous research consistent with this possibility suggests that infants tend to prefer fixating more complex stimuli over less complex stimuli (Courage, Reynolds, & Richards, 2006; Martin, 1975). Given this alternative interpretation of these mixed results, it remains unclear if location-based inferences influence performance in this paradigm. Similarly, Simon and colleagues (1995) found varying effects in a replication and extension using Ernie and Elmo dolls to assess the influence of violations of object identity on infants' performances. Their test display contained a combination of arithmetically and physically possible and impossible "solutions" (i.e., Elmo + Elmo = Elmo [impossible arithmetic]; Elmo + Elmo = Ernie and Ernie [impossible identity]; Elmo + Elmo = Ernie [impossible arithmetic and identity]). Although the observed effects were consistently positive, they were stronger in the subtraction than the addition condition. Thus, the status of these alternative explanations based on location reasoning or knowledge of the physical properties of the objects has not been resolved.

Moreover, Uller and colleagues (1999) considered an explanation based on the timing and placement of the screen relative to the objects on the stage, and reported mixed results. Infants looked longer at the "incorrect" test displays in the addition condition only when they saw the following series of events: The first object was placed on the stage, a screen was then raised that covered the object, and finally the second object was introduced (as used by Koechlin et al., 1997; Simon et al., 1995; and Wynn, 1992). However, when the infants saw two objects placed on the stage one-by-one behind a screen that was raised from the start, no such differences in looking times were found (Uller et al., 1999). These results suggest that infants may be creating mental models of the actual objects on the stage floor, referred to as "object files" (Kahneman, Treisman, & Gibbs, 1992), rather than counting the number of objects placed on the stage.

Experimental Variations: Numerical Characteristics of Stimuli

Wakeley and colleagues (2000) and Cohen and Marks (2002) used a different number of objects in the Wynn paradigm and reported mixed results. Wakeley and colleagues reported no significant effects in computerized versions of Wynn's experiments, including a substitution variant that began with a different number of items (3 − 1 = 1 or 2). However, Cohen and Marks (2002) included other "solutions" to the original Wynn addition (0, 2, and 3) and subtraction (0, 1, and 3) problems and still found that infants looked longer at the "incorrect" arithmetic outcomes (with the exception of zero; infants will not typically fixate empty displays). Although several of the Cohen and Marks results were consistent with Wynn's original results, the inconsistent results reported by Wakeley and colleagues suggested that changing the number of objects used may challenge the reliability of Wynn's findings as well as her arithmetical interpretations of those findings.

Experimental Variations: Familiarization Effects

Several researchers (Clearfield & Westfahl, 2006; Cohen & Marks, 2002; Moore & Cocas, 2006; Slater et al., 2010) explored the possibility that infants' behaviors in Wynn-style experiments reflect not mathematical competence, but a preference for the initial number of objects displayed. After all, infants in subtraction conditions typically see two objects before the screen obstructs their view, and the two-object result is the "incorrect" display. Likewise, infants in addition conditions typically see one object before the obstruction event, and the one-object result is the "incorrect" display. Perhaps infants look at the incorrect displays not because they are "mathematically surprising," but because they are the same displays seen just before the obstruction event. Although infants who have become habituated to repeatedly presented stimuli typically prefer *novel* stimuli in subsequent preference tests (Flom & Pick, 2012; Kavšek, 2013; Sirois & Mareschal, 2004), there is also evidence for familiarity preferences in infancy, where infants prefer previously seen displays (for review, see Houston-Price & Nakai, 2004); these kinds of effects typically occur when infants are *familiarized* with a stimulus set but are not allowed to become fully habituated to it.

Several primary investigations examined the possibility that the Wynn results might reflect such familiarization effects (Clearfield & Westfahl, 2006; Cohen & Marks, 2002; Moore & Cocas, 2006; Slater et al., 2010). These studies involved providing infants with varying levels of exposure to one or two objects before exposing them to the standard addition or subtraction events. Whereas Wynn included only two pretrials that may have familiarized the infants to one or two objects, Cohen and Marks as well as Clearfield and Westfahl used eight familiarization trials, and Slater and colleagues used six familiarization trials. Moore and Cocas used an infant-controlled procedure in which a single trial continued for up to 30 s (unless the infant looked away from the object for two cumulative seconds); this procedure increased the likelihood that the infants were habituated to—rather than merely familiarized with—the test displays. Despite including more familiarization events relative to the original Wynn study design, neither Cohen and Marks nor Slater and colleagues reported increased preferences for the nonsurprising ("correct") test displays; instead, infants looked longer at the "incorrect" test display, supporting the Wynn effect.

However, familiarity influenced the results in some of the conditions in Clearfield and Westfahl (2006), as well as in Moore and Cocas (2006). Although Clearfield and Westfahl reported longer looking times for the "incorrect" outcome in 3- to 5-month-olds in their exact replication of Wynn's addition condition, when infants were shown eight familiarization trials in subsequent experiments, the results were inconsistent. Infants familiarized with the incorrect outcome looked longer at the correct outcome, while infants familiarized with the correct outcome had no preference (Clearfield & Westfahl, 2006). Looking data from the male infants in Moore and Cocas supported an effect of habituation; they spent less time looking at the same number of items on display as previously seen in lengthy familiarization trials. Considering the data from Cohen and Marks (2002) and Slater and colleagues (2010), and partly from Clearfield and Westfahl (2006) and Moore and Cocas (2006), simple familiarity preferences alone do not seem to account for the Wynn (1992) effect.

Current Study

This meta-analysis aimed to systematically compile and quantitatively synthesize all of the primary replications and extensions

of Wynn (1992). Unfortunately, although Wynn reported the mean looking times for each group of infants in her first two experiments, she determined statistical significance using tests that compared two sets of difference scores (each reflecting infants' visual preferences for 2 items vs. 1 item) generated by infants in the addition versus the subtraction group. However, this omnibus statistical analysis does not directly assess the statistical difference between looking times at the correct versus the incorrect test displays. Moreover, although means for the addition and subtraction conditions were reported, standard deviations were not available to properly compute the effect size. These circumstances prevented us from using Wynn's first two experiments in the current meta-analysis. Nonetheless, Wynn's third experiment was included as it involved only an addition group and *t* test results comparing the infants' preferences for 2 versus 3 items. This experiment was incorporated in the synthesis with subsequent studies that assessed statistical differences between looking at the incorrect versus the correct displays.

Based on a review of methodological variations in previous research, a number of potential theoretical and methodological metamoderators were coded: replications versus extensions, mathematical or numerical characteristics of the stimuli, whether or not familiarization trials were included, infant age, stimulus and display type, attrition, and publication year. Considering the inconsistencies in the experimental findings to date, no direction regarding the summary effect size was hypothesized.

Method

Literature Search

Primary studies were identified using a combination of approaches. PsycInfo, Google Scholar, and ProQuest were searched to identify conceptual replications and extensions of Wynn's (1992) original study. The extensive search to identify pertinent studies was conducted by using the following keywords: infants, infancy, add, subtract, number, numbers, and math. Two keywords were entered into each search: one representing the population (infants or infancy) and one representing the phenomenon (add, subtract, number, numbers, or math). This approach resulted in 10 independent searches (2 population terms × 5 phenomenon terms).[1] After relevant studies were located, the papers and their reference sections were reviewed to perform ascendency and descendancy searches until no further relevant investigations were located (Crano, Brewer, & Lac, 2015).

Five researchers who published work related to Wynn's original study were also contacted via e-mail: Three were asked to help identify relevant literature; one was contacted regarding an unpublished paper but the information to calculate the effect size was unavailable; and another was the author of a published paper who did not respond to our request. No dissertations or unpublished studies were found on Proquest or reported by the other five researchers contacted.

Inclusion and Exclusion Criteria

The main inclusion criterion for the meta-analysis was the use of an experimental method designed to replicate or extend Wynn's (1992) original study. Each included study had to contain at least one sample in which infants responded to a mathematically incorrect and correct number of items (incorrect and correct conditions), and each study had to compare infants' looking times at these items. Most studies tested multiple samples.[2] If a study featured several experiments administered to the same set of participants, only one effect size was culled to ensure statistical independence of samples in the synthesis. A sample was excluded if insufficient information was reported to compute the effect size, or if the researcher did not provide required statistics when contacted. Two of the three independent samples from Wynn's (1992) original study were excluded for both reasons. Another study (Kobayashi et al., 2004) was excluded because its cross-modal method involved auditory tones as well as visual objects, and therefore was very different than the Wynn paradigm and unlike any of the other replications or extensions.

Variable Coding

The majority of the primary publications (9 out of 12 studies) contained more than one sample of infants, so an independent effect size was calculated for each sample, consistent with the guidelines for the assumption of independence (Hedges & Olkin, 1985; Hedges & Vevea, 1998). Thus, independent samples represented the unit of analysis in this synthesis. Accordingly, effect sizes and potential moderators were coded for each sample. Based on a coding guide, two coders independently coded characteristics and calculated the effect size for each sample. Discrepancies were discussed and reconciled by the two coders.

Calculating Effect Sizes (Cohen's *d*)

The effect size index, Cohen's *d*, represents the difference between the mean looking times (dependent variable) at the incorrect and correct test displays (independent variable) divided by the pooled/common standard deviation (Cohen, 1992; Hedges & Olkin, 1985). Longer average looking times at the incorrect over the correct test display yielded a positive Cohen's *d* value, whereas longer looking times at the correct over the incorrect test display yielded a negative Cohen's *d* value. All effects were corrected for sample size bias using Hedges' *G* adjustment (Hedges & Vevea, 1998).

Dunlap, Cortina, Vaslow, and Burke (1996) recommended that means and standard deviations should be the primary statistical information used to compute the effect size for paired or matched designs in a meta-analysis. Therefore, means and standard deviations, when available, were used to derive effect sizes; these are the rawest values typically reported in papers. If the means and standard deviations were not reported, the effect size was calculated using the paired *t* test, repeated *F* test (two levels with 1 degree of freedom in the numerator), or exact *p* value along with the sample size. None of the studies reported the autocorrelation

[1] To state this in Boolean logic terms, our search involved seeking true solutions to the following statement: (add OR subtract OR number OR numbers OR math) AND (infant OR infancy).

[2] Wynn (1992)—like many of the subsequent replications and extensions—used a between-subjects design in which some infants were assigned to an addition condition and others were assigned to a subtraction condition. For the purposes of this meta-analysis, these are considered to be two separate samples, because each group of infants yielded an independent effect size.

between the two dependent conditions, so the standard $r = .50$ was used.

Model Estimation and Heterogeneity Tests

The weighted summary effect was computed using a random-effects model (Borenstein, Hedges, Higgins, & Rothstein, 2009; Lac, 2014). Homogeneity of the distribution of effect sizes was evaluated to determine whether effect sizes across samples varied more than might be expected due to random sampling variability (Gliner, Morgan, & Harmon, 2003). The Q test assessed the presence versus absence of effect size heterogeneity, and the I^2 index assessed the extent of such heterogeneity in the summary effect (Hedges & Olkin, 1985).

Meta-Moderator Analysis

To help explain systematic variation across the distribution of effect sizes, 10 potential metamoderators of the summary effect were tested to determine differences between subgroups of samples (Borenstein & Higgins, 2013). The first metamoderator examined the effect of exact/close replications versus extensions of the seminal experiment. Samples were rated as exact/close replications if they included the original Wynn (1992) addition and/or subtraction conditions ($1 + 1 = 1$ or 2; $2 - 1 = 1$ or 2; $1 + 1 = 2$ or 3) and if the experimental design did not differ discernibly. Samples coded as extensions differed appreciably, for example by using a different number of outcomes or familiarization trials, by using a different order of screen or object placement, or by including rotating objects, and so forth.

Three potential metamoderators concerned mathematical or numerical characteristics of the stimuli: the mathematical operation represented, the number of objects presented in the addition versus subtraction displays, and the number of objects presented in the test displays. The mathematical-operation moderator allowed comparison of infants' behaviors in addition versus subtraction conditions. The number-of-objects moderator compared infants' behaviors after viewing the addition or subtraction displays containing only 1 or 2 items versus addition or subtraction displays that included other numerosities (e.g., $3 - 1$). The final moderator concerning numerical characteristics of the stimuli compared infants' behaviors when presented with Wynn's original test displays (1 and/or 2) versus test displays containing other numbers of items (0 and/or >2).

Another potential metamoderator tested was the inclusion (or not) of familiarization trials. All samples that used familiarization trials presented infants with both correct and incorrect test displays prior to the experimental trials. The results of the studies using familiarization trials were mixed; thus, this characteristic was coded to determine if the effects generated by the samples that saw familiarization trials differed from those generated by the samples that did not.

An additional metamoderator, infant age, was assessed by median split, to investigate if infants younger than or equal to 162 days (about 5 months) behaved differently than older infants. For the moderator analysis, studies that tested infants this age or younger represented one subsample and studies that tested infants older than this age represented the other subsample. Prior research examining various aspects of infant cognitive development has shown differences in results for younger versus older infants. For example, Moore and Johnson (2008, 2011) found novelty preferences in a spatial task used with 5-month-olds, but familiarity preferences in the same task used with 3-month-olds. Thus, age was tested as a potential moderator, as infants younger than 5 months old may respond differently than older infants.

The next two metamoderators assessed were stimulus type (geometric vs. toy) and display type (puppet show vs. computer screen). These two metamoderators were examined because less attractive stimuli may have drawn less attention, thereby influencing performance. For example, if three-dimensional geometric shapes presented in a puppet show display (Wynn, 1992) did not attract the infants' attention, the attrition rate may have been higher (because infants whose attention is lost tend to become fussy). These moderators could influence how engaging the infants found the stimulus displays and how much attention was paid, which may alter the size of the effect.

Furthermore, an additional moderator assessed one of the most common reasons for attrition in infant studies: the number of participant exclusions due to fussiness (e.g., Moore & Cocas, 2006; Wakeley et al., 2000). Some papers only reported a single value denoting the total number of infants who were excessively fussy across all the independent samples. In such instances, the total number of fussiness exclusions was divided by the number of samples to obtain an estimate of fussiness-exclusions per sample. Next, the fussiness moderator was converted to odds: The number of infants who were excluded for fussiness was divided by the number who eventually participated in the experiment. Larger odds represented greater likelihood of exclusion versus remaining in the sample, with a median split ($odds = 0.35$) applied for the moderation analysis.

A final moderator—publication year by median split (published before vs. after January 1, 2003)—was assessed to explore possible trends in earlier versus more recent publications. This moderator is commonly used in meta-analyses to evaluate paradigm shifts in research over time. All of the aforementioned metamoderators were analyzed by comparing the mean effect sizes of the two subsets of studies and testing for statistically significant differences (Hedges & Olkin, 1985).

Results

Common Sample Features

Several methodological features characterized the majority of samples in this area of research. First, each independent sample used within-subjects designs, in which infants served in both the incorrect and correct conditions (independent variable). That is, infants viewed displays in a sequence of test trials, a sequence that presented both mathematically "correct" and "incorrect" numbers of items in different test trials. Second, infants were exposed to multiple trials, and their looking-time scores were averaged across trials in each condition. Third, mean looking time (in seconds) served as the dependent measure.

Characteristics of Samples

The inclusion procedures yielded 12 published primary studies, containing 26 independent samples, with 550 total infants for the

quantitative synthesis. Table 1 lists, by publication date, the names of the authors, sample size, mean and standard deviation (in seconds) for the incorrect and correct conditions, effect size, and *p* value for each sample. A positive Cohen's *d* denotes longer looking times at the incorrect compared to correct test display, whereas a negative value indicates a preference in the opposite direction. Of the 26 samples, 22 yielded positive effect size values, and 12 attained statistical significance, $p < .05$ (two-tailed) with two samples that were significant at $p < .06$.

The *p* values computed based on the effect sizes (reported in Table 1) were similar to those reported in most of the journal articles, but they were not always identical. As in any meta-analytic review that requires inferences based on information reported in primary studies, discrepancies may have occurred for several reasons. First, in some cases, rounding error may have occurred due to calculating effect sizes from studies reporting their statistics up to about two decimal places. Second, in some cases, a one-tailed statistical test was applied in the study, but a more conservative two-tailed test was used for all analyses in the current synthesis. Third, the quality and amount of statistical information reported varied across journal articles. For instance, the sample size reported in the *Participants* section was sometimes discrepant from the sample size deduced from the degrees of freedom (*df*) reported for the analysis. Such a discrepancy might be attributable to listwise deletion of missing values or misreporting by the study's authors. To compute effect size, we used the sample sizes listed in the *Participant* sections, which were probably least susceptible to misreporting. Fourth, in some cases, the statistical test was evaluating data generated in a complex experimental design that included additional factors and covariates (e.g., gender), which might have reduced the error term and thereby made the reported result more likely to attain statistical significance. The current meta-analysis focused on the main effect of the independent variable on looking times at the incorrect and correct test displays.

Summary Effect Size

First, a hierarchical meta-analysis was conducted to determine the extent of clustering effects for samples within each paper; this was an important first step because any given study conducted in the same laboratory by one team of authors might be susceptible to generating samples with effect sizes of more similar magnitudes than those generated in studies conducted across different laboratories (Stevens & Taylor, 2009). Following the recommended guidelines for estimating a meta-analytic model using a hierarchical weighting scheme (Hedges, Tipton, & Johnson, 2010; Tanner-Smith & Tipton, 2014), the variance components for tau-squared (study-to-study) and omega-squared (samples within studies) were 0.09 and 0.00, respectively. Thus, the independence of sample-to-sample effects within the same paper was supported, so standard (nonhierarchical) meta-analytic procedures were then implemented.

Table 1
Characteristics for Each Sample

Study	Study author(s)	Sample	*N*	Mean (*SD*)		Cohen's *d*	*p*-value
				Incorrect condition	Correct condition		
1	Wynn (1992)	1	16	11.89	9.96	.49	.055
2	Simon, Hespos, and Rochat (1995)	2	10	11.25 (8.47)	8.21 (4.93)	.38	.210
2	Simon, Hespos, and Rochat (1995)	3	10	10.76 (5.98)	6.77 (4.24)	.68	.036
3	Koechlin, Dehaene, and Mehler (1997)	4	8	10.80	10.60	.03	.927
3	Koechlin, Dehaene, and Mehler (1997)	5	7	14.50	9.80	1.58	.003
3	Koechlin, Dehaene, and Mehler (1997)	6	7	13.50	8.40	1.28	.007
3	Koechlin, Dehaene, and Mehler (1997)	7	7	10.00	10.90	−.33	.330
4	Uller et al. (1999)	8	32			.64	<.001
4	Uller et al. (1999)	9	32			.16	.371
5	Wakeley, Rivera, and Langer (2000)	10	22	8.85 (5.45)	9.76 (5.78)	−.16	.451
5	Wakeley, Rivera, and Langer (2000)	11	22	9.70 (5.18)	8.85 (5.45)	.16	.437
5	Wakeley, Rivera, and Langer (2000)	12	24	9.50 (5.98)	8.72 (5.68)	.13	.514
6	Cohen and Marks (2002)	13	40	10.94 (7.36)	7.82 (4.97)	.47	.004
6	Cohen and Marks (2002)	14	40	9.95 (7.67)	7.09 (6.96)	.38	.018
7	McCrink and Wynn (2004)	15	13	10.28 (3.87)	7.35 (3.04)	.78	.010
7	McCrink and Wynn (2004)	16	13	9.13 (5.93)	8.00 (6.16)	.17	.504
8	Berger, Tzur, and Posner (2006)	17	24	8.04 (4.66)	6.94 (4.37)	.24	.240
9	Clearfield and Westfahl (2006)	18	16	4.26	3.04	.54	.036
10	Moore and Cocas (2006)	19	31	12.08 (5.90)	11.13 (6.28)	.15	.389
10	Moore and Cocas (2006)	20	31	12.87 (6.51)	11.88 (6.65)	.15	.405
10	Moore and Cocas (2006)	21	46	8.30 (4.39)	9.62 (4.80)	−.28	.057
10	Moore and Cocas (2006)	22	43	8.59 (4.52)	8.85 (6.20)	−.05	.759
11	McCrink and Wynn (2009)	23	12	4.37	2.77	.70	.021
11	McCrink and Wynn (2009)	24	12	7.42	4.51	.73	.017
12	Slater et al. (2010)	25	16			1.11	<.001
12	Slater et al. (2010)	26	16			.74	.006

Note. Empty cells indicate the information was not reported in the study. If a study featured several experiments administered to the same set of participants, only one effect size was culled to ensure statistical independence of samples in the synthesis. Furthermore, a sample was excluded if insufficient information was reported to compute the effect size, or if the researcher did not provide required statistics when contacted.

The random-effects model produced a summary effect of $d = +0.34$, $Z = 4.59$, $p < .001$ (95% confidence interval [CI] 0.19 to 0.48) across the 26 separate samples.[3] The fixed-effects model resulted in a summary effect of $d = +0.27$, $Z = 6.14$, $p < .001$ (95% CI [0.18 to 0.35]). All subsequent meta-analytic procedures, including moderation analyses, were weighted based on the random-effects model, as it accounts for more sources of variance and is less prone to Type I error compared to a fixed-effect model (Hunter & Schmidt, 2000; Lac, 2014).

Evaluation of Publication Bias

Investigations with statistically significant findings are more likely than investigations with null results to be published. Consequently, publication bias tends to support the direction of the hypothesized effects and contributes to an overestimation of the magnitude of the summary effect (Rosenthal, 1979). Several approaches were undertaken to evaluate the possibility of publication bias. First, the "fail-safe N" formula (Rosenthal, 1979) was applied to calculate the number of potentially unpublished studies exhibiting a null effect ($d = 0.00$) that would have to be incorporated into the meta-analysis to render the summary effect nonsignificant at $p > .05$. This formula suggested that 336 unpublished null-result samples would need to exist to render the summary effect we detected statistically nonsignificant.

Second, the exploratory test for an excess of significant findings was conducted (Ioannidis & Trikalinos, 2007), applying the variant for random-effects models (Schimmack, 2012, 2016), to assess the extent of publication bias by comparing the number of findings observed to be significant with the number of findings expected to be significant from power analyses. Across all the samples, the median observed power was .46 based on $p < .05$ (two-tailed), indicating that research in this domain tends to be underpowered in comparison to the recommended power of .80 (Cohen, 1992). The meta-analysis revealed that 12 samples attained significance, with the excess test of significant findings instead proposing that 11 were expected to attain significance based on the observed power. A binomial probability test indicated that these two frequencies were not significantly different, suggesting that selective publication of significant findings was minor.

Third, a funnel plot of the distribution of effect sizes was interpreted as a visual aid to detect publication bias (see Figure 2). To produce a funnel plot, the effect size for each sample is plotted on the X-axis, and the standard error (inversely related to sample size) is plotted on the Y-axis. The vertical line above the diamond in Figure 2 represents the location of the summary effect. Visual inspection of the dispersion of effect sizes in the funnel plot revealed some asymmetry, suggesting the likelihood of at least minor publication bias (Borenstein et al., 2009).

Fourth, a supplementary nonparametric technique—a trim and fill analysis (Duvall & Tweedie, 2000)—was pursued to estimate the number of missing "file drawer" samples that may exist and assess the impact of these findings if hypothetically incorporated into the meta-analysis. This test indicated that incorporating five additional samples with negative effect sizes (opposite to the hypothesized direction of effects) would be sufficient to produce a symmetrical distribution in the funnel plot of Figure 2. As part of this statistical procedure, if the "missing" samples were hypothetically included, the adjusted summary effect would decrease but still remain significant, $d = +0.24$ (95% CI [0.09 to 0.39]), $p < .01$. Thus, despite the evidence for publication bias revealed by the trim and fill analysis, the cumulative effects in the synthesis were sufficient to produce a significant summary effect even after statistically imputing the missing samples.

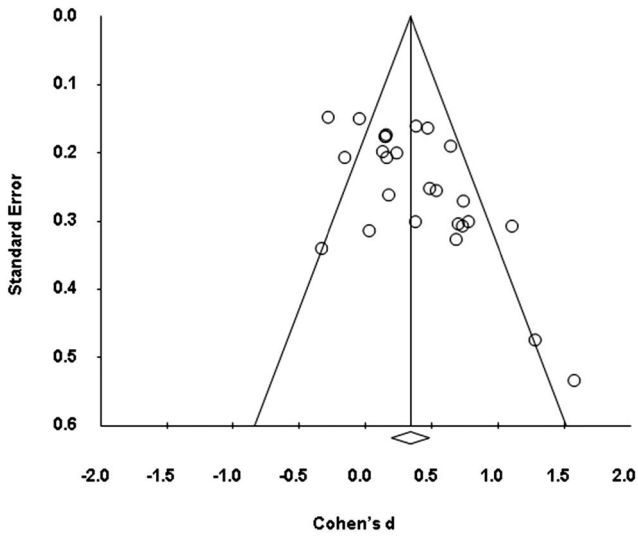

Figure 2. Funnel plot of effect sizes.

Fifth, the PET-PEESE test (Stanley & Doucouliagos, 2014), based on a weighted least squares metaregression approach, was performed. Adhering to interpretational guidelines, the PEESE (precision effect estimate of standard error) was ascertained to be more appropriate compared to the PET (precision effect test). The technique involves building a regression equation (weighted by the inverse of the variances) in which the variance (squared standard errors) of each sample is used to predict the distribution of effect sizes. The intercept of the equation is then interpreted based on the theory that this would furnish information about the hypothetical effect size given a standard error near zero (i.e., large sample size). The PEESE test indicated the presence of publication bias, $p < .05$, and that Cohen's d would be $-.01$ after correcting for sample size, suggesting strong publication bias.

Summary of Meta-Moderators

The distribution of effects across the samples was tested to determine whether the dispersion might be explained by metamoderators, $Q(25) = 66.03$, $p < .001$; $I^2 = 62.14$. Given that the Q test exhibited significance in sample-to-sample variability beyond sampling error and that the I^2 index corroborated the lack of homogeneity, the exploration of potential metamoderators was warranted. As shown in Table 2, 10 possible sample-level characteristics were tested to determine whether subgroups of samples

[3] Sensitivity analyses applying other autocorrelation values were examined. An $r = 0.25$ indicated a summary effect (random-effects model) of $d = +0.38$, $Z = 4.45$, $p < .001$. An $r = 0.75$ indicated a summary effect (random-effects model) of $d = +0.29$, $Z = 4.73$, $p < .001$. These effects are consistent with the direction of the summary effect reported in the results section when $r = 0.50$.

Table 2
Meta-Moderators

Moderator	Subgroups	*k*	*d*	95% CI LL	95% CI UL	*Q* test of moderation
Replications vs extensions						.28
	Replications	8	.40*	.12	.68	
	Extensions	18	.31**	.14	.49	
Addition vs. subtraction						.26
	Addition	15	.31*	.11	.51	
	Subtraction	11	.38**	.17	.60	
Mathematical operations						.62
	1 + 1 or 2 − 1	21	.31**	.15	.48	
	Other	5	.45*	.15	.75	
Number of "solution" items						2.50
	1 and/or 2	19	.28*	.10	.46	
	0 and/or >2	7	.48**	.31	.65	
Familiarization trials						.81
	Yes	13	.40**	.19	.61	
	No	13	.27*	.07	.47	
Age by median split						.14
	≤162 days	16	.36**	.16	.57	
	>162 days	10	.31*	.12	.50	
Stimulus type						.31
	Toys	12	.38**	.21	.55	
	Geometric figures	14	.30*	.08	.53	
Display type						1.25
	Puppet show	18	.39**	.23	.56	
	Computer Screen	8	.22	−.03	.47	
Fussiness exclusion odds by median split						.08
	≤.35	13	.32*	.12	.51	
	>.35	12	.36*	.12	.61	
Publication year by median split						.02
	≤2002	14	.33**	.15	.51	
	>2002	12	.35*	.12	.58	

Note. The *Q* tests of moderation indicate that none of the meta-moderators were significant. The notation "*k*" refers to the number of samples in each subgroup. The upper limit (UL) and lower limit (LL) of the 95% confidence intervals for each subgroup's effect size are listed.
* $p < .01$. ** $p < .001$.

generated statistically different effect sizes. None of the characteristics statistically moderated looking times at the incorrect versus correct test displays. Thus, the variations in sample and methodological characteristics tested were not significantly related to studies' results (Hall & Rosenthal, 1991).

Although none of the tests of moderation indicated a significant *difference* between any two subgroups, it was informative to ascertain if the effect size generated by each subgroup was significantly greater than Cohen's $d = 0.00$. Such significant effect sizes were generated by the following subgroups: both replications and extensions, both addition and subtraction conditions, both types of mathematical operations (1 + 1 or 2 − 1, and other), both numbers of "solution" items (1 and/or 2; and 0 and/or > 2), both studies with and without familiarization trials, both age groups (≤162 days and >162 days), both stimulus types (toys and geometric figures), both fussiness groups (≤0.35 and >0.35), and both publication year groups (≤2002 and >2002). For the display type moderator, the puppet show subgroup generated an effect size that was significantly greater than Cohen's $d = 0.00$. The computer screen subgroup generated an effect size that was not significantly greater than Cohen's $d = 0.00$.

Discussion

The primary objective of this meta-analysis was to accumulate empirical evidence to evaluate the extent to which Wynn's (1992) original findings were reliably reproduced in other laboratories, using different samples and variations in methods. The overall summary effect size found was statistically significant, $d = +0.34$, $p < .001$, suggesting that infants prefer to fixate the test displays presenting the "mathematically incorrect" number of items. Although no metamoderators were identified, 19 out of 20 of the subgroups generated effect sizes that were significantly different from a null effect. Specifically, statistically significant summary effects were generated in replications and extensions of Wynn's study involving infants of various ages who were exposed to addition and subtraction paradigms that varied with respect to the inclusion (or exclusion) of familiarization trials as well as the type and number of items displayed. Notably, these results respond to one of the debates in the field regarding the influence of familiarization trials, suggesting that prior experience with the stimuli did not influence looking times during the test trials. Overall, these findings provide evidence that most of the variations in methods

used by different researchers did not significantly influence the direction or significance of the summary effect size.

None of the metamoderators were statistically significant, but one subgroup generated a statistically significant effect size and the corresponding subgroup generated a nonsignificant effect size. Specifically, infants in studies that used puppet show displays tended to exhibit the Wynn effect reliably. In contrast, infants in studies that used computerized displays did not. These findings have implications for future investigators studying Wynn's effect, as they indicate that puppet show displays in particular are able to elicit the effect. At least two potential explanations may account for this observation, one suggesting that puppet show stimulus displays may elicit attention from infants, and one suggesting that these types of displays may be characterized by reduced experimental control.

Viewing stimuli in a puppet display versus a computer screen may be more interesting to infants, and might therefore elicit more attention to the ongoing events and support any processing related to the stimuli. For the 11 samples that reported sufficient information to support an analysis to address this question, a meta-ANOVA was performed to compare the mean looking times at the test stimuli seen in puppet shows versus on computerized displays. The mean looking times for the test stimuli within each sample were pooled. Infants exposed to the puppet shows (*meta mean* = 9.14) versus the computerized displays (*meta mean* = 10.23) did not significantly differ on looking time at the test stimuli, Q *test* = 1.22, $p = .269$. This analysis did not support the proposed explanation.

Another possible explanation for this phenomenon considers the fact that puppet show displays naturally offer less experimental control and are potentially more susceptible to experimenter effects than are computerized displays. That is, experimenters manipulating the objects on puppet stages cannot be blind to experimental conditions, because they need to know what objects to place and remove from the displays and in which order. Ideally, these experimenters would be blind to the *hypotheses* of the studies to minimize biased behavior, but of the studies utilizing puppet show displays in the current meta-analysis, none of them addressed this question in their reports, so there is no way to know the extent to which experimenter effects might have influenced the outcomes of these studies. Unlike puppet show displays, using computer screens to present stimuli offers the possibility to control for such experimenter effects, as well as for different factors confounded with number (e.g., spatial extent) that may by manipulated on a computer screen in ways that would be difficult using puppet show displays. Thus, the finding that puppet show displays elicited the Wynn effect reliably, whereas computerized displays did not, likely reflects differences in experimental control and/or factors related to the ways in which the stimuli were manipulated in the various studies.

A potential limitation of this meta-analysis was the statistics available in the primary studies. Ideally, more reliable estimates of effect sizes can be calculated using the means and standard deviations of the comparison groups (Hedges & Olkin, 1985). However, only 54% (14 out of 26) of the samples reported these statistics; the remainder reported t statistics (27%, or 7 out of 26), or main effect F statistics (19%, or 5 out of 26). Although authors were consulted to obtain the raw descriptive statistics, they were not always made available.

Meta-analysis affords greater statistical power than a single primary study (Cohn & Becker, 2003). However, the detection of metamoderators tends to be underpowered in published quantitative syntheses (Harris, Hedges, & Valentine, 2009; Hedges & Pigott, 2004). Metamoderation analysis might not furnish sufficient power if combining experiments of small sample sizes. The current synthesis included all available primary research conducted to date, but data collection is often slow with infants, so infant studies contain smaller sample sizes compared to other fields. Although the effect sizes for some of the subgroups appeared to be different as a function of the metamoderators (e.g., display type or number of solution items), the associated Q tests of moderation did not reach the threshold of statistical significance.

The fail-safe N, excess test of significance, funnel plot, and trim and fill analysis all converged in suggesting minor to moderate publication bias. However, trim and fill analyses can be influenced when meta-analyses include a small number of studies, as in the current situation. Even with a large number of studies ($n > 200$), the algorithm used in the current analysis to detect asymmetry can be influenced by a single deviant study (Duval & Tweedie, 2000). Identifying and excluding such outlier studies is especially difficult in cases with a small number of studies. Thus, the trim and fill analysis might have underestimated publication bias. Furthermore, the PET-PEESE test disclosed that the significant summary effect was entirely attributable to publication bias (after adjusting for sample size). The discrepant findings from the five statistical procedures that were implemented to evaluate publication bias stem from the underlying statistical theory, models, and assumptions of the various approaches. For instance, the PET-PEESE has been criticized on grounds that it severely penalizes samples with a small N (Cunningham & Baumeister, 2016), is inappropriate for syntheses involving a limited number of studies (Cunningham & Baumeister, 2016), is sometimes inferior in performance compared to estimation methods that do not correct for publication bias (Reed, Florax, & Poot, 2015), and is premised on acceptance of the assumption that large sample sizes confer unbiased effect size estimates (Inzlicht, Gervais, & Berkman, 2015). Each of the other four tests used have been criticized on various grounds as well (e.g., Cunningham & Baumeister, 2016).

Future simulation studies conducted by methodologists should compare and contrast these various approaches of evaluating publication bias, for the purposes of identifying their strengths and weaknesses as well as the scenarios in which these procedures yield similar versus discrepant conclusions. Although the majority of the five implemented procedures revealed minor to moderate publication bias, one indicated extreme bias; the truth probably lies somewhere in the middle of this continuum, meaning that publication biases are still probable even if not supported by all of these statistical tests. Based on the overall evidence of these publication assessments as a set, the extent of publication bias is likely moderate, with the overall summary effect appearing to be positive even after accounting for this issue. Future meta-analyses are recommended when more empirical studies have been conducted on numerical transformations in infancy.

The summary effect size we discovered is reliable, but variability was found across the effect sizes reported in replications and extensions conducted since Wynn (1992). Furthermore, although infants behave reliably as Wynn predicted in these protocols, the interpretation that infants are conducting mathematical operations

remains open to debate, as noted in the Introduction; there are still several possible ways to interpret the finding that infants prefer to look at one object after a Wynn-style 1 + 1 display and at two objects after a Wynn-style 2 − 1 display (for review, see Cantrell & Smith, 2013). Interpreting the data in all but one of the samples included in this meta-analysis (i.e., Berger et al., 2006) required inferences based on infants' visual behaviors, and looking paradigms like the ones that generated these data were originally developed to address sensory and perceptual processing in infancy; only later were they applied to assess cognitive processing. Because looking behaviors are influenced by many display properties including perceptual features, novelty, familiarity, recency, predictability, and time lapse between stimulus exposures, several theorists have argued that interpretations of data based on looking times should consider possible simple perceptual explanations before accepting more complex cognitive explanations (Charles & Rivera, 2009; Haith, 1998; Kagan, 2013; Moore & Cocas, 2006).

Several extensions of Wynn (1992) reported mixed results when evaluating such alternative explanations, for example explanations based on the predictability of object location (Koechlin et al., 1997), or identity features of the stimuli (Simon et al., 1995). Likewise, although some extensions provided some support for alternative explanations based on familiarity preferences (Moore & Cocas, 2006), others did not (Cohen & Marks, 2002; Slater et al., 2010). Uller and colleagues' (1999) mixed results suggested that infants' behaviors in Wynn-style experiments might reflect their ability to build strong visual images of the objects (i.e., object files) rather than their ability to reason arithmetically. This suggestion is consistent with more recent results from Charles and Rivera (2009), who reported that infants' looking behaviors are sensitive to different methods of making objects disappear and reappear. Because of the difficulties associated with interpreting looking time data as meaningful about cognitive mechanisms, additional primary research including different methodologies will be required to evaluate various alternative interpretations of Wynn's phenomenon, as well as to evaluate the contribution of an early number sense to later mathematical learning.

For example, one replication and extension of Wynn's study reported looking time data as well as brain activity consistent with the detection of an error (Berger et al., 2006). Infants who had a visual preference for the incorrect test display also had greater negative activity in their time-locked event-related potentials when they were presented with the incorrect test display than when they were presented with the correct test display. This finding is consistent with brain activity found in adults when detecting an error (e.g., Brown & Braver, 2005), supporting Wynn's interpretation that infants' visual preferences for the incorrect test display are related to a violation of expectation or detection of a mathematical error. However, electrophysiological data were collected only for the infants who exhibited the behavioral effect, so it was not possible to compare the brain activity of the infants who did and did not prefer the incorrect test display. Future research using the violation-of-expectation paradigm and measuring both behavioral and electrophysiological activity may help clarify error detection in infants who are being tested for numerical or mathematical knowledge (or who are being tested in other domains in which detecting errors may contribute to understanding early cognitive development).

Furthermore, longitudinal measures assessing stability of individual differences in numerical competence are warranted to explore the contribution of a putative early numerical sense to mathematical understanding later in life. While this meta-analysis was not designed to rule out or rule in any possible *explanations* for infants' behaviors in Wynn-style experiments, the current quantitative review discovered that the Wynn effect is probably real and reliable, encouraging further research on this important topic.

References

References marked with an asterisk indicate studies included in the meta-analysis.

Baillargeon, R. (1994). How do infants learn about the physical world? *Current Directions in Psychological Science, 3,* 133–140. http://dx.doi.org/10.1111/1467-8721.ep10770614

Baillargeon, R., & DeVos, J. (1991). Object permanence in young infants: Further evidence. *Child Development, 62,* 1227–1246. http://dx.doi.org/10.2307/1130803

*Berger, A., Tzur, G., & Posner, M. I. (2006). Infant brains detect arithmetic errors. *Proceedings of the National Academy of Sciences of the United States of America, 103,* 12649–12653. http://dx.doi.org/10.1073/pnas.0605350103

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley. http://dx.doi.org/10.1002/9780470743386

Borenstein, M., & Higgins, J. P. (2013). Meta-analysis and subgroups. *Prevention Science, 14,* 134–143. http://dx.doi.org/10.1007/s11121-013-0377-7

Brannon, E. M., Abbott, S., & Lutz, D. J. (2004). Number bias for the discrimination of large visual sets in infancy. *Cognition, 93,* B59–B68. http://dx.doi.org/10.1016/j.cognition.2004.01.004

Brown, J. W., & Braver, T. S. (2005). Learned predictions of error likelihood in the anterior cingulate cortex. *Science, 307,* 1118–1121. http://dx.doi.org/10.1126/science.1105783

Cantrell, L., & Smith, L. B. (2013). Open questions and a proposal: A critical review of the evidence on infant numerical abilities. *Cognition, 128,* 331–352. http://dx.doi.org/10.1016/j.cognition.2013.04.008

Charles, E. P., & Rivera, S. M. (2009). Object permanence and method of disappearance: Looking measures further contradict reaching measures. *Developmental Science, 12,* 991–1006. http://dx.doi.org/10.1111/j.1467-7687.2009.00844.x

*Clearfield, M. W., & Westfahl, S. M. C. (2006). Familiarization in infant's perception of addition problems. *Journal of Cognition and Development, 7,* 27–43. http://dx.doi.org/10.1207/s15327647jcd0701_2

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159. http://dx.doi.org/10.1037/0033-2909.112.1.155

*Cohen, L. B., & Marks, K. S. (2002). How infants process addition and subtraction events. *Developmental Science, 5,* 186–201. http://dx.doi.org/10.1111/1467-7687.00220

Cohn, L. D., & Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological Methods, 8,* 243–253. http://dx.doi.org/10.1037/1082-989X.8.3.243

Cordes, S., & Brannon, E. M. (2009). Crossing the divide: Infants discriminate small from large numerosities. *Developmental Psychology, 45,* 1583–1594. http://dx.doi.org/10.1037/a0015666

Courage, M. L., Reynolds, G. D., & Richards, J. E. (2006). Infants' attention to patterned stimuli: Developmental change from 3 to 12 months of age. *Child Development, 77,* 680–695. http://dx.doi.org/10.1111/j.1467-8624.2006.00897.x

Crano, W. D., Brewer, M. B., & Lac, A. (2015). *Principles and methods of social research* (3rd ed.). Mahwah, N.J: Routledge.

Cunningham, M. R., & Baumeister, R. F. (2016). How to make nothing out of something: Analyses of the impact of study sampling and statistical interpretation in misleading meta-analytic conclusions. *Frontiers in Psychology, 7,* 1639. http://dx.doi.org/10.3389/fpsyg.2016.01639

Dunlap, W. D., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1,* 170–177. http://dx.doi.org/10.1037/1082-989X.1.2.170

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56,* 455–463. http://dx.doi.org/10.1111/j.0006-341X.2000.00455.x

Fagan, J. F., III. (1970). Memory in the infant. *Journal of Experimental Child Psychology, 9,* 217–226. http://dx.doi.org/10.1016/0022-0965(70)90087-1

Fagan, J. F., III. (1990). The paired-comparison paradigm and infant intelligence. *Annals of the New York Academy of Sciences, 608,* 337–364. http://dx.doi.org/10.1111/j.1749-6632.1990.tb48902.x

Flom, R., & Pick, A. D. (2012). Dynamics of infant habituation: Infants' discrimination of musical excerpts. *Infant Behavior and Development, 35,* 697–704. http://dx.doi.org/10.1016/j.infbeh.2012.07.022

Gliner, J. A., Morgan, G. A., & Harmon, R. J. (2003). Meta-analysis: Formulation and interpretation. *Journal of the American Academy of Child & Adolescent Psychiatry, 42,* 1376–1379. http://dx.doi.org/10.1097/01.chi.0000085750.71002.01

Haith, M. M. (1998). Who put the cog in infant cognition? Is rich interpretation too costly? *Infant Behavior and Development, 21,* 167–179. http://dx.doi.org/10.1016/S0163-6383(98)90001-7

Hall, J. A., & Rosenthal, R. (1991). Testing for moderator variables in meta-analysis: Issues and methods. *Communication Monographs, 58,* 437–448. http://dx.doi.org/10.1080/03637759109376240

Harris, C., Hedges, L., & Valentine, J. (2009). *Handbook of research synthesis and meta-analysis*. New York, NY: Russell Sage Foundation.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods, 9,* 426–445. http://dx.doi.org/10.1037/1082-989X.9.4.426

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1,* 39–65. http://dx.doi.org/10.1002/jrsm.5

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3,* 486–504. http://dx.doi.org/10.1037/1082-989X.3.4.486

Hespos, S. J., & Baillargeon, R. (2001). Infants' knowledge about occlusion and containment events: A surprising discrepancy. *Psychological Science, 12,* 141–147. http://dx.doi.org/10.1111/1467-9280.00324

Houston-Price, C., & Nakai, S. (2004). Distinguishing novelty and familiarity effects in infant preference procedures. *Infant and Child Development, 13,* 341–348. http://dx.doi.org/10.1002/icd.364

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment, 8,* 275–292. http://dx.doi.org/10.1111/1468-2389.00156

Inzlicht, M., Gervais, W., & Berkman, E. (2015). *Bias-correction techniques alone cannot determine whether ego depletion is different from zero: Commentary on Carter, Kofler, Forster, & McCullough, 2015*. SSRN. http://dx.doi.org/10.2139/ssrn.2659409

Ioannidis, J. P., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials, 4,* 245–253. http://dx.doi.org/10.1177/1740774507079441

Kagan, J. (2013). *The human spark: The science of human development*. New York, NY: Basic Books.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology, 24,* 175–219. http://dx.doi.org/10.1016/0010-0285(92)90007-O

Kavšek, M. (2013). The comparator model of infant visual habituation and dishabituation: Recent insights. *Developmental Psychobiology, 55,* 793–808. http://dx.doi.org/10.1002/dev.21081

Kobayashi, T., Hiraki, K., Mugitani, R., & Hasegawa, T. (2004). Baby arithmetic: One object plus one tone. *Cognition, 91,* B23–B34. http://dx.doi.org/10.1016/j.cognition.2003.09.004

*Koechlin, E., Dehaene, S., & Mehler, J. (1997). Numerical transformations in five-month-old human infants. *Mathematical Cognition, 3,* 89–104. http://dx.doi.org/10.1080/135467997387425

Lac, A. (2014). A primer for using meta-analysis to consolidate research. *Substance Use & Misuse, 49,* 1064–1068. http://dx.doi.org/10.3109/10826084.2014.862025

Lightfoot, C., Cole, M., & Cole, S. R. (2012). Counting. *The development of children* (pp. 161–200). New York, NY: Macmillan.

Lipton, J. S., & Spelke, E. S. (2004). Discrimination of large and small numerosities by human infants. *Infancy, 5,* 271–290. http://dx.doi.org/10.1207/s15327078in0503_2

Martin, R. M. (1975). Effects of familiar and complex stimuli on infant attention. *Developmental Psychology, 11,* 178–185. http://dx.doi.org/10.1037/h0076448

McCrink, K., & Birdsall, W. (2015). Numerical abilities and arithmetic in infancy. In R. C. Kadosh & A. Dowker (Eds.), *The oxford handbook of numerical cognition* (pp. 258–263). New York, NY: Oxford University Press.

*McCrink, K., & Wynn, K. (2004). Large-number addition and subtraction by 9-month-old infants. *Psychological Science, 15,* 776–781. http://dx.doi.org/10.1111/j.0956-7976.2004.00755.x

*McCrink, K., & Wynn, K. (2009). Operational momentum in large-number addition and subtraction by 9-month-olds. *Journal of Experimental Child Psychology, 103,* 400–408. http://dx.doi.org/10.1016/j.jecp.2009.01.013

Moore, D., Benenson, J., Reznick, J. S., Peterson, M., & Kagan, J. (1987). Effect of auditory numerical information on infants' looking behavior: Contradictory evidence. *Developmental Psychology, 23,* 665–670. http://dx.doi.org/10.1037/0012-1649.23.5.665

*Moore, D. S., & Cocas, L. A. (2006). Perception precedes computation: Can familiarity preferences explain apparent calculation by human babies? *Developmental Psychology, 42,* 666–678. http://dx.doi.org/10.1037/0012-1649.42.4.666

Moore, D. S., & Johnson, S. P. (2008). Mental rotation in human infants: A sex difference. *Psychological Science, 19,* 1063–1066. http://dx.doi.org/10.1111/j.1467-9280.2008.02200.x

Moore, D. S., & Johnson, S. P. (2011). Mental rotation of dynamic, three-dimensional stimuli by 3-month-old infants. *Infancy, 16,* 435–445. http://dx.doi.org/10.1111/j.1532-7078.2010.00058.x

Myers, D. G., & Dewall, N. (2016). Developing through the lifespan. *Exploring psychology* (pp. 119–170). New York, NY: Worth.

Reed, R. W., Florax, R. J., & Poot, J. (2015). *A Monte Carlo analysis of alternative meta-analysis estimators in the presence of publication bias*. Economics Discussion Papers, No 2015–9, Kiel Institute for the World Economy. Retrieved from http://www.economicsejournal.org/economics/discussionpapers/2015-9

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86,* 638–641. http://dx.doi.org/10.1037/0033-2909.86.3.638

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17,* 551–566. http://dx.doi.org/10.1037/a0029487

Schimmack, U. (2016). *A revised introduction to the R-Index*. Retrieved from https://replicationindex.wordpress.com/2016/01/31/a-revised-introduction-to-the-r-index/

Schöner, G., & Thelen, E. (2006). Using dynamic field theory to rethink infant habituation. *Psychological Review, 113,* 273–299. http://dx.doi.org/10.1037/0033-295X.113.2.273

*Simon, T. J., Hespos, S. J., & Rochat, P. (1995). Do infants understand simple arithmetic? A replication of Wynn (1992). *Cognitive Development, 10,* 253–269. http://dx.doi.org/10.1016/0885-2014(95)90011-X

Sirois, S., & Mareschal, D. (2004). An interacting systems model of infant habituation. *Journal of Cognitive Neuroscience, 16,* 1352–1362. http://dx.doi.org/10.1162/0898929042304778

*Slater, A. M., Bremner, J. G., Johnson, S. P., & Hayes, R. A. (2010). The role of perceptual and cognitive processes in addition-subtraction studies with 5-month-old infants. *Infant Behavior and Development, 33,* 685–688. http://dx.doi.org/10.1016/j.infbeh.2010.09.004

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods, 5,* 60–78. http://dx.doi.org/10.1002/jrsm.1095

Starkey, P., & Cooper, R. G., Jr. (1980). Perception of numbers by human infants. *Science, 210,* 1033–1035. http://dx.doi.org/10.1126/science.7434014

Starkey, P., Spelke, E. S., & Gelman, R. (1983). Detection of intermodal numerical correspondences by human infants. *Science, 222,* 179–181. http://dx.doi.org/10.1126/science.6623069

Starkey, P., Spelke, E. S., & Gelman, R. (1990). Numerical abstraction by human infants. *Cognition, 36,* 97–127. http://dx.doi.org/10.1016/0010-0277(90)90001-Z

Stevens, J. R., & Taylor, A. M. (2009). Hierarchical dependence in meta-analysis. *Journal of Educational and Behavioral Statistics, 34,* 46–73. http://dx.doi.org/10.3102/1076998607309080

Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in Stata and SPSS. *Research Synthesis Methods, 5,* 13–30. http://dx.doi.org/10.1002/jrsm.1091

*Uller, C., Carey, S., Huntley-Fenner, G., & Klatt, L. (1999). What representations might underlie infant numerical knowledge? *Cognitive Development, 14,* 1–36. http://dx.doi.org/10.1016/S0885-2014(99)80016-1

*Wakeley, A., Rivera, S., & Langer, J. (2000). Can young infants add and subtract? *Child Development, 71,* 1525–1534. http://dx.doi.org/10.1111/1467-8624.00244

*Wynn, K. (1992). Addition and subtraction by human infants. *Nature, 358,* 749–750. http://dx.doi.org/10.1038/358749a0

Xu, F., & Arriaga, R. I. (2007). Number discrimination in 10-month-old infants. *British Journal of Developmental Psychology, 25,* 103–108. http://dx.doi.org/10.1348/026151005X90704

Received October 2, 2015
Revision received January 30, 2017
Accepted March 9, 2017 ■