Theory of Mind and Prosocial Behavior in Childhood: A Meta-Analytic Review

Kana Imuta, Julie D. Henry, and Virginia Slaughter
University of Queensland

Bilge Selcuk
Koc University

Ted Ruffman
University of Otago

It has been argued that children who possess an advanced theory of mind (ToM) are more likely to act prosocially, yet the empirical findings are mixed. To address this issue definitively, a meta-analytic integration of all prior literature that met appropriate inclusion criteria was conducted. In total, 76 studies including 6,432 children between 2 and 12 years of age contributed to these analyses.Collapsed across all studies, a significant association emerged ($r = .19$), indicating that children with higher ToM scores also received higher scores on concurrent measures of prosocial behavior. The magnitude of this effect was similar across ToM assessments requiring identification of others’ cognitions versus emotions, and it existed irrespective of whether the ToM measure imposed demands on false belief reasoning or not. The association with ToM was also evident for different subtypes of prosocial behavior (helping, cooperating, comforting). ToM had a similar effect for boys and girls, but was slightly stronger in children aged 6 years or older, relative to their younger peers. Taken together, these findings provide the strongest evidence to date that being able to explicitly consider what other people are thinking and feeling is related to children’s tendencies to act prosocially, although the magnitude of the association is relatively weak.

*Keywords:* theory of mind, prosocial behavior, perspective taking, children, meta-analysis

*Supplemental materials:* http://dx.doi.org/10.1037/dev0000140.supp

In developmental psychology, the construct of theory of mind (ToM) grew out of the Piagetian notion of perspective taking, which is the ability to take another person’s point of view. This involves identifying what others see (Salatas & Flavell, 1976), think (Ruby & Decety, 2003), and feel (Lamm, Bateson, & Decety, 2007)—visual, cognitive, and affective perspective-taking (VPT, CPT, APT), respectively. ToM has classically been defined as involving insight into other people’s minds and reasoning about how mental states influence behavior (e.g., Perner, 1991; Wellman, 1990). More recently, ToM has been defined broadly as involving general social insights without a commitment as to whether the tasks used to measure ToM really do tap mental state understanding (e.g., Perner, 2010; Ruffman, 2014). In the present review, we adhere to this second formulation.

Our aim was to determine whether children’s ToM performance relates to their prosocial behavior. Many have claimed that the social insights indexed by success on ToM tasks will profoundly affect children’s interpersonal behavior (e.g., Hughes & Cutting, 1999; Wellman, 1990). However, empirical support for this idea, and for a link between ToM and prosocial behavior in particular, has not yet accumulated in a consistent or convincing manner. Therefore, in the current study, we conduct the largest meta-analysis to date, to comprehensively evaluate the association between performance on various ToM tasks and prosocial behavior in childhood.

Prosocial behavior involves any form of voluntary action that is performed to benefit another individual (Eisenberg, 2003). The first signs of prosociality emerge shortly after children’s first birthdays, when they begin to help others complete simple goal-directed behaviors (Warneken & Tomasello, 2006, 2007). Then, over the course of early childhood, children’s repertoire of prosocial behaviors gradually expands and increases in frequency to include sharing to meet others’ unmet material desires, comforting another person’s emotional distress, and cooperating to attain mutual goals (Dunfield & Kuhlmeier, 2013; Eisenberg & Fabes, 1998; Paal & Bereczkei, 2007). Recognition of others’ needs, desires, feelings, and intentions should facilitate children’s engagement in these various prosocial behaviors (Dunfield, 2014; Hay & Cook, 2007; Hoffman, 2000). Furthermore, engaging in prosocial behaviors may, in turn, lead to the development of a more nuanced ToM (Astington, 2003; Eisenberg & Fabes, 1998). That is, by carrying out (or failing to carry out) prosocial behaviors, children experience and learn about the emotions and
thoughts that arise in response to these situations (e.g., Sy, De-Meis, & Scheinfield, 2003; Weller & Lagattuta, 2014). Thus, the reciprocal link between prosocial behavior and ToM development seems theoretically compelling; however, when researchers have directly assessed this association in children, the results are surprisingly inconsistent. Some studies indicate strong positive concurrent associations, yet others report null or even negative associations (see Figure 1).

In an attempt to bring order to this literature, two meta-analyses assessing the relation between ToM and prosociality have been conducted. In the first, Underwood and Moore (1982) analyzed mean correlations between prosocial behavior and perspective-taking (including VPT, CPT, and APT) in samples of 3- to 13-year-old children. Small but significant associations emerged between prosocial behavior and each type of perspective taking ($r_s = .25$ and .28 in each category). However, the number of contributing studies here was small, with 4, 10, and 2 studies assessing VPT, CPT, and APT, respectively.

A second larger meta-analysis (Carlo, Knight, McGinley, Goodvin, & Roesch, 2010) was then conducted, which included 47 studies. Of these, 12 involved teenagers, college students, or adults, and the remainder ($k = 35$) examined children aged 12 years or younger. Carlo et al. (2010) tested whether greater similarity between the social insight and prosocial behavior assessed would lead to a stronger correlation between the two measures. For instance, compared with VPT and comforting, the relation between APT and comforting might be larger because both involve emotion. Across all studies, they found an overall weighted mean correlation between perspective taking and prosocial behavior of .16, which increased to around .40 to .50 when similarity ratings were high. However, these effects were moderated by the age group studied. Specifically, the overall mean association was close to zero ($r_s = .01$) among 3- to 5-year-olds and participants over 18 years of age, indicating that the significant overall association was being driven by studies that included children in the 6- to 17-year-old range. This finding appears to be at odds with Underwood and Moore’s (1982) study, which identified a significant association among 3- to 13-year-olds. It is also puzzling from a theoretical perspective in that ToM development is particularly pronounced in the preschool period (Wellman, Cross, & Watson, 2001), and it has been widely assumed that this transition should be specifically associated with developments in social functioning, including prosocial behavior (e.g., Eisenberg, Fabes, & Spinrad, 2006; Hoffman, 2000).

These previous meta-analyses provide initial insights into the relationship between ToM and prosocial behavior in childhood but the picture is still not entirely clear and many important questions remain. Therefore, we conducted the most comprehensive meta-analytic evaluation of this association to date. Relative to prior meta-analyses, we drew together a larger set of studies ($K = 76$) and included a wider range of ToM measures (traditional perspective taking as well as more recent ToM tasks). In addition, we included only studies that assessed children (2 to 12 years), in order to quantify the extent to which ToM is related to prosocial behavior during a period of substantial developmental change for both constructs. Furthermore, with such a large number of studies, it was possible to look at the degree to which any significant association varied as a function of several important variables (see below for more details). Finally, we separately evaluated the strength of the associations between language ability and each of our two key constructs to begin to consider the role of this important, but often ignored, third variable.

**Type of ToM Task**

Following on from previous meta-analyses, we distinguished between two broad categories of ToM tasks that assessed either children’s identification of others’ (a) emotions (APT); or (b) cognitions (CPT; e.g., desire, intention, knowledge, etc.). Within the CPT category, we made a finer distinction, examining tasks that involved reasoning about another’s false belief (FB) and predicting subsequent behavior (i.e., first-order FB). Understanding FB is often regarded as a hallmark measure of ToM because it requires children to predict another’s action/belief independent of their own belief (e.g., Astington, 2003; Denett, 1978; Perner, 1991). For this reason, many have hypothesized that FB tasks are particularly good indicators of the ability to consider another’s perspective and as such, may be specifically associated with the production of prosocial behavior (Cassidy, Stetson Werner, Rourke, & Zubernis, 2003; Lalonde & Chandler, 1995; Moore, Barresi, & Thompson, 1998; Slomkowski & Dunn, 1996). Conversely, one could argue that FB taps insight into cognitive mental states (as opposed to affective states, which might be regarded as more central to prosocial behavior), so that it might be less strongly related to prosocial behavior compared with APT. Our dataset, which included a large number of FB tasks, provides the first opportunity to examine whether FB is specifically related to prosociality in children.

Our comprehensive meta-analysis also included some ToM tasks that could not be classified as APT or CPT (e.g., interpretive understanding, tactile perspective taking, VPT). These measures are labeled in the Appendix (see Appendix in online supplemental materials) and included in the overall analysis, but were not included in subanalyses of the ToM measures.

**Type of Prosocial Behavior**

There have been numerous theoretical and empirical attempts to identify different types of prosocial behavior. For this meta-

---

**Figure 1.** Funnel-plot for the overall analysis providing a plot of standard error on the vertical axis as a function of effect size (expressed as Fisher’s Z) on the horizontal axis. As recommended by Sterne, Becker, and Egger (2005), guidelines were superimposed on these funnel plot to show the expected distribution of studies in the absence of bias, and were also used to help to identify outliers. The vertical line in the middle of the distribution denotes the average, aggregate mean.
analysis, we adopted Dunfield and Kuhlmeier’s (2013) classification of prosocial behaviors: responses to instrumental needs (helping), emotional distress (comforting), and material desires (sharing). We also created a separate category of prosocial tasks involving cooperation; whereas helping tasks required children to assist on behalf of a target person, usually the experimenter, cooperation tasks measured children’s willingness to assist in tandem with a target person, usually another child.

**Motivation for Prosocial Behavior**

Some tests prompt children with gestures, verbal hints, or even direct instructions to perform a prosocial act, whereas others give the opportunity for spontaneous prosociality. We distinguished compliant (prompted) versus spontaneous prosociality because prompting may cue the child to engage in a socially normative response that they would not have generated of their own accord. Compliant prosocial behavior, therefore, may reflect different cognitive underpinnings (e.g., social scripts) or other elements of ToM (e.g., communicative competence) as opposed to children’s inclination to act for the benefit of others, specifically.

**Age**

Another variable that may systematically influence the ToM-prosocial relation is age. ToM becomes increasingly sophisticated between the ages of 2 and 12 years (Peterson, Wellman, & Slaughter, 2012), with a significant developmental shift in the preschool period (Wellman et al., 2001). If ToM is related to prosocial behavior, then the preschool period in particular might see an advance in prosocial behavior, consistent with Hoffman’s (2000) suggestion that ToM equips children to engage in prosocial acts. However, there is mixed evidence as to whether prosocial behavior is stable from the school years or increases. For instance, one study indicates an increase over both the preschool and school years (Eisenberg & Fabes, 1998), whereas another found stability of prosocial behaviors from kindergarten (roughly 5 years) to Grade 6 (roughly 11 years) in a large sample of American children (Nagin, 2005). For these reasons, we separately analyzed data from samples of preschoolers (age 2 to 5) versus older children.

**Gender**

None of the previous meta-analyses included gender as a potential moderator, yet girls tend to be more prosocial than boys (Eisenberg & Fabes, 1998). Furthermore, girls tend to do better than boys on ToM tasks, although the advantage is slight (Chorman, Ruffman, & Clements, 2002). These findings suggest that gender might be a moderator of the ToM-prosocial behavior relation, so we evaluated findings separately for boys and girls.

**Language**

Children’s language ability is a consistent, positive correlate of performance on ToM tasks (Milligan, Astington, & Dack, 2007), as well as their prosocial behavior (e.g., Durkin & Conti-Ramsden, 2007). Therefore, it is important to consider language ability when examining the ToM-prosocial behavior relation, although most individual studies have not reported relations with language. We addressed this issue by calculating mean effect sizes for associations between language ability, which included assessments of receptive or productive vocabulary or tests of grammar, and the two key target variables of interest.

**Overview of Research Questions**

In sum, this meta-analysis seeks to answer the following questions:

1. Is there an association between children’s performance on different types of ToM tasks and their tendency to engage in prosocial behavior? This is an important, unresolved question. As noted, prior research provides inconsistent results about the nature and magnitude of any association between these constructs.

2. Is any observed association with prosocial behavior stronger for different ToM tasks, namely, APT versus CPT, or FB reasoning in particular? Given the importance attributed to FB reasoning measures in the ToM literature, this is a critical question, and one that has not yet been addressed at the level of meta-analysis.

3. Is any observed association stronger for different types of prosocial behaviors or for differently motivated prosocial behavior (compliant vs. spontaneous)?

4. Is the strength of any observed association influenced by children’s age? Carlo et al.’s (2010) findings suggest that an association will not be present in preschoolers, yet it is widely proposed that ToM developments during the preschool period should be related to increases in prosocial tendencies.

5. Does the strength of any observed association vary by children’s gender? Girls do somewhat better on ToM tasks and tend to be more prosocial, suggesting links might be stronger in girls.

6. Are relations between ToM and prosocial behavior similar to relations between language and prosocial behavior?

**Materials and Method**

**Literature Search**

A systematic search of the existing literature on the relation between children’s ToM and prosocial behavior was completed in July 2014. Identification of studies eligible for inclusion was achieved by searching the Web of Science (Thomson Reuters), *Psych INFO* (American Psychological Association), *Scopus* (Science), and *Google-Scholar* (Google) databases. The keywords searched were: “prosocial behavior,” “caring,” “comforting,” “cooperating,” “donating,” “helping,” “sharing,” “supporting,” in combination with, “social understanding,” “perspective taking,” “theory of mind,” “mindreading,” “mentalizing,” “false belief,”
Inclusion Criteria

Studies included in the meta-analysis met the following criteria:

1. Published in English or Turkish (i.e., languages in which members of our team were fluent).
2. Research design included healthy preschool or elementary school-age children. Healthy was defined as the absence of psychiatric or neurological illness or severe sensory impairment (e.g., profound deafness).
3. Included a behavioral measure of ToM with a primary focus on mental state attribution and/or distinguishing between two different mental perspectives. Tasks assessing empathy (vicarious experience of another’s emotional state) and/or sympathy (sadness in reaction to another’s misfortune) were not included.
4. The studies had to include a measure of prosocial behavior, assessed concurrently with children’s ToM. Children’s prosocial behavior was assessed through laboratory-based behavioral tasks, observations, and questionnaires. Specifically, we included measures of children’s comforting, cooperating, helping, and sharing. Measures of prosocial behavior were categorized and analyzed separately, as (a) produced in response to a prompt to act prosocially (i.e., compliant); or (b) produced without a prompt (i.e., spontaneous).
5. Included studies had to report precise statistics convertible to effect sizes. Studies that reported correlations between the two key dependent measures were only considered to be eligible where raw correlations were reported (i.e., partial correlations did not contribute). For studies published within the last decade that did not present all necessary statistics, authors were contacted directly to request the relevant data. Lastly, effect sizes were derived from $M$s and $SD$s provided where studies divided children into those who had high versus low ToM, such that the data did not include continuous ToM scores that could be correlated with children’s prosocial behavior scores.

In addition to the 76 studies that were included in the present meta-analysis, the initial search revealed a further 51 studies. Of these, 30 were eliminated because they did not supply appropriate data and were published more than a decade ago. Another 16 studies were eliminated because the prosociality measure did not assess a specific prosocial behavior (for instance, it tapped into broader social competence, social skill, etc.) or because they were self-report measures of prosocial behavior based on hypothetical situations. A further four studies were eliminated because they did not meet the criterion for ToM assessment, and one because the study focused on deaf children.

Sample Characteristics

The 6,432 children who participated in the 76 studies included in our meta-analysis were from 12 different countries: Australia, Canada, China, Colombia, Germany, Israel, Italy, the Netherlands, Switzerland, Turkey, the United Kingdom, and the United States. In all but seven studies (where the majority were African American, Hispanic, Asian, and a variety of Jewish family backgrounds), the majority of children were Caucasian. Across all 76 studies, the primary language for the majority of children was English, but for children in 13 of the studies, their primary languages were Arabic/Hebrew, Chinese, Dutch, German, Italian, Spanish, and Turkish. The majority of children were recruited from day care centers, kindergarten, preschool, and primary schools in the cities in which data was collected. Children were from working-, middle-, and upper-class families, with the majority from middle-class backgrounds.

Statistical Analysis

Meta-analysis is a rigorous, quantitative alternative to the traditional review process, as it involves statistical integration of results. The basis of this methodology is the effect size, a standardized statistic that quantifies the magnitude of an effect. Although two mathematically equivalent types of metric exist for quantifying effect size (often referred to as the $r$ and $d$ families), as a consequence of its greater generality of interpretation, consistency of meaning and more salient practical meaning, it has been argued that $r$ is the more useful effect size estimate (Rosenthal & DiMatteo, 2001). Consequently, in the present study, $r$ was used for statistical analyses. This corresponds to the degree of correlation between the two variables of interest (i.e., ToM and prosocial behavior).

For each construct, effects were pooled to derive an estimate of the mean, with each effect weighted for sample size to correct for sampling error. To do so, the random effects model was selected in preference to the more commonly used fixed effects model as it yields more generalizable parameter estimates (Shadish & Haddock, 1994). The mean in the fixed effects model is presumed to reflect a common underlying effect parameter that gives rise to the sample observations. However, in the random effects model the mean represents a hyperparameter, as it allows for substantive differences beyond sampling error that differentiate the effects contributing to each respective mean. Statistically, the crucial difference between these methodologies is in the calculation of standard errors and confidence intervals, which for the random effects model are typically larger. For this reason, the National Research Council (1992) suggests that the latter model is preferable, as fixed effects models may lead to inappropriately strong conclusions.

The first stage in combining effect sizes involves estimating $Q$, a statistic used to test the homogeneity of effects that have been pooled (i.e., the extent to which the effects contributing to each respective mean can be regarded as measuring the same underlying parameter). When $Q$ exceeds the upper-tail critical value of chi-square at $k-1$ df ($k =$ number of effects pooled), the variance is significantly greater than would be expected if the effects in question shared a common population effect size. In such circumstances, assumptions underpinning use of random (but not fixed) effects analyses would be met (Shadish & Haddock, 1994). How-
ever, even if $Q$ does not prove significant in a particular analysis, there is typically a good deal of meaningful covariation between study characteristics and findings. Thus, estimates of $Q$ were derived to provide an estimate of the degree of variability in the effects contributing to each particular mean, but were not used as a decision rule to guide choice of analytical model. As a heuristic for interpretative purposes, where $k$ is held constant, larger values of $Q$ are indicative of greater between-groups variability.

To interpret how important a particular effect was in practical terms, Cohen’s (1977) guidelines were used (small = .10, medium = .30, large = .50). Squares of the effect size multiplied by 100 were also presented as these latter quantities represent the percentage of the variance shared by the two dependent measures (i.e., ToM and prosocial behavior). In addition, for each mean effect size, the combined $Z$ was calculated, which indicates the significance (in the present study, two-tailed) of the mean effect size estimate.

An important consideration in integrating this literature was that almost all eligible datasets contributed multiple dependent measures, often for the same construct. Importantly, for each mean effect size (reported in Table 1), each participant dataset was only permitted to contribute once. This strategy was achieved by calculating the mean of the relevant effects for the dataset in question, and entering this single value. However, the same participants were allowed to contribute to different mean effects. Since these mean effects are not independent (i.e., the same participants contribute to both), comparing them statistically is not recommended (Hedges & Olkin, 1985). Allowing each sample to only contribute to one dependent variable (to avoid violating the independence assumption) would in principle bypass this problem, but it would do so at the cost of inflating the Type II error rate to an unacceptable level (given the relatively small number of studies contributing to several of the contrasts). Consequently, we refer primarily to differences in effect size magnitude, as well as in the percentage of the variance accounted for ($\Delta PVAF$) when interpreting results.

Finally, a number of validity threats have been identified that may lead to imprecise conclusions in both nonquantitative and meta-analytic reviews. Particularly problematic is “the file drawer problem” (Rosenthal, 1979), which refers to the fact that significant results are more likely to be published than nonsignificant results (Easterbrook, Gopalan, Berlin, & Matthews, 1991; Ferguson & Heene, 2012; Sterling, 1959). A number of procedures were used to address this potential source of bias. In line with recommendations by Borenstein, Hedges, Higgins, and Rothstein (2005),

### Table 1

<table>
<thead>
<tr>
<th>Contrast</th>
<th>Mean $r$</th>
<th>95% CI</th>
<th>PVAF</th>
<th>$K$</th>
<th>$N$</th>
<th>$Q$</th>
<th>Orwin Failsafe $N$</th>
<th>Begg’s method</th>
<th>Egger’s method</th>
<th>Trim and fill imputed $r$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prosocial-ToM</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All studies</td>
<td>.19</td>
<td>.15 .23</td>
<td>3.6</td>
<td>76</td>
<td>6,432</td>
<td>205.1*</td>
<td>.151</td>
<td>1.52*</td>
<td>.12*</td>
<td></td>
</tr>
<tr>
<td><strong>By measure:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>APT</td>
<td>.24</td>
<td>.15 .33</td>
<td>5.8</td>
<td>27</td>
<td>1,994</td>
<td>102.0*</td>
<td>32</td>
<td>.32*</td>
<td>2.23</td>
<td>.24*</td>
</tr>
<tr>
<td>CPT</td>
<td>.16</td>
<td>.12 .20</td>
<td>2.6</td>
<td>57</td>
<td>4,899</td>
<td>111.4*</td>
<td>23</td>
<td>.12</td>
<td>1.39*</td>
<td>.10*</td>
</tr>
<tr>
<td><strong>By FB demands:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FB task</td>
<td>.25</td>
<td>.19 .30</td>
<td>6.3</td>
<td>12</td>
<td>1,041</td>
<td>8.86</td>
<td>19</td>
<td>-.15</td>
<td>-.10 .25*</td>
<td></td>
</tr>
<tr>
<td>Non-FB task</td>
<td>.21</td>
<td>.15 .26</td>
<td>4.4</td>
<td>55</td>
<td>4,219</td>
<td>186.5*</td>
<td>42</td>
<td>.22*</td>
<td>2.05* .11</td>
<td></td>
</tr>
<tr>
<td><strong>By behavior:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comforting</td>
<td>.17</td>
<td>.08 .26</td>
<td>2.9</td>
<td>13</td>
<td>835</td>
<td>20.2</td>
<td>8</td>
<td>-.01</td>
<td>.95</td>
<td>.16*</td>
</tr>
<tr>
<td>Cooperating</td>
<td>.24</td>
<td>.10 .37</td>
<td>5.8</td>
<td>15</td>
<td>921</td>
<td>59.6</td>
<td>22</td>
<td>.15</td>
<td>-.12</td>
<td>.24*</td>
</tr>
<tr>
<td>Helping</td>
<td>.17</td>
<td>.10 .24</td>
<td>2.9</td>
<td>26</td>
<td>1,904</td>
<td>58.7</td>
<td>14</td>
<td>.09</td>
<td>1.36</td>
<td>.16*</td>
</tr>
<tr>
<td>Sharing</td>
<td>.11</td>
<td>.05 .18</td>
<td>1.2</td>
<td>33</td>
<td>2,478</td>
<td>75.1*</td>
<td>0</td>
<td>.25*</td>
<td>1.57</td>
<td>.02</td>
</tr>
<tr>
<td><strong>By motivation:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spontaneous</td>
<td>.19</td>
<td>.12 .26</td>
<td>3.6</td>
<td>39</td>
<td>2,991</td>
<td>128.7*</td>
<td>28</td>
<td>.21*</td>
<td>1.50</td>
<td>.19*</td>
</tr>
<tr>
<td>Compliant</td>
<td>.10</td>
<td>-.08 .27</td>
<td>1.0</td>
<td>7</td>
<td>587</td>
<td>23.9*</td>
<td>1</td>
<td>.10</td>
<td>-.86</td>
<td>.06</td>
</tr>
<tr>
<td><strong>By age group:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age 5 and under</td>
<td>.17</td>
<td>.12 .21</td>
<td>2.9</td>
<td>35</td>
<td>2,757</td>
<td>44.4</td>
<td>23</td>
<td>.05</td>
<td>.74</td>
<td>.14*</td>
</tr>
<tr>
<td>Age 6 plus</td>
<td>.24</td>
<td>.14 .33</td>
<td>5.8</td>
<td>26</td>
<td>1,803</td>
<td>96.5*</td>
<td>32</td>
<td>.17</td>
<td>.61</td>
<td>.24*</td>
</tr>
<tr>
<td><strong>By gender:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boys</td>
<td>.12</td>
<td>.02 .21</td>
<td>1.4</td>
<td>13</td>
<td>817</td>
<td>18.9</td>
<td>1</td>
<td>.26</td>
<td>1.23</td>
<td>.12*</td>
</tr>
<tr>
<td>Girls</td>
<td>.23</td>
<td>.07 .38</td>
<td>5.3</td>
<td>12</td>
<td>641</td>
<td>42.1*</td>
<td>11</td>
<td>.26</td>
<td>2.52</td>
<td>.24*</td>
</tr>
<tr>
<td><strong>ToM and language:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All studies</td>
<td>.33</td>
<td>.25 .40</td>
<td>10.9</td>
<td>17</td>
<td>2,023</td>
<td>54.6*</td>
<td>35</td>
<td>.32</td>
<td>1.84</td>
<td>.27*</td>
</tr>
<tr>
<td><strong>Prosocial-language</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All studies</td>
<td>.13</td>
<td>.09 .17</td>
<td>1.7</td>
<td>20</td>
<td>2,351</td>
<td>15.5</td>
<td>6</td>
<td>-.04</td>
<td>.22</td>
<td>.13*</td>
</tr>
</tbody>
</table>

Note. A positive effect size indicates a positive association between the two constructs of interest, a negative effect size the reverse. $PVAF$ refers to percentage of variance shared between the two dependent measures (for instance, prosocial behavior and theory of mind understanding). $K$ refers to the number of studies and $N$ refers to the number of participants that contributed to the effect. Failsafe $N$ refers to the number of missing studies with null results that would be required to nullify the significance of each computed mean effect.

Trim and fill: (look for missing studies to left of mean; using random effects model. Imputed mean is random effects. The mean effects for the association between ToM with language, and between prosocial behavior and language are also reported.

$p < .05$. 

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
we proceeded through a logical sequence in which we first obtained a visual sense of the data by constructing funnel plot diagrams for all aggregate mean effect sizes (see Figures 1 and 2). These provide a plot of standard error on the vertical axis as a function of effect size (expressed as Fisher’s Z) on the horizontal axis. As recommended by Sterne et al. (2005), guidelines were superimposed on these funnel plots to show the expected distribution of studies in the absence of bias (see, e.g., Egger, Davey Smith, Schneider, & Minder, 1997; Sterne & Egger, 2001).

Next, we used rank correlation and regression procedures to establish whether there is any evidence of bias. First, the Begg and Mazumdar rank correlation test was used. Here, the correlation between study size and effect size is calculated; a significant correlation suggests that bias may exist, as it implies that small studies are more likely to be included when they show a relatively large effect size. Egger’s test of the intercept provides a test of the same potential bias, but here uses precision (the inverse of the standard error) to predict the standardized mean effect size. As with the rank correlation test, a significant intercept indicates that bias may be present.

In addition to testing for the possible presence of bias in the dataset, we were also interested in judging the robustness of each mean effect size to any potential bias, should it exist. For these analyses, we first calculated Orwin’s “Fail safe N.” This statistic quantifies the number of filed studies averaging null results \( (r = 0) \) that must exist to reduce the mean correlation past a specified threshold (in the present study, a value of .1, i.e., a small effect). In addition, we used trim-and-fill analyses to establish how each respective mean effect size would change if any identified bias were to be removed. Thus, for each aggregate mean, we also report an imputed point estimate (for more details on each of these procedures, see Borenstein, Hedges, Higgins, & Rothstein, 2005) and if this differs from the weighted mean \( r \) then we interpret the imputed mean \( r \). All meta-analytic calculations were conducted using Version 2.0 of the software, Comprehensive Meta-Analysis (Borenstein et al., 2005).

**Results**

A total of 76 published and unpublished studies conducted between 1973 and 2014 met the inclusion criteria. These studies are indicated by an asterisk in the reference section. They included a total of 6,432 children aged between 2 and 12 years.

The Appendix (see Appendix in online supplemental materials) presents a study-by-study breakdown of effect sizes for the association between the measure of ToM and prosocial behavior. The first stage in the analysis of these data involved calculating an overall mean weighted effect size (with a single effect size permitted to contribute for each independent dataset, and these effects in turn weighted for sample size). A positive effect indicates that relatively advanced ToM was associated with higher levels of prosocial behavior; a negative effect indicates the reverse. As can be seen in Table 1, the overall mean effect size was significant, but small in magnitude \( (r = .19) \). Thus, collapsed across all studies, children’s performance on ToM tasks was concurrently associated with their prosocial behavior, but this association was a relatively weak one. Although Egger’s test of the intercept identified significant potential publication bias in the studies contributing to this mean effect, the trim and fill imputed mean effect remained significant, although was smaller \( (r = .12) \).

**Analyses by Type of ToM Measure**

**Cognitive versus affective perspective taking.** The first subanalysis focused on the type of ToM measure. Here, we separately calculated mean associations between APT and prosocial behavior \( (k = 27) \), and between CPT and prosocial behavior \( (k = 57) \); see Table 1. The mean effect sizes for both types of ToM measure were significant, but the mean effect size for the APT-prosocial association was greater \( (r = .24) \) than that for CPT and prosocial behavior \( (r = .16) \). However, the \( Q \)-statistics indicated that both of these effects were significantly heterogeneous, and for both, significant potential bias was identified (for the former using Begg’s method, and for the latter, Egger’s method). When imputed means were estimated that took into account identified bias, the APT-prosocial association was unchanged \( (r = .24) \), but the CPT-prosocial association was still significant, although reduced \( (r = .10) \). These data therefore indicate that an association exists between both APT and CPT assessments with prosocial assessments, but the magnitude of this association is greater when the ToM task measures an understanding of affect.

**False belief versus other ToM tasks.** Next, we focused on whether performance on FB tasks, compared with all other types of ToM tasks, is significantly associated with prosocial behavior \( (k = 12 \) for FB tasks and 55 for other ToM tasks). Thus, in this analysis, APT tasks were combined with CPT tasks that did not measure FB to create a non-FB category. The mean effect sizes for both ToM categories were statistically significant \( (rs = .25 \) for FB and .21 for non-FB tasks). The imputed mean was unchanged from the calculated mean for the FB tasks, but a significant potential publication bias was identified for the mean effect for non-FB tasks and the imputed mean effect was substantially reduced \( (r = .11) \), although still remained statistically significant. Overall, the data indicate the following: (a) APT tasks on their own are a good correlate of prosocial behavior, (b) CPT tasks that do not measure FB are a less significant correlate of prosocial behavior, and (c) FB understanding on its own is significantly associated with prosocial behavior.

**Analyses by Type of Prosocial Behavior: Comforting, Cooperating, Helping, Sharing**

The next subanalyses evaluated associations between children’s ToM and different expressions of prosociality. The mean effect sizes between ToM and comforting, cooperating, helping, and sharing \( (k = 13, 15, 26, \) and 33, respectively) were all statistically significant, and ranged from .11 (for sharing) to .24 (for cooperating). No significant potential publication bias was identified for comforting, cooperating, and helping, and imputed means for these three behaviors were very similar to the calculated means \( (rs = .16, .24, \) and .16, respectively), and all remained statistically significant. However, significant potential bias was identified for sharing behavior using Begg’s method, and the imputed mean was substantially smaller, and no longer statistically significant \( (r = .02) \). These data indicate that, with the exception of sharing behavior, children’s ToM is consistently associated with different types of prosocial behavior.
Figure 2. Funnel-plots for each of the subanalyses providing a plot of standard error on the vertical axis as a function of effect size (expressed as Fisher’s $Z$) on the horizontal axis. As recommended by Sterne et al. (2005), guidelines were superimposed on these funnel plot to show the expected distribution of studies in the absence of bias, and were also used to help to identify outliers. The vertical lines in the middle of the distributions denote the average, aggregate means.
Analyses by Motivation: Spontaneous Versus Compliant

Next, we ran a subanalysis to evaluate whether the association between children’s ToM and their prosocial behavior varied according to the way in which the prosocial behavior was motivated. The mean effect size for ToM and spontaneous prosociality was significant \( r = .19; k = 39 \) but this was not the case for compliant prosociality \( r = .10; k = 7 \). Although significant potential publication bias was identified for spontaneous social behaviors using Begg’s method, the imputed mean effect was unchanged in magnitude, and remained significant.

Analyses by Age Group: Preschoolers Versus Older Children

The next subanalysis focused on age group, specifically preschool-aged cohorts between 2 and 5 years of age versus cohorts aged 6 and older. As shown in Table 1, the mean effect sizes for both age groups were significant. The association for younger children \( r = .17; k = 35 \) was smaller than that for older children \( r = .24; k = 26 \). No significant potential publication bias was identified for either age group, and the imputed mean was slightly reduced for the younger age group \( r = .14 \), but remained unchanged for the older age group, with both imputed means remaining statistically significant. Overall, then, the positive association between ToM and prosocial behavior is significant in both preschool and school-age children, but is slightly stronger in the older cohort.

Analyses by Gender

To examine gender, analyses were rerun, but this time, only with studies that reported effect sizes separately for males and females. Although effect sizes for both genders were significant, the magnitude of the effect for females was substantially larger than that for males \( r_s = .23 \) vs. \( .12 \), respectively; \( k_s = 12 \) and 13, respectively). No significant publication bias was identified for either of these mean effects, and the imputed mean was unchanged for males \( r = .12 \), and slightly increased for females \( r = .24 \). These data indicate that although a significant relationship between ToM and prosocial behavior is evident for both groups, this relationship is stronger for females.

Associations With Language

Within the sample of studies included in this meta-analysis, 17 reported associations between children’s language ability and their ToM scores, and 20 reported associations between language and prosocial behavior. Aggregating over these studies, we found a mean ToM-language correlation of .33 \( k = 17 \), which is similar in magnitude to that reported in Milligan, Astington, and Dack’s (2007) meta-analysis. We also found a mean prosocial-language correlation of .13 \( k = 20 \), which is similar in magnitude to the association identified between ToM and prosocial behavior. For neither of these mean effects was significant potential publication bias identified, so that imputed means were very similar to calculated means \( r_s = .27 \) and .13, respectively), with both remaining significant.

Discussion

It has long been argued that children’s ToM should relate to real world functioning, and in particular, their propensity to engage in prosocial actions. Reflecting on the importance attributed to this idea, a large literature has emerged, focused on the association between children’s ToM and their prosocial behavior; the current meta-analysis examined 76 such studies. The first major finding to emerge was that ToM was significantly correlated with prosociality in children aged 2 to 12 years, with the overall small but significant mean effect of .19 largely commensurate with the overall results from prior meta-analyses (Carlo et al., 2010; Underwood & Moore, 1982). Nevertheless, a key strength of the current meta-analysis was the large number of contributing studies, which allowed us to complete a number of novel corollary meta-analyses, the results of which substantially enrich our understanding of how these two constructs are related in childhood.

In the first of these analyses, it was shown that two broadly distinguished aspects of ToM, namely APT and CPT, were differentially related to prosocial behavior, with APT a stronger correlate. This finding is surprising in some ways, since both APT and CPT fundamentally require recognition of another individual’s perspective. Yet in other ways it is not surprising since some types of APT tasks may access “hot” ToM (as opposed to “cold” ToM; see Davis & Stone, 2003), by inviting children both to identify a protagonist’s affective state and to vicariously experience it. This type of task may bring a motivational component that could increase the tendency to act prosocially (cf. Carlo et al., 2010). By contrast, the social insights that are tapped in CPT tasks are likely to be important in identifying situations that require prosocial assistance, but this does not necessarily prompt a motivation to act prosocially (Dunfield, 2014). The results of our analysis differ from Underwood and Moore’s (1982) meta-analytic findings that APT and CPT were similarly related to prosocial behavior; this may be explained by the starkly smaller number of studies that contributed to their subanalyses \( k = 2 \) APT and \( k = 10 \) CPT for Underwood and Moore vs. \( k = 27 \) APT and \( k = 57 \) CPT in the present study).

We also examined the subset of CPT tasks that tested FB reasoning versus the ToM tasks that did not (including both APT and CPT tasks). FB tasks have been regarded as a hallmark measure of ToM, based on the fact that these tasks are structured such that children must put aside their own perspective in order to pass (Wimmer & Perner, 1983). Many have argued that, as such, performance on FB tasks should be a particularly powerful indicator of children’s interpersonal sensitivity and therefore should relate to their social functioning. After imputing mean effects that removed any identified publication bias, there was some evidence that the association with prosocial behavior was substantially stronger for FB measures compared to all other types of ToM tasks: 6.3% of the variance in prosocial behavior was accounted for by FB tasks versus 1.2% for the combination of APT and non-FB ToM tasks. This result reinforces the long-held suspicion that FB understanding is relevant for real world social functioning (Hughes & Leekam, 2004; Lalonde & Chandler, 1995). Likewise, our analyses showed that APT tasks, on their own, are related to prosocial behavior, and CPT tasks that do not involve FB (e.g., those that measure knowledge understanding) are less strong predictors of prosocial behavior.
Recent theories of prosocial behavior development suggest that different types of prosocial behaviors may rely upon different social–cognitive insights (Dunfield, 2014). To address this issue, we investigated whether the four different types of prosocial behaviors examined in the present study (comforting, cooperating, helping, sharing) related differently to children’s ToM. Small but significant associations were identified across the different types of prosocial tasks, with the exception of sharing behaviors for which there was evidence of publication bias and a nonsignificant imputed mean effect. On the face of it, sharing would seem to require an understanding of others’ mental states; however, sharing tends to be a social convention that young children have many opportunities to acquire because (a) they are often concerned with an accumulation of external resources so that adults must frequently emphasize the necessity to share, and (b) sharing is a relatively straightforward concept to convey because it applies routinely whenever salient, observable resources are available and relies on the equal distribution of such resources. Thus, the weak association between ToM and sharing likely stems from the routinely enforced parental instruction that children should share, thereby eliminating the need for social insight to act as a trigger for sharing behavior.

Helping, comforting, and cooperating were all significantly associated with children’s ToM, with cooperation showing the largest mean effect of .24. Overall these associations support the proposal that acting prosocially is underpinned by a developing sensitivity to others’ perspectives (Dunfield, 2014). The uniquely strong association between ToM and cooperation may relate to the fact that, unlike other forms of prosocial behavior, cooperation requires ongoing, continuous interpretation of socioemotional signals and adjustment of one’s own desires to maintain. Furthermore, cooperation may provide particularly salient interpersonal feedback that facilitates the development of children’s ToM, because it necessarily involves sharing emotional states and outcomes with another person (Johnson, 1975b).

Our analysis indicated that ToM was significantly correlated with spontaneous but not compliant prosocial behaviors. This is not surprising, as most likely the “compliant” prosocial tasks in which children are explicitly prompted to comfort, cooperate, help, or share, simply tap their propensity to follow rules rather than their ability to independently identify situations calling for prosocial behavior and, in turn, to motivate themselves to act.

Our initial prediction with respect to age was that associations may be stronger in the younger cohort because there is substantial development in ToM during the preschool years (Wellman et al., 2001). If ToM enables prosocial behavior, then one might expect the association to increase as children’s ToM becomes more sophisticated. In contrast to this prediction, there was some indication that the link between ToM and prosocial behavior was stronger for 6- to 12-year-old children than for preschool children, as estimated by both actual and imputed mean effects.

On the one hand, our finding of a significant association between ToM and prosocial behavior in the youngest group ($r = .17$) contrasts with Carlo et al.’s (2010) study, which identified no significant association between these constructs in preschoolers. The explanation for this difference is not obvious, although Carlo et al.’s (2010) study included only five preschool studies, whereas ours included 35. With 35 studies, our findings stand to be less susceptible to null results obtained through chance. On the other hand, the general pattern of our findings was similar to Carlo et al. (2010), in that both studies found stronger relations in older compared with younger children. One possibility for this pattern of findings may be that it reflects the influence of method variance. Specifically, ToM is operationalized differently at different developmental periods, with the tasks given in middle childhood tending to be more varied and tapping a wider range of social insights compared to those typically administered in early childhood. Another possibility is that there is simply more error variance when tasks are given to young children because their wandering attention leads to somewhat random performance on tasks compared with older children.

It is possible, however, that the age-related pattern of findings is not artifactual. ToM may play less of a role in younger children’s prosociality because at least some early prosocial acts may be motivated by normative ideas of how one should respond when a person is in need (Paulus, 2014), or through the mere desire to have a social interaction with other people (Paulus & Moore, 2012). With age, children become more selective about when and how to behave prosocially (Hay & Cook, 2007), by considering the perspective of the other individual in need and weighing it against that of their own (Sierksma, Thijs, Verkuyten, & Komter, 2014). Additionally, the stronger relation found in older children in comparison to younger children may indirectly support the idea that engaging in prosocial behaviors helps foster children’s ToM. That is, children who are more inclined to act prosocially gain more opportunities to experience and witness the emotional benefits of these acts for both themselves and others. This outcome will be experienced as rewarding, and for young children, reward tends to be more motivating than punishment (van Duijvenvoorde, Zanoli, Rombouts, Rijmakers, & Crone, 2008).

Finally, the results from our gender subanalyses indicated that the ToM-prosocial association differed substantially for girls and boys, with this relation stronger for girls. This finding makes sense in light of evidence that girls tend to outperform boys on both measures (Charman et al., 2002; Eisenberg & Fabes, 1998) and given that girls’ peer relationships tend to be more intimate and more communicatively complex day-to-day (Banerjee, Rieffe, Terwogt, Gerlein, & Voutsina, 2006). This would suggest that in childhood, girls may be more highly tuned to others’ perspectives as well as more practiced in producing prosocial acts. Our finding of a gender effect in the present meta-analysis is therefore consistent with this broader literature. Nevertheless, it is important to note that, although smaller, a significant ToM-prosocial association was evident for boys as well as girls. This underscores the overall suggestion from this meta-analysis, that the ToM–prosociality correlation emerges in early childhood and maintains across diverse expressions of ToM and of prosociality, as well as distinct interpersonal norms characteristic of boys and girls in the preschool and early school years.

Other Influences on Prosocial Behavior

Over all analyses, the correlation between ToM and prosocial behavior indicates a consistent, though relatively modest association. There are two ways of interpreting this pattern. On the one hand, correlations between ToM and prosocial behavior may underrepresent the true relation outside of the lab (see Carlo et al., 2010; Eisenberg & Miller, 1987 for detailed arguments). On the
other hand, as mentioned above, ToM may be a necessary but not a sufficient condition for prosocial behavior (Astington, 2003). Specifically, motivation, over and above understanding, must play a substantial role in whether one acts prosocially (Carlo, Knight, Eisenberg, & Rotenberg, 1991). Motivation to act prosocially is affected by a range of variables, including parenting behavior and sibling influences (Bryant & Crockenberg, 1980), social exclusion (Twenge, Baumeister, DeWall, Ciarocco, & Bartels, 2007), religiosity (Shariff & Norenzayan, 2007), empathy (Eisenberg & Miller, 1987), and positive mood (George, 1991), in addition to the other factors we have already discussed above (e.g., social conventions, reciprocity). Such findings suggest two possible conclusions, both of which potentially have validity: (a) ToM enables but does not necessitate prosocial behavior, and (b) ToM is one contributor to prosocial behavior among many in children, but is likely not the most important contributor. Further research that separately assesses ToM and children’s motivations to behave prosocially, as well as other cognitive and situational variables, will be necessary to empirically test these two conclusions.

With respect to other potential variables that might influence prosociality in childhood, the current analyses reinforce the need to consider language ability. Specifically, separate meta-analyses of the association between language ability and both ToM and prosocial behavior revealed significant associations. Language ability was not considered in prior meta-analyses but our finding was that greater language capacity was related to higher ToM scores, and also to an increased likelihood of engaging in prosocial behavior. It is possible, therefore, that language development drives the associations that we found between ToM and prosocial behavior. It was unfortunately not possible to control for language in the meta-analysis because of the relatively low number of studies reporting the necessary statistics (the precise association between ToM and prosocial behavior controlling only for language, and no other variable). This does, however, point to the need for future research focused on the relation between ToM and prosocial behavior to carefully take into consideration any potential mediating role of language. For instance, language fluency might facilitate children’s engagement in social interactions, and this social experience might then facilitate both their ToM and their understanding of when to help another person.

Another limiting factor in interpreting our findings is that the vast majority of studies have been cross-sectional, and therefore, do not speak to whether ToM plays a causal role in facilitating prosocial behavior. As discussed previously, it is possible that a relatively advanced ToM could lead to more prosocial behavior, by enabling children to accurately identify another’s mindset and enabling children to accurately identify another’s mindset and reason about engaging in prosocial behaviors (Hoffman, 2000). Alternatively, it might be that producing prosocial behavior leads to rewarding social experiences, which could in turn facilitate children’s ToM. Once again, the relation could also be mediated by a third variable, such as temperament or language. Further studies, ideally with longitudinal assessments of ToM and prosociality, will help to gain a more complete understanding of the relation between these constructs. If ToM is causal, then one might anticipate that early ToM correlates with later prosociality, independent of early prosociality (i.e., using a cross-lagged design), whereas there would be no relation between early prosociality and later ToM.

The few studies that have examined longitudinal relations have sometimes found no relation between early ToM and later prosocial behavior (Ruffman, Slade, Devitt, & Crowe, 2006), and sometimes found that early ToM is related to later prosocial behavior (e.g., Broeren, Muris, Diamantopoulou, & Baker, 2013; Caputi, Lecce, Pagnin, & Banerjee, 2012; Eggum et al., 2011). However, these studies have generally failed to examine bidirectional relations in a cross-lagged study with measures of ToM and prosocial behavior at both time points, a methodology essential to examining causality. Intervention studies also have rich promise in advancing this literature, in which it could be tested whether children who are provided with training in ToM are more likely to subsequently engage in prosocial behavior.

Finally, it is important to note that although in the present study a number of methodological approaches were used to try and identify and correct for publication bias, new and ever more sophisticated publication bias methodologies continue to be developed and refined (e.g., Simonsohn, Nelson, & Simmons, 2014; Simonsohn, Simmons, & Nelson, 2015). Given that the present meta-analysis identified significant sources of potential bias for many of the calculated mean effects, use of these approaches should be considered for future meta-analytic reviews on related topics.

To conclude, this study reports the largest meta-analysis of the relation between ToM and prosocial behavior in children aged between 2 and 12 years. The results show that ToM is significantly related to children’s prosocial behavior, although the effect is small, accounting for approximately 3.6% of the variance overall. However, this small effect is robust, because it generalizes across all types of ToM task, and three types of prosocial behavior. Furthermore, several potentially important differences emerged with respect to the magnitude of the observed effect, for instance, with suggestions of a stronger relation for ToM measures that imposed demands on FB understanding, for children aged 6 years or older, and for girls. In sum, this comprehensive meta-analysis moves the field substantially forward by providing the clearest picture to date of the relation between ToM and prosocial behavior during this critical developmental period. It also points to important avenues for future investigation to more clearly understand the nature of the link.

References

Asterisks indicate the 76 studies that were included in the meta-analysis.


Received March 19, 2015
Revision received January 12, 2016
Accepted April 7, 2016