Rapid Automatized Naming (RAN) as a Kindergarten Predictor of Future Reading in English: A Systematic Review and Meta-analysis

Sean McWeeny
Soujin Choi
June Choe
Roxelyn and Richard Pepper Department of Communication Sciences and Disorders, Northwestern University, Evanston, Illinois, USA

Alexander LaTourrette
Department of Psychology, Northwestern University, Evanston, Illinois, USA

Megan Y. Roberts
Elizabeth S. Norton
Roxelyn and Richard Pepper Department of Communication Sciences and Disorders, Northwestern University, Evanston, Illinois, USA

Department of Medical Social Sciences and Institute for Innovations in Developmental Sciences, Feinberg School of Medicine, Northwestern University, Evanston, Illinois, USA

ABSTRACT

Rapid automatized naming (RAN) has been shown to be a strong correlate of reading abilities. RAN also predicts future reading across different ages, ability levels, and languages, and is often used in literacy screening. Thus, understanding the specific relations between early RAN and later reading difficulties is important, particularly for screening. This systematic review and meta-analysis (with N = 60 samples; k = 373 effect sizes; n = 10,513 participants), was the first to test the extent to which measures of RAN assessed before grade school predict future reading performance in English-speaking children. We also tested whether characteristics of the RAN tasks, reading measures, or sample demographics moderate this relationship. We found that overall, kindergarten/preschool RAN is correlated with grade-school reading at $r = -.38$, similar in magnitude to previous concurrent meta-analyses that included various ages and languages. We found that alphanumeric RAN tasks were more strongly related to future reading than were non-alphanumeric tasks, as well as that RAN significantly predicts all types of reading measures tested, but more strongly predicts real word than nonword reading. To assess the role of RAN’s unique predictive power, we also meta-analyzed the semipartial correlations of early RAN with later reading when controlling for phonological awareness (PA); the result of $r_{sp} = -.25$ demonstrates RAN’s significant, unique contribution beyond PA. These results support shared cognitive resource models in which the similarity between RAN and reading tasks accounts for their correlation. We provide practical guidelines for based on these data for early screening for reading difficulties and dyslexia.

Introduction

Reading is a complex process that requires the automatic integration of multiple cognitive and linguistic abilities. Reading-related skills such as rapid automatized naming (RAN), phonological awareness, and letter knowledge can all be measured at the pre-reading stage and predict later reading ability (Byrne et al., 1997; Pennington and Lefly, 2001; Scarborough, 1998; Schatschneider et al., 2004). However, it is currently a major challenge to accurately identify reading difficulties early in reading development, when intervention is likely more effective (Al Otaiba et al., 2014; Blachman et al., 2014; Cavanaugh et al., 2004; Lovett et al., 2017; Torgesen, 2004; Vellutino et al., 1998). Optimizing screening batteries that allow early identification of reading problems at the outset of schooling, and therefore earlier intervention, is critical to optimizing long-term outcomes for children with reading difficulties (Connor et al., 2014).

Numerous studies have examined pre-school and kindergarten-age predictors of later reading ability and how various factors can modify the
relationship between predictors and reading outcomes (e.g., Hjetland et al., 2017). Across studies, the measures that are most commonly identified as strong predictors of later reading in English include phonological awareness (PA), RAN, letter name and sound knowledge, and language ability (for reviews, see National Early Literacy Panel, 2008; Ozernov-Palchik and Gaab, 2016). Though RAN shares some processes with these other predictors, it has consistently been shown to uniquely relate to reading, beyond the contribution of phonological awareness (Kirby et al., 2003; Manis et al., 2000; Wolf and Bowers, 1999), and beyond similar measures of general processing speed and single (discrete) item naming (Altani et al., 2020; Logan et al., 2011). Whereas some measures such as letter knowledge are only predictive of reading for a short interval until they are mastered (Paris, 2005), RAN retains its concurrent and predictive relation with reading over time (Wagner et al., 1997). Further, early RAN predicts reading over long time intervals, at least a decade into the future (Adlof et al., 2010; Mazzocco and Grimm, 2013). Importantly, the RAN-reading relationship persists across varying ages, reading abilities and alphabetic and non-alphabetic languages and orthographies of varying depth (Araújo et al., 2015; Araújo and Faisca, 2019; Caravolas et al., 2019; Furnes and Samuelsson, 2011).

Gaining a nuanced understanding of the relation between RAN and reading ability is important for two major reasons: informing educational/clinical practice and informing theory. In terms of informing practice, understanding the circumstances under which RAN best predicts later reading is crucial for screening and early identification of reading difficulties. For example, little is known about when the optimal time is to screen and whether the exact type of RAN test matters (in terms of number of items, type of items, use of raw or standardized score, and more). Identifying children with reading difficulties as early as possible, when intervention is more effective, would mitigate the compounding negative consequences that poor readers face under the predominant “wait to fail” model, such as reduced educational attainment, poorer socio-emotional well-being, and higher rates of entry into the juvenile justice system (Humphrey and Mullins, 2002; Richardson and Wydell, 2003; Svensson et al., 2001; Torgesen, 2004).

Understanding the nature of the RAN-reading relationship also informs understanding of the nature of reading ability and development as well as theory related to reading. Multiple-deficit models, pioneered by Wolf & Bowers’ (1999) Double Deficit Hypothesis, consider naming speed to be one causal factor in reading ability (Menghini et al., 2010; Pennington, 2006; Pennington et al., 2012). However, in other prominent accounts such as the Simple View of Reading (Gough and Tunmer, 1986), the constructs of speed and automaticity as measured by RAN are considered to play a minor role at best (as part of the decoding component, Johnston and Kirby, 2006). Another longstanding question in the field is how unique RAN is as a predictor, and its relationship to phonological processing (a construct that includes PA; e.g., Wagner et al., 1994; Wagner et al., 1997). Many individual studies find that RAN is a unique predictor of reading, distinct from or beyond the contributions of phonological and letter knowledge or orthographic measures (Landerl et al., 2019; Norton and Wolf, 2012), and that they have distinct neural correlates (Norton et al., 2014, 2021). However, no meta-analysis to date has directly tested RAN’s unique contribution above and beyond other pre-reading measures. Understanding the relationship between RAN, reading, and other pre-reading variables is thus key to clarifying RAN’s role in reading development.

Defining RAN Tasks

RAN is measured by the time it takes a child to name an array of familiar items, such as objects, colors, numbers, or letters (Denckla and Rudel, 1976; Norton and Wolf, 2012), reflecting the automaticity of the multiple processes that are involved in this process (Wolf et al., 2000). There are several important parameters that define a true RAN task. First, the items to be named must be highly familiar or automatized. For example, when children are typically still learning their letters in kindergarten, the RAN letters task may not relate closely to reading because the naming is not automatized. However, once children have learned the names of letters and numbers with automaticity, these alphanumeric RAN tasks are completed faster than non-alphanumeric tasks (such as objects or colors) and are more strongly related to reading (Cardoso-Martins and Pennington, 2004; Schatschneider et al., 2004; Torgesen et al., 1997). Second, the items must be arranged in an array or grid and named in the left-to-right, row-by-row fashion that is analogous to reading in English. (In rare cases, the items can be named top-to-bottom in columns, e.g., Van den Bos et al., 2002). Naming items that are presented one at a time in a speeded manner (discrete naming) is not the same as the serial process of a true RAN task (Altani et al., 2020; de Jong, 2011; Logan et al., 2011; Protopapas et al., 2013), even though some studies call this “discrete RAN.” Third, the RAN measure is usually based on time to complete the task. Some studies use the number of items/second or seconds/item (e.g., Schatschneider et al., 2004). Errors and self-corrections are not typically used in calculating a RAN score, but they may increase the time to name the array and thus be reflected in the naming time. Other factors can be calculated from a RAN task, such as pause time or change row-by-row (Amtmann et al., 2007; Georgiou et al., 2006; Georgiou et al., 2008a, 2008b), but these are less widely used in practice.
Theories of Mechanisms Underlying the RAN-Reading Relationship

Many potential explanations for why RAN relates so strongly to reading have been posited, including their shared processes of global processing speed (e.g., Kail and Hall, 1994), phonological processing (e.g., Wagner et al., 1997), serial visual processing and orthographic access (Sunseth and Bowers, 2002), and articulation (Papadopoulos et al., 2016; Wolf et al., 2000). These variables, along with many other shared cognitive processes, change over the course of development, and therefore the model explaining the relationship between RAN and reading must account for this. For example, as children gain accuracy and automaticity in reading, RAN speed becomes more strongly correlated with reading speed (Juul et al., 2014). This relationship varies depending on orthographic transparency, with accuracy measures plateauing much earlier in transparent than opaque orthographies (Seymour et al., 2003).

No matter how dynamic and multi-faceted the model between RAN and reading can be, there are specifications of how variables such as processing speed, serial processing, and articulation may relate to RAN and reading. Path models have been extensively tested, with each study finding slightly different model specifications (Cutting and Denckla, 2001; Georgiou et al., 2016; Papadopoulos et al., 2016). For example, the relationships among general processing speed, RAN, phonological processing, and orthographic processing change based on whether the orthographic processing measures are speeded or not (Georgiou et al., 2016). Another key specification is that the RAN-reading relationship is driven by not only serial processing or left-to-right eye movements (Protopapas et al., 2013), but cascading processing (i.e., processing multiple items simultaneously in overlapping fashion and effectively looking ahead at items to be named next; Gordon and Hoedemaker, 2016; Nayar et al., 2018). RAN may also have a unique relationship with oral reading fluency as opposed to silent word reading fluency (i.e., word-chains), suggesting that articulation plays an important role in the relationship between RAN and oral reading fluency (Georgiou et al., 2013; Papadopoulos et al., 2016). Though these studies were in Greek, it may hold that these models would replicate in English, as RAN shows similar patterns of relation with reading across languages (Araújo et al., 2015) and is considered more general to cognition than specific to a given language (Papadopoulos et al., 2016).

Ultimately, most current models suggest that RAN and reading are related because they share multiple underlying linguistic and non-linguistic cognitive processes (Georgiou and Parrila, 2020; Norton and Wolf, 2012; Wolf et al., 2000). The paths of these models may be “common cause” with RAN and reading both directly affected by processes like working memory, or through mediation, in which RAN ability may affect reading indirectly through improved orthographic processing or phonological awareness (Papadopoulos et al., 2016). Thus, within an individual, a profile of strengths and weaknesses of underlying cognitive processes will affect both RAN, reading, and other mediating variables to account for their relationship. Although the exact role of some processes such as articulation is debated (Cutting and Denckla, 2001; Georgiou and Parrila, 2020; Lervåg and Hulme, 2009), it is agreed that multiple shared neural and cognitive processes underlie both RAN and reading (as demonstrated with fMRI; Cummine et al., 2015).

Insights on how RAN Relates to Reading from Meta-Analyses

Previous meta-analyses have documented the significant correlation between RAN and reading across various reading constructs and languages. In the first published meta-analysis of RAN and reading, Swanson et al. (2003) found a strong concurrent relationship between RAN and single word reading \( r = -0.41 \), when looking across a range of ages, reading abilities, and languages\(^1\). Two subsequent meta-analyses have found a similar magnitude of relationship between RAN and reading, while providing new contributions. Araújo et al. (2015) found the overall concurrent RAN-reading relationship across languages to be \( r = -0.43 \), with slightly higher correlations in opaque orthographies like English. Their analyses included substantially more studies, and thus provided greater statistical power than earlier work by Swanson and colleagues. In turn, Hjetland et al. (2017) found the longitudinal correlation from early RAN to later reading to range from \( r = -0.34 \) to \( -0.37 \), depending on the reading measures used. Thus, they demonstrated that longitudinal correlations with RAN have similar effect sizes to concurrent correlations.

Differences in RAN ability have also been identified in two meta-analyses of children with reading difficulties. In a meta-analysis of various cognitive and reading-related skills, Kudo et al. (2015) found that the effect size difference for RAN in children without versus with reading difficulties was \( d = 0.89 \) (equivalent to \( r = 0.41 \)), however only 10 samples were included in that analysis. In a much larger meta-analysis with 216 effect sizes analyzed, Araújo and Faisca (2019) documented an even larger RAN deficit in individuals with dyslexia (\( d = 1.19 \), equivalent to \( r = 0.51 \)). These documented RAN deficits in children with reading difficulties/dyslexia support its use as an early screener.

In addition to demonstrating consistent correlations between RAN and reading, these meta-analyses also demonstrated that various factors (i.e., moderators), such as the type of stimuli used, the orthographic depth of the language studied, and the type of reading measure, affect the strength of the RAN-reading correlation. Swanson et al.
Motivations and Goals for the Current Study

The purpose of this meta-analysis is to assess the longitudinal relationship from RAN measured in kindergarten or preschool to later reading abilities in English. Measuring the longitudinal relationship, as opposed to the concurrent relationship, is essential not only for investigating RAN’s utility as an early screener for reading difficulties, but also essential for understanding the changing relationship between RAN and reading as reading transitions from a focus on accuracy to efficiency (Seymour et al., 2003). We consider a variety of reading constructs, including measures of nonword decoding (i.e., reading nonsense words like “sorpi”), sight word reading (i.e., reading single words that can be recognized without decoding), reading comprehension (i.e., reading paragraphs or sentences and being able to answer questions about the writing’s content) and reading fluency (i.e., reading sentences or paragraphs aloud as accurately and quickly as possible). This work thus extends a previous meta-analysis (Hjetland et al., 2017) to include articles that use all reading constructs rather than only reading comprehension as an outcome. We also directly test early RAN’s unique contribution to later reading, above and beyond the contribution of PA. PA and RAN share considerable variance and interest in parsing their respective effects has only grown since the formulation of the double deficit hypothesis (Norton and Wolf, 2012). This question serves practical and theoretical purposes in understanding how much RAN contributes to our understanding of early reading development. Finally, we perform extensive forward and backward snowball searching, as more papers were available to include beyond those identified in the Hjetland et al. (2017) dataset.

Practical Motivations

The key considerations for this design, including its focus on work in English-speaking children, early measures of RAN, and longitudinal relationships, are driven by a goal for this meta-analysis to inform specific policy recommendations for educators and administrators. It is clear that state- and local-level policymakers are looking for ways to best implement RAN in screening, as evidenced by the creation of measures such as the Arkansas Rapid Naming Screener and its use by other states (Arkansas Department of Education, 2017). As in previous meta-analyses examining the concurrent RAN-reading relationship, we also test several potential moderators, which address key practical questions. Practical questions, such as “how many items should a RAN task include?” and “at what age should I evaluate RAN?”, may help educators and clinicians choose effective screening measures. Policymakers are also interested in RAN’s unique contribution to predicting reading outcomes, which is why we have considered it alongside PA measures.

Theoretical Motivations

Most meta-analyses of RAN focus on documenting the relationship between RAN and reading while generally not trying to explain why RAN and reading are related. Here, we will test several questions related to why RAN and reading are correlated. Theoretical questions, such as “do timed reading measures more strongly relate to RAN than untimed reading measures?” and “do nonword decoding tasks relate less strongly to RAN than sight word tasks?” may help researchers further converge on theory for why RAN and reading relate.

Summary

Because RAN ability develops considerably during the school-age years (Denckla and Rudel, 1976; Georgiou...
et al., 2006), its relationship to later reading ability may be different than the concurrent relations between RAN and reading at older ages. However, if early RAN reliably predicts later reading, it further increases the motivation to include RAN in kindergarten or preschool literacy screening. However, there is a lack of understanding of the theoretical and practical questions about how early RAN task performance relates to later reading abilities. As such, quantifying the average relationship between early RAN and later reading is the primary research question in this meta-analysis. Secondary questions are whether factors related to the RAN task, reading measure, or child participant sample, moderate the RAN-reading relationship. These specific questions and their rationale are explained in depth, and specific analyses are proposed in the Method section.

**Method**

This study followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moher et al., 2009). Data collection and extraction processes are described in text and in Figure 1. The PRISMA checklist is provided as Supplemental Material. Our data, protocols, processing and analyses scripts, and other related documents are available via Open Science Framework: https://osf.io/awpqk/. This meta-analysis was considered exempt by the Institutional Review Board at Northwestern University.

**Study Inclusion Criteria**

For the present study, we focused on articles in which English was the primary language of the participants, as
consistency of orthography can moderate the RAN-reading relationship (Araújo et al., 2015) and the largest number of published studies are in English. We acknowledge that English is not a representative orthography (Share, 2008; Share, 2021), but that this analysis serves as a starting point and allows specific conclusions to be drawn in at least this one language. As we were interested in early predictors of reading ability, we only included articles in which the initial timepoint with RAN assessment was in (the US equivalent of) kindergarten or preschool (the earliest stage at which RAN can be measured reliably) and reading was subsequently measured at some point in Grades 1–5. Thus, we only included studies that spanned at least one school year. For studies that only reported the sample’s age rather than grade, we included the study if the mean age was ≤78 months (age 6.5 years, or the middle of first grade in the US). Studies with children who spoke other languages were excluded; however, studies with bilingual children were included if (a) the language of instruction was English and (b) the children were described as fluent in English. All eligibility criteria can be found in Table 1. Examples of specific decisions regarding inclusion/exclusion can be found in the Supplemental Materials.

### Data Collection

On September 26, 2019, we identified possible sources through full-text database searches of EBSCO (PsychINFO, PsycARTICLES, and ERIC) and PubMed. We used the search terms: (reading OR dyslexia) AND (“rapid naming” OR “naming speed” OR “rapid automat* naming” OR “RAN” OR “rapid serial naming”) AND (“preschool*” OR “kindergart*” OR “pre- school*” OR “pre k*” OR “pre- k*” OR “prek*” OR “child*”), see Table 1. This search returned 4497 titles, 4088 of which were unique. We re-ran this search on November 8, 2021, to include articles published since September 2019. Figure 1 shows the number of articles at each stage.

### Abstract and Title Screening

As a first step, one of two authors reviewed the title of each article from the database search; titles that were deemed to be clearly irrelevant were screened out. This title screening step resulted in 2098 potentially relevant articles with abstracts to be screened. These abstracts were then each reviewed by two different screeners. Three individuals contributed to abstract screening and consensus was reached in all cases of conflict. Abstract screening for full-text inclusion agreement was 85% and all disagreements were resolved with consensus of three coders. And 437 of these articles were deemed relevant and were then full-text screened. Seven trained coders screened full texts for inclusion, with 89% agreement and resolution of all
disagreements. From these, 94 articles met the eligibility criteria. After contacting authors to obtain some that were not included in articles, 52 had relevant effect sizes. These articles were then each coded for various measures of interest twice, by two of five trained coders. There was 94% agreement across all variables and any disagreements were reviewed by the first and second author and resolved through consensus.

**Snowball Search**

After the database search and screening, a snowball search was conducted using references and citations of the 52 included studies with relevant effect sizes. For this snowball search, we used Microsoft Academic Graph (Wang et al., 2019), which is a database that tracks connections between published papers, such that every backward reference is also a forward citation, similar to Web of Science. All articles that were identified by the snowball search were title and abstract screened using the same processes as those described above. Snowball searching returned 43 articles that met the eligibility criteria. And 15 of these studies had relevant effect sizes (after contacting authors) and added 10 unique samples. The search also returned 28 studies without relevant effect sizes, 14 of which were related to samples already contained in the corpus.

**Contacting Authors for Additional Information/Data**

Authors from either the database search or snowball search whose paper had no relevant effect sizes (e.g., because of reporting regressions or grouped analyses rather than correlations) were contacted via email to request raw data or correlation matrices so that the information could be included in the current analysis. For the papers where this was the case, 9 authors responded to our request, providing data on 10 unique samples.

**Data Extraction**

Data for this study were collected and managed using REDCap (Research Electronic Data Capture) tools hosted at Northwestern University (Harris et al., 2009, 2019). REDCap is a secure, web-based software platform designed to support data capture for research studies.

Data (including relevant information on the sample, tasks, and Pearson correlations) from each paper/sample were entered in REDCap by two independent coders, and consensus was reached in case of any discrepancy. For longitudinal studies that measured RAN and/or reading at multiple timepoints, we extracted only one kindergarten/preschool time point and only one grade school timepoint. This design consideration intentionally minimizes variance, as our primary question is focused on the utility of RAN as an early screener. However, a side effect of this approach is that it limits the variability that can be explained by age of testing. Timing of initial and follow-up assessments were coded in terms of the sample’s grade, as papers predominantly reported grade rather than age. Exceptions and further details are listed in Supplemental Materials, and a full list of sources included in the meta-analysis is also available in Supplemental Materials.

**Effect Size Extraction**

The scoring of the RAN task affected whether the Pearson correlation with reading would be positive or negative. If a raw score (i.e., time) or rate (time/item) was used, the correlation was entered as negative. If a standard score or rate (item/time) was used, this value was multiplied by −1. There were a few exceptions to this rule, in which a reading measure was either based on time or rate (e.g., Wolf et al., 1986) or expressed as a chronological age lag (Heath and Hogben, 2004). In addition, there were several ambiguous cases that were carefully considered, see details in Supplemental Materials.

Many studies assess RAN as part of a large battery of reading-related measures that potentially predict later reading. Due to the many constructs measured in these large and longitudinal studies, many researchers created latent RAN or reading measures through factor or principal components analysis (Dally, 2006; Macdonald et al., 2013). We decided to extract these correlations between one or two latent variables as they qualify as Pearson correlations, and later test whether including them would change our results.

**RAN Measure Categories**

The stimuli used in a RAN task are typically restricted to one of five types: colors, objects, letters, digits, or occasionally animals. Even more rarely, studies have used colored animals (e.g., Catts et al., 1999). The ‘colored animals’ task (e.g., naming "blue cow," "red dog," etc.) is included here as a RAN task, but not compared with other stimulus types in moderator analyses due to the very few studies that employed it. We also excluded tasks with multiple stimulus types in the array, such as letters and numbers, in order to focus on the classic RAN task. Previous meta-analyses have found that the relationship with reading is stronger between alphanumeric (i.e., letters or numbers) than non-alphanumeric stimuli (such as colors or objects; Araújo et al., 2015). However, this was assessed concurrently, whereas different results may be seen with early RAN predicting later reading. Further, many children do not know their letters accurately or automatically in kindergarten or preschool, making a RAN letters task inappropriate for these younger children. Thus, in the current study we quantified each RAN task’s relationship with later reading and whether alphanumeric RAN tasks are a stronger predictor of later reading than non-alphanumeric RAN.
Reading Measure Categories

Here, we operationalized three primary types of reading measures: reading fluency, reading comprehension, and single word reading measures. Fluency measures had to measure either a rate or total number of words read correctly in a pre-determined time limit in connected text (sentences or passages). This definition differs from fluency measures in Araújo et al. (2015), who used “items per second” as a measure of fluency. Single word reading included real and nonword reading tasks and was further broken down into single word efficiency (i.e., timed single word and nonword reading) and single word accuracy (i.e., untimed single word and nonword reading) measures. The full categorization of each reading measure is located in files available on the Open Science Framework site for this project.

Previous meta-analysis of children of all ages indicates that RAN is associated with single word reading accuracy (i.e., word ID) at \( r = -0.41 \) and reading comprehension at \( r = -0.45 \) (Swanson et al., 2003). Hjetland et al. (2017) found mean effect sizes of \( r = -0.37 \) for word reading and \( r = -0.34 \) for reading comprehension with earlier RAN measures. However, the specific correlations between RAN and reading vary considerably between and within studies. For example, in one study (Cronin and Carver, 1998), kindergarten RAN scores related to Grade 1 Word ID scores at \( r = -0.37 \) to \(-0.60\), depending on the RAN task, and to passage comprehension at \( r = -0.31 \) to \(-0.57\). Thus, we quantified RAN’s relationship with 3 primary types of reading: fluency, comprehension, and single word reading. Single word reading was further analyzed as accuracy versus efficiency measures.

Timed Measures

Because RAN is a speeded task, it is typically more closely related to timed or speeded reading measures (Savage and Frederickson, 2005; Schatschneider et al., 2004). This is evident in studies of older students; for example, RAN speed in grade 3 significantly predicted performance on a timed single word reading task in grades 3, 4, and 5, but did not reliably predict untimed single word reading (Georgiou et al., 2009). Further, one theoretical account posits that processes underlying RAN constrain the development of reading fluency (Lervåg and Hulme, 2009). Thus, we quantified RAN’s relationship with timed and untimed reading measures.

Nonword Reading

Nonword reading task have extra phonological demands that sight words do not. Previous meta-analyses (Araújo et al., 2015) found a weaker correlation between nonword reading and RAN than real word reading and RAN. This difference may exist because nonword reading is much less automatic than real word reading, even early in reading development. Therefore, we quantified RAN’s relationship with real word reading and nonword reading, with the prediction that the relationship between RAN and nonword measures would be weaker than RAN and real word reading.

Participant Characteristics

Reading Ability

Among older students, there is mixed evidence regarding whether RAN is a stronger correlate or predictor of reading ability among children who are poor readers than typical or skilled readers. Some studies find a stronger concurrent RAN-reading relation in poor readers (Araújo et al., 2011; Bowers et al., 1988; Felton and Brown, 1990; McBride-Chang and Manis, 1996). One study found that RAN in 3rd grade significantly predicted later single word reading in 8th grade among poor readers, but that there was no such significant relation in good readers (Meyer et al., 1998). On the other hand, meta-analyses of concurrent RAN-reading relations in older children reveal that the correlation between RAN and reading is similar in samples of typical readers and poor readers; Swanson et al. (2003) found correlations of \( r = -0.41 \) for typical readers and \(-0.43\) for poor readers, and Araújo et al. (2015) found no significant differences in the magnitude of the concurrent relations between RAN and reading whether the sample of readers was poor/impaired \( (r = -0.49\), \text{typical/average} \( (r = -0.45\), or unselected \( (r = -0.43\). It is not known whether these differences across studies are due to a restricted range or “ceiling” effect in RAN among good readers with greater variability among poor readers (McBride-Chang and Manis, 1996) or whether differential relations truly exist in good versus poor readers.

Due to the focus here on young children, we are not able to examine the full range of reading ability and how it may correlate with RAN. We can probe whether children at risk for dyslexia may have a different RAN-reading relationship than peers without risk for dyslexia. Children with familial risk for dyslexia tend to have poorer RAN skills than their peers (Pennington and Lefly, 2001; van Bergen et al., 2012), yet not all children with familial risk or poor RAN scores go on to be poor readers. Some studies find a weaker RAN-reading relationship in those at risk for dyslexia; for example, Heath & Hogben (2004) found that pre-kindergarten RAN correlated with Grade 2 Word ID at \( r = -0.03 \) for children with poor PA skills, compared with \( r = -0.38 \) for children with good PA skills. Other studies find quite similar effect sizes across risk status; for example, Hulme et al. (2015) found children with versus without risk for dyslexia had correlations between kindergarten RAN Objects and Grade 3 reading of \( r = -0.21 \) and \( r = -0.22 \), respectively. Here, we used a three-tier classification system of risk: low, medium, and high-risk. Any sample from the general population or an explicitly low-risk group was considered low-risk (e.g., Cardoso-Martins and Pennington,
A medium risk sample was one where the study oversampled for dyslexia risk using family history and/or poor performance on pre-reading measures, but still included many low-risk participants (e.g., Ozernov-Palchik et al., 2017). High-risk samples were explicitly stated as such, categorized using family history and pre-reading measure performance, and were often analyzed as subgroups in studies (e.g., Cardoso-Martins and Pennington, 2004). Thus, we tested whether early RAN is a better predictor in samples of primarily typically developing children as opposed to samples with larger proportions of children identified as at-risk for reading difficulties.

**Practical Considerations**

**RAN Task Publication, Standardization and Test Length**
There are a number of published, standardized and normed RAN measures that are used widely, including the Comprehensive Test of Phonological Processing (CTOPP and CTOPP-2; Wagner et al., 2013) and the RAN/RAS Tests (Wolf and Denckla, 2005), among others. However, many studies use researcher-created RAN tasks that have not necessarily been standardized or normed. Among these tests, the format of the RAN task, including how many different unique items (types) and total number of items included (tokens), also varies. A previous meta-analysis found no moderating effect for the total number of items in a RAN task on concurrent relations with reading (Araújo et al., 2015). Thus, we tested whether using a published, standardized measure influenced the RAN-reading relationship, as well as whether RAN measures with different numbers of items per set or total items, were more strongly related to reading.

**Timing of Initial RAN Assessment and Later Reading Assessment**
Dyslexia is typically not diagnosed before the end of grade 2 because the heterogeneity of reading development profiles makes it difficult to reliably identify children who will have ongoing reading difficulty. Thus, it would be helpful to know when RAN assessment is effective for predicting later reading. In the US, kindergarten screening often includes literacy; thus, many studies that investigate longitudinal relations with RAN measure it at the start of kindergarten. However, some studies have assessed RAN in children as young as age 3.5 (McBride-Chang and Kail, 2002; Su et al., 2017). Widely used normed measures of RAN are available for children age 4 and up (e.g., CTOPP-2). Thus, we tested how the timing of RAN assessment (i.e., preschool versus kindergarten) differentially impacts the RAN-reading correlation.

Another important consideration is the timing of the later or ‘outcome’ reading measure, as the nature of the relations between early RAN and subsequent reading may change over the course of reading development. For example, early in reading development, children are developing accuracy in reading, and over time, they become accurate and build automaticity; thus, RAN may relate to fluency-based reading more strongly when reading is more automatized. In a practical sense, for early identification of reading problems, it may be important to know when this relation becomes stable. Wolf et al. (2000) suggested that RAN may play an attenuated role in predicting reading for typical readers after grade 2, because so many children achieve automaticity in naming and reading. Thus, we tested the extent to which the timing of reading assessment moderated the RAN-reading relationship.

**Distinct Associations with Reading from Phonological Awareness**
There is substantial shared variance between RAN and PA; thus, understanding each one’s unique longitudinal relationship with reading is essential to understand the broader picture of how pre-reading skills relate to reading ability (Schatschneider et al., 2004; Vander Stappen and Reybroeck, 2018). The double deficit hypothesis (Wolf and Bowers, 1999) generated considerable interest in this topic. Sufficient studies exist to extract and meta-analyze their intercorrelations, yet no meta-analysis has done so. We operationalized PA measures as any task that required a participant to manipulate or isolate phonemes in words or nonwords (phonological memory tasks such as nonword repetition were excluded). Our categories of PA measures were thus elision/deletion, isolation, blending, and matching/rhyming, as well as composite PA measures testing these subcategories. Thus, we tested the unique relationship between RAN and reading controlling for PA, using semipartial correlations.

**Outlier Handling**
Due to the nature of nested effect sizes, we examined outliers at the study level. We did this by taking the mean of each effect size and moderator variable at the study level and then testing whether any observations fell above the 97.5%ile or below the 2.5%ile. If a study fell outside of these values, it was further investigated and considered for inclusion on a case-by-case basis; importantly, this was done before analysis so as not to bias results. All studies/samples were retained for intercept-only models. For moderator analyses, several studies were excluded as they were outliers for the variable of interest. These cases are described in Supplemental Materials.

**Study Quality and Risk of Bias**
Study quality measures can be helpful in identifying whether certain designs, such as double-blind randomized control trials, yield less-biased estimates of effect sizes. Features that reflect study quality are less clear for correlational,
longitudinal research designs. Here, we use three measures of study quality and risk of bias: use of a published standardized RAN test, use of latent variables, and the study’s sample size. These were all separately analyzed as moderators of the RAN-reading relationship, as there is no gold-standard or guidance for doing so, we felt it was not appropriate to create a composite study quality and risk of bias measure.

**Statistical Power**

Power was calculated for each moderator analysis and is reported alongside each moderator analysis. As in Araújo et al. (2015), we used the value of 0.1 difference between Fisher’s z values as the smallest difference that would be meaningful. For the sample risk proportion analysis (e.g., low, medium, and high risk proportion), we used 1 Fisher’s z difference on either side of za = 0.4, as this is a typical RAN-reading correlation reported in other meta-analyses. As there is no widely accepted methodology for calculating moderator analyses’ power in robust variance estimation (RVE) models, we used the degrees of freedom from each moderator analysis (rounded to the nearest integer, which is effectively a sample size). We used the *metapower* package (Griffin, 2020, 2021) to calculate power for each moderator tested, using the mean sample size of n = 176 and an F value of 75%. Because this uses an *a priori* effect size estimate, this is not a *post hoc* power calculation. Power values for each analysis are presented alongside each model in Table 4. To calculate power for moderator analyses of semipartial correlations, we used a nearly identical procedure to the Pearson correlation power calculation. The only difference was that instead of using an F value of 75%, we used an F value of 50%, as this was much closer to the F of the intercept-only model of the semipartial correlations.

**Analysis Process and Plan**

**Meta-Analysis of RAN-Reading Correlations**

Reported effects in the literature were transformed from Pearson correlations to Fisher’s z-scores, which normalizes their distribution for analysis. They were then transformed back to Pearson correlations in results here, for ease of interpretation and comparison with other meta-analyses. To accommodate multiple effect sizes per study, we used correlated effects models using robust variance estimation (RVE) with the r (R Core Team, 2013) package *robumeta* (Fisher et al., 2017; Hedges et al., 2010). These models allow for correlated effects within a study, maximizing data retention. Furthermore, these models allow the grouping of multiple studies that share a sample (e.g., the International Longitudinal Twin Study; Furnes and Samuelsson, 2009, 2011). Intercept-only and moderator analyses were performed using the *robu* function. Moderators were tested in separate meta-regression models (e.g., separate models testing alphanumeric stimuli as a moderator and testing dyslexia risk as a moderator), except for time of assessment, in which the initial and outcome timepoints were considered together.

**Meta-Analysis of RAN-PA-Reading Semipartial Correlations**

To address the practical question of RAN’s unique contribution to reading, we coded the associations among PA, RAN, and reading. Correlation matrices from included studies were examined and the correlations between RAN-PA, PA-reading, and RAN-reading were extracted. For the semipartial analyses, correlations were not z-transformed, as semipartial correlations cannot be z-transformed (Aloe and Thompson, 2013). Pearson correlations (RAN-PA, PA-reading, RAN-reading) were used to calculate the semipartial correlations between RAN and reading, with the variance of PA partialled out. In order to pool these semipartial correlations, there needed to be equal numbers of RAN, PA, and reading measures per matrix. Because each study varied greatly in the number of measures for each construct, the simplest case of one measure for each construct (e.g., RAN, PA, or reading) was used to calculate each semipartial correlation. If multiple RAN, PA, or reading measures were used, the number of semipartial correlations calculated for each study could be represented by the formula nsp = nr × np × nreading. These semipartial correlations were then pooled using the methods outlined by Aloe & Becker (2012). The variance component for each semipartial correlation was calculated using equation 5 from Aloe & Becker (2012).

**Risk of Bias**

To test for funnel plot asymmetry, which is indicative of publication or reporting bias, we used a technique that allows for multiple effect sizes per study. Traditional methods for examining funnel plot asymmetry, such as Egger’s Regression or trim-and-fill analyses, only accommodate one effect size per study. “Sandwich” estimators (Rodgers and Pustejovsky, 2020) expand these methods to correlated effects models. We therefore used an “Egger’s Sandwich Regression” to test for funnel plot asymmetry. As our data came from a variety of sources, we also ran a moderator analysis to test whether published effect sizes were larger than unpublished effect sizes (e.g., an unpublished dissertation, data emailed from authors).

**Results**

**Sample Description**

The final analytic sample (n = 10,513) was drawn from 60 independent samples across 67 papers. Whereas the largest
TABLE 2
Descriptive Statistics for Samples Included in the Full Meta-Analysis

<table>
<thead>
<tr>
<th>Model</th>
<th>N</th>
<th>k</th>
<th>n</th>
<th>Mean</th>
<th>SD</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Initial timepoint</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>47</td>
<td>295</td>
<td>8552</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-K</td>
<td>13</td>
<td>51</td>
<td>2508</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mixed K/Pre-K</td>
<td>1</td>
<td>27</td>
<td>139</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Final timepoint</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 1</td>
<td>27</td>
<td>134</td>
<td>5972</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 2</td>
<td>28</td>
<td>164</td>
<td>3902</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 3</td>
<td>11</td>
<td>51</td>
<td>1621</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 4</td>
<td>8</td>
<td>24</td>
<td>2050</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time between measures</td>
<td>60</td>
<td>373</td>
<td>10513</td>
<td>27.38</td>
<td>11.16</td>
<td>12–57</td>
</tr>
<tr>
<td><strong>RAN task</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Published/Standardized</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>16</td>
<td>86</td>
<td>4526</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>46</td>
<td>287</td>
<td>6305</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Stimuli</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alphanumeric</td>
<td>22</td>
<td>109</td>
<td>4425</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-alphanumeric</td>
<td>50</td>
<td>255</td>
<td>9068</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAN colors</td>
<td>22</td>
<td>69</td>
<td>4044</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAN objects</td>
<td>29</td>
<td>118</td>
<td>5689</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAN letters</td>
<td>16</td>
<td>63</td>
<td>3196</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAN numbers</td>
<td>12</td>
<td>35</td>
<td>3232</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Composition</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAN total items</td>
<td>48</td>
<td>297</td>
<td>8457</td>
<td>72.28</td>
<td>43.77</td>
<td>24–216</td>
</tr>
<tr>
<td>RAN unique items</td>
<td>46</td>
<td>288</td>
<td>7136</td>
<td>5.84</td>
<td>2.50</td>
<td>4–20</td>
</tr>
<tr>
<td><strong>Sample risk proportion</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low risk</td>
<td>42</td>
<td>238</td>
<td>8528</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium risk</td>
<td>7</td>
<td>72</td>
<td>1579</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High risk</td>
<td>12</td>
<td>63</td>
<td>487</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Latent variable(s) used</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>5</td>
<td>12</td>
<td>1809</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>58</td>
<td>361</td>
<td>9879</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. N = number of samples/studies; k = number of effect sizes; n = number of participants.
The N for some sections may not sum to 10,513 as a result of these factors not being mutually exclusive within a study.

Sample size in the Hjetland et al. (2017) longitudinal RAN analyses was 3,746, the current sample is thus nearly three times greater, even though we restricted the language of the participants to English and the initial timepoint to before grade 1. For studies that reported age of participants at the initial timepoint, the mean age was 67.51 months (SD of 4.02) and a range of mean ages from 54–75 months across studies. The mean interval between initial and final
timepoint was 27.41 months, which is consistent with our prioritization of the Grade 2 timepoint. Other descriptive statistics for the samples included are presented in Table 2.

Intercept-Only Models

We calculated an intercept-only model to assess our main research question, the overall correlation between pre-school/Kindergarten RAN scores and later reading scores. The intercept-only model yielded a mean effect size of $z = -0.40$ (95% CI: $-0.37$ to $-0.44$, $p < .001$), equivalent to a Pearson correlation of $r = -0.38$. This indicates that on average, children with faster RAN time before grade school have stronger grade school reading performance. The forest plot for the overall intercept-only model is presented in Supplemental Material. Excluding studies that reported latent variables for RAN or reading resulted in nearly identical model results ($r = -0.38$). There was substantial variability in studies’ effect sizes ($F = 74.09$; $r^2 = 0.018$), indicating that analysis of moderators may further clarify the RAN-reading relationship. We also tested intercept-only models including only a subset of studies based on what types of RAN tasks and reading measures the study used. These results are presented in Table 3. All models were significant at $p < .001$, indicating that the relationship between various RAN and reading measures is quite robust.

Many papers that report a RAN-reading correlation also measured PA and reported its correlations with RAN and reading. The meta-analysis of the semipartial correlations ($r_{sp}$) calculated from these matrices had large samples ($N = 32$; $k = 353$; $n = 5,452$). The intercept-only model of the semipartial correlations yielded an effect of $r_{sp} = -0.25$; 95% CI $-0.28$ to $-0.22$.

Moderators and Meta-Regression

Primary practically and theoretically motivated moderators were analyzed and are presented in Table 4. We also tested whether partialling PA out of the RAN-reading relationship. We also tested intercept-only models including only a subset of studies based on what types of RAN tasks and reading measures the study used. These results are presented in Table 3. All models were significant at $p < .001$, indicating that the relationship between various RAN and reading measures is quite robust.

<table>
<thead>
<tr>
<th>Model</th>
<th>N</th>
<th>k</th>
<th>F</th>
<th>$r^2$</th>
<th>r</th>
<th>t</th>
<th>df</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>All studies/samples</td>
<td>60</td>
<td>373</td>
<td>74.09</td>
<td>0.018</td>
<td>-0.38</td>
<td>-22.35</td>
<td>50.21</td>
<td>[-0.44, -0.37]</td>
</tr>
<tr>
<td>RAN type</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colors</td>
<td>22</td>
<td>69</td>
<td>66.39</td>
<td>0.012</td>
<td>-0.32</td>
<td>-11.60</td>
<td>19.69</td>
<td>[-0.40, -0.27]</td>
</tr>
<tr>
<td>Objects</td>
<td>29</td>
<td>118</td>
<td>74.21</td>
<td>0.012</td>
<td>-0.34</td>
<td>-15.67</td>
<td>25.75</td>
<td>[-0.41, -0.31]</td>
</tr>
<tr>
<td>Letters</td>
<td>16</td>
<td>63</td>
<td>68.11</td>
<td>0.017</td>
<td>-0.46</td>
<td>-15.01</td>
<td>10.81</td>
<td>[-0.57, -0.42]</td>
</tr>
<tr>
<td>Digits</td>
<td>12</td>
<td>35</td>
<td>76.94</td>
<td>0.015</td>
<td>-0.45</td>
<td>-11.60</td>
<td>10.42</td>
<td>[-0.58, -0.39]</td>
</tr>
<tr>
<td>Reading measure types</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading comprehension</td>
<td>39</td>
<td>87</td>
<td>74.43</td>
<td>0.021</td>
<td>-0.38</td>
<td>-15.91</td>
<td>31.05</td>
<td>[-0.46, -0.35]</td>
</tr>
<tr>
<td>Reading fluency</td>
<td>23</td>
<td>54</td>
<td>77.84</td>
<td>0.036</td>
<td>-0.35</td>
<td>-7.95</td>
<td>17.60</td>
<td>[-0.47, -0.28]</td>
</tr>
<tr>
<td>Single word reading</td>
<td>50</td>
<td>193</td>
<td>69.30</td>
<td>0.015</td>
<td>-0.38</td>
<td>-22.28</td>
<td>40.59</td>
<td>[-0.44, -0.36]</td>
</tr>
<tr>
<td>Reading measure splits</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single word reading</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Real word reading</td>
<td>45</td>
<td>109</td>
<td>70.24</td>
<td>0.015</td>
<td>-0.41</td>
<td>-24.43</td>
<td>38.85</td>
<td>[-0.46, -0.39]</td>
</tr>
<tr>
<td>Nonword reading</td>
<td>38</td>
<td>84</td>
<td>66.59</td>
<td>0.013</td>
<td>-0.33</td>
<td>-16.05</td>
<td>28.48</td>
<td>[-0.38, -0.29]</td>
</tr>
<tr>
<td>Timing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Timed reading</td>
<td>33</td>
<td>137</td>
<td>81.27</td>
<td>0.032</td>
<td>-0.37</td>
<td>-11.70</td>
<td>26.86</td>
<td>[-0.46, -0.32]</td>
</tr>
<tr>
<td>Untimed reading</td>
<td>57</td>
<td>223</td>
<td>70.25</td>
<td>0.014</td>
<td>-0.37</td>
<td>-21.78</td>
<td>48.48</td>
<td>[-0.43, -0.36]</td>
</tr>
<tr>
<td>Efficiency and accuracy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Efficiency</td>
<td>22</td>
<td>57</td>
<td>52.52</td>
<td>0.009</td>
<td>-0.40</td>
<td>-19.15</td>
<td>14.12</td>
<td>[-0.47, -0.38]</td>
</tr>
<tr>
<td>Accuracy</td>
<td>48</td>
<td>155</td>
<td>70.07</td>
<td>0.015</td>
<td>-0.37</td>
<td>-20.48</td>
<td>41.42</td>
<td>[-0.43, -0.35]</td>
</tr>
</tbody>
</table>

Note. $N =$ number of samples/studies; $k =$ number of effect sizes. All models were significant at $p < .001$. 

**TABLE 3**

Main Effects: Intercept-Only Models
<table>
<thead>
<tr>
<th>Model</th>
<th>N</th>
<th>k</th>
<th>I²</th>
<th>r</th>
<th>t</th>
<th>df</th>
<th>p</th>
<th>95% CI</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Practical considerations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unique RAN tokens</td>
<td>45</td>
<td>285</td>
<td>68.49</td>
<td>0.016</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td></td>
<td></td>
<td>-0.34</td>
<td>-3.84</td>
<td>5.94</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unique tokens</td>
<td></td>
<td></td>
<td>-0.01</td>
<td>-0.81</td>
<td>4.52</td>
<td>0.46</td>
<td></td>
<td>[-0.04 0.03]</td>
<td>0.11</td>
</tr>
<tr>
<td>Total RAN items</td>
<td>46</td>
<td>290</td>
<td>63.80</td>
<td>0.012</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td></td>
<td></td>
<td>-0.46</td>
<td>-13.5</td>
<td>25.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total items</td>
<td></td>
<td></td>
<td>0.00</td>
<td>1.19</td>
<td>10.7</td>
<td>0.26</td>
<td></td>
<td>[-0.00 0.00]</td>
<td>0.19</td>
</tr>
<tr>
<td>Standardized RAN test</td>
<td>60</td>
<td>373</td>
<td>73.94</td>
<td>0.019</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td></td>
<td></td>
<td>-0.40</td>
<td>-21.07</td>
<td>37.81</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Published/STD test</td>
<td></td>
<td></td>
<td>0.06</td>
<td>1.37</td>
<td>23.52</td>
<td>0.18</td>
<td></td>
<td>[-0.03 0.15]</td>
<td>0.36</td>
</tr>
<tr>
<td>Age at assessments</td>
<td>60</td>
<td>373</td>
<td>73.97</td>
<td>0.019</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td></td>
<td></td>
<td>-0.31</td>
<td>-3.28</td>
<td>21.25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Initial (RAN) age (mos.)</td>
<td></td>
<td></td>
<td>-0.01</td>
<td>-1.95</td>
<td>19.52</td>
<td>0.07</td>
<td></td>
<td>[-0.01 0.00]</td>
<td>0.31</td>
</tr>
<tr>
<td>Final (Reading) age (mos.)</td>
<td></td>
<td></td>
<td>0.00</td>
<td>0.04</td>
<td>19.51</td>
<td>0.97</td>
<td></td>
<td>[-0.00 0.00]</td>
<td></td>
</tr>
<tr>
<td><strong>Theoretical considerations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alphanumeric vs. non-alphanumeric</td>
<td>58</td>
<td>364</td>
<td>69.45</td>
<td>0.015</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td></td>
<td></td>
<td>-0.46</td>
<td>-11.05</td>
<td>14.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-alphanumeric</td>
<td></td>
<td></td>
<td>0.13</td>
<td>2.78</td>
<td>21.83</td>
<td>0.01</td>
<td>[0.03 0.23]</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>Nonword vs. real word reading</td>
<td>50</td>
<td>193</td>
<td>66.55</td>
<td>0.013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td></td>
<td></td>
<td>-0.33</td>
<td>-15.62</td>
<td>28.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Real word measure</td>
<td></td>
<td></td>
<td>-0.09</td>
<td>-3.73</td>
<td>37.09</td>
<td>&lt;.001</td>
<td></td>
<td>[-0.14 -0.04]</td>
<td>0.51</td>
</tr>
<tr>
<td>Timed vs. untimed reading</td>
<td>58</td>
<td>360</td>
<td>72.56</td>
<td>0.017</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td></td>
<td></td>
<td>-0.37</td>
<td>-23.08</td>
<td>40.97</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Timed reading</td>
<td></td>
<td></td>
<td>0.00</td>
<td>0.13</td>
<td>31.74</td>
<td>0.90</td>
<td>[-0.06 0.07]</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>Efficiency vs. accuracy</td>
<td>56</td>
<td>212</td>
<td>69.45</td>
<td>0.015</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td></td>
<td></td>
<td>-0.38</td>
<td>-20.23</td>
<td>39.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Efficiency</td>
<td></td>
<td></td>
<td>-0.01</td>
<td>-0.035</td>
<td>18.83</td>
<td>.73</td>
<td>[-0.09 0.06]</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>Sample risk proportion</td>
<td>60</td>
<td>373</td>
<td>74.50</td>
<td>0.019</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td></td>
<td></td>
<td>-0.35</td>
<td>-5.62</td>
<td>7.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(continued)
relationship changed the theoretically motivated moderator effects; these analyses will be referred to as semipartial moderator analyses, as opposed to the primary moderator analyses, and are presented in Table 5. Several moderators changed considerably when PA was partialled out. To ensure that these changes were not due to the specific subset of studies included in semipartial analysis, the primary meta-analysis models were re-run with the same subset of studies as the semipartial correlation analyses. This subset of studies will be referred to as the subset of semipartial studies, for which the sample size is \(n = 5452\) compared to \(n = 10,513\) for the full sample.

**Table 4**

<table>
<thead>
<tr>
<th>Model</th>
<th>(N)</th>
<th>(k)</th>
<th>(I^2)</th>
<th>(t^2)</th>
<th>(d_f)</th>
<th>(p)</th>
<th>95% CI</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low risk</td>
<td>-0.05</td>
<td>-0.78</td>
<td>9.19</td>
<td>.46</td>
<td>-0.20</td>
<td>.10</td>
<td>[-0.20 0.10]</td>
<td>.36</td>
</tr>
<tr>
<td>Medium risk</td>
<td>-0.01</td>
<td>-0.17</td>
<td>11.78</td>
<td>.87</td>
<td>-0.19</td>
<td>.16</td>
<td>[-0.19 0.16]</td>
<td></td>
</tr>
</tbody>
</table>

Note. \(N = \) number of studies; \(k = \) number of effect sizes, All intercepts were significant at \(p < .01\). Moderator effects indicated in bold are \(p < .05\).

**Practical Moderators**

**Unique RAN Items and Total RAN Items**

We tested whether specific features of the RAN task administered in each study, such as the number of total items or the number of unique items, were differentially predictive of later reading. We found that neither the number of total items, nor the number of unique items moderated the RAN-reading relationship (all \(p > .26\)). This indicates that RAN test length and item composition, within the limits of what has been studied, does not meaningfully modify the RAN-reading relationship.

**Standardized RAN Measure**

Next, we tested whether using published assessments that are standardized and normed, such as the RAN/RAS Tests or the RAN subtests from the CTOPP, affected the RAN-reading relationship. We found that using a published assessment had no effect (\(\Delta r = 0.06; p = .18\)) on the strength of the RAN-reading relationship. This also was an indicator of risk of study bias, indicating that study quality may be less likely to bias these results.

**Age at Assessments**

We tested whether the timing of the RAN or reading assessments (e.g., earlier or later than initial assessment at early kindergarten for RAN assessment or than Grade 2 for reading assessment) moderated the RAN-reading relationship. We found that age at reading assessment had no moderating effect (\(\Delta r = 0.00; p = .97\)), but that age at RAN assessment did have a marginally significant effect (\(\Delta r = -0.01; p = .07\)), in the direction of later assessment having a stronger RAN-reading relationship. We considered that this result may be conflated with whether alphanumeric RAN was assessed or not, as younger children are less likely to be able to complete alphanumeric RAN, and alphanumeric RAN has been a stronger predictor than non-alphanumeric RAN in previous meta-analyses (Araújo et al., 2015). After controlling for whether the RAN task was alphanumeric or not, there was no effect of age at initial assessment (\(\Delta r = 0.00; p = .15\)). This result indicates that the exact timing of early RAN measurement does not differentially affect the RAN-reading relationship.

**Theoretical Moderators**

**Alphanumeric versus Non-alphanumeric RAN**

The correlations for RAN letters and RAN digits with reading were nearly identical (\(r = -0.46\) and \(r = -0.45\), respectively), as were correlations for RAN colors and RAN objects with reading (\(r = -0.32\) and \(r = -0.34\), respectively). Based on these values, the fact that studies find RAN digits to be automatized even earlier than letters (Åvall et al., 2019) and to be consistent with previous meta-analyses that combined these categories (e.g., Araújo et al., 2015), we collapsed the RAN types into alphanumeric and non-alphanumeric RAN. We then directly tested whether alphanumeric RAN was a better predictor of reading than non-alphanumeric RAN. We found that alphanumeric RAN is a significantly stronger predictor of reading (\(\Delta r = 0.13; p = .01\)), meaning that RAN tasks with letters or numbers had a stronger correlation with reading than did tasks with colors or objects. To consider the possibility that this relationship was conflated with initial age (because younger children may be less likely to have completed an alphanumeric task successfully), we ran the same analysis controlling for initial age, and the effect was unchanged (\(\Delta r = 0.13; p = .01\)). In sum, for our samples’ ages, alphanumeric RAN was a stronger predictor of future reading regardless of age. However, it may be the case that studies considered age when selecting their RAN measures and tended to administer alphanumeric measures for children who were already automatic with those stimuli, as is intended. To test whether partialling out PA affected this relationship, we tested the moderator effect for Pearson correlations in the subset of semipartial studies (\(\Delta r = 0.09; p = .02\)), which was again significant. With PA partialled...
out, whether the RAN task was alphanumeric or not had a marginal effect on reading ability ($\Delta r_{sp} = 0.07; p = .07$).

**Real versus Nonword Reading**

Next, we directly tested whether measures of nonword reading had a weaker relationship with RAN than measures of single, real word reading. We found a significant effect ($\Delta r = -0.09; p < .001$), with measures of nonword reading having a weaker relationship with RAN than measures of single, real word reading. This effect was unchanged in the subset of semipartial studies ($\Delta r = -0.10; p = .01$). However, with PA partialled out, real word and nonword reading did not have a differential relationship with RAN ($\Delta r_{sp} = 0.04; p = .19$).

**Timed versus Untimed Reading**

We then tested whether timed reading measures were more related to RAN than untimed measures. We found no difference ($\Delta r = 0.00; p = .90$) between timed and untimed reading measures as they relate to RAN. In the subset of semipartial studies ($\Delta r = 0.01; p = .88$), as well as with PA partialled out, timed and untimed reading tasks had no significant moderating effect ($\Delta r_{sp} = 0.05; p = .11$).

**Reading Efficiency versus Reading Accuracy**

As there were no differences in timed versus untimed reading measures, we also tested whether measures of reading efficiency were more related to RAN than measures of reading accuracy only. We found no difference ($\Delta r = -0.01; p = .73$) between how measures of reading efficiency and reading accuracy relate to RAN. This effect was unchanged in the subset of semipartial studies ($\Delta r = 0.03; p = .57$). However, with PA partialled out, reading efficiency measures had a significantly stronger relationship with RAN than reading accuracy measures ($\Delta r_{sp} = 0.08; p = .03$).

**Dyslexia Risk Proportion in the Sample**

Using the three-level classification of dyslexia risk of the sample (low, medium, or high proportion of children at

### TABLE 5

**Main and Moderator Effects for Semipartial Correlation Meta-Analysis**

<table>
<thead>
<tr>
<th>Model</th>
<th>$N$</th>
<th>$k$</th>
<th>$I^2$</th>
<th>$r^2$</th>
<th>$r$</th>
<th>$t$</th>
<th>$df$</th>
<th>$p$</th>
<th>95% CI</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>All studies/samples</td>
<td>32</td>
<td>353</td>
<td>60.94</td>
<td>0.007</td>
<td>-0.25</td>
<td>-17.7</td>
<td>27</td>
<td>&lt;.001</td>
<td>[-0.28,-0.22]</td>
<td>0.99</td>
</tr>
<tr>
<td><strong>Theoretical considerations</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>Alphanumeric vs. Non-Alphanumeric</td>
<td>31</td>
<td>350</td>
<td>50.95</td>
<td>0.015</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.30</td>
<td>-9.66</td>
<td>6.71</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Alphanumeric</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.07</td>
<td>2.07</td>
<td>9.45</td>
<td>.07</td>
<td>[-0.01, 0.14]</td>
<td>0.28</td>
</tr>
<tr>
<td>Nonword vs. Real Word Reading</td>
<td>26</td>
<td>203</td>
<td>65.11</td>
<td>0.008</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.23</td>
<td>-8.29</td>
<td>15.40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Real Word Measure</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.04</td>
<td>-1.35</td>
<td>20.80</td>
<td>.19</td>
<td>[-0.09, 0.02]</td>
<td>0.56</td>
</tr>
<tr>
<td>Timed vs. Untimed Reading</td>
<td>32</td>
<td>347</td>
<td>58.89</td>
<td>0.006</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.23</td>
<td>-16.68</td>
<td>22.80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Timed Reading</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.06</td>
<td>-1.92</td>
<td>18.80</td>
<td>.07</td>
<td>[-0.12, 0.01]</td>
<td>0.52</td>
</tr>
<tr>
<td>Efficiency vs. Accuracy</td>
<td>31</td>
<td>223</td>
<td>53.53</td>
<td>0.005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.24</td>
<td>-14.49</td>
<td>22.67</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Efficiency</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.08</td>
<td>-3.06</td>
<td>9.37</td>
<td>.01</td>
<td>[-0.14, 0.02]</td>
<td>0.28</td>
</tr>
<tr>
<td>Sample Risk Proportion</td>
<td>32</td>
<td>353</td>
<td>60.78</td>
<td>0.007</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.26</td>
<td>-15.86</td>
<td>18.30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Risk</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.04</td>
<td>1.32</td>
<td>15.20</td>
<td>.21</td>
<td>[-0.23, 0.08]</td>
<td>0.43</td>
</tr>
</tbody>
</table>

*Note. $N$ = number of studies; $k$ = number of effect sizes. All intercepts were significant at $p < .05$. Moderator effects indicated in **bold** are $p < .05$.***
risk) in a single model, we tested whether the RAN-reading relationship was affected by dyslexia risk. There was no significant moderating effect of level of dyslexia risk (all $\Delta r \leq 0.05; \forall p > 0.46$). In order to ensure that this was not specific to this grouping categorization, we also ran a model using a dichotomous categorization of risk (i.e., general population versus any type of high-risk sample) and found highly similar results ($\Delta r = 0.05; p = .31$). There was also no effect of dichotomized risk in the subset of semipartial studies ($\Delta r = 0.07; p = .11$), and there was no significant moderating effect of dyslexia risk with PA partialled out ($\Delta r = 0.04; p = .21$). These results indicate that the RAN is a similar predictor of reading across samples of children that vary in risk for dyslexia.

**Risk of Bias Analysis**

To assess risk of bias, we ran an Egger’s Sandwich Regression, in which the standard deviation estimates from each study were used as the moderator. We found no risk of bias in our effect size estimates ($p = .32$). Sample size is often used as a study quality measure as well; this result indicates that sample size has no significant effect on effect size estimates. However, because our data were composed of peer-reviewed studies, unpublished theses, and emailed data from published studies, we also ran moderator analyses with whether data were from a published paper or not (i.e., an unpublished dissertation or emailed data). These analyses revealed strong evidence of reporting bias, with published effect sizes being stronger than unpublished effect sizes ($\Delta r = 0.09; p = .02$). This effect was not driven by the inclusion of dissertation manuscripts ($\Delta r = 0.003; p = .97$), but rather by other types of unpublished data (e.g., emailed data). Due to the highly nested nature of these data, a funnel plot visualization is not provided, given that plotting up to 27 effect sizes with the same standard error would result in essentially a horizontal line on the funnel plot and be difficult to interpret.

**Discussion**

This meta-analysis expands on previous findings by documenting the longitudinal relationship between early RAN and various measures of later reading abilities in English-speaking children. Consistent with previous research and meta-analyses, RAN tasks were found to be a strong predictor of all types of reading. The mean effect size found for RAN predicting reading overall ($r = −0.38$) is similar to meta-analyses of concurrent RAN-reading correlations, with $r$ ranging from −0.38 to −0.45 depending on reading measure in Swanson et al. (2003), $r = −0.43$ Araújo et al. (2015), $r = −0.34$ for reading comprehension, and $r = −0.37$ for Word ID in Hjetland et al. (2017). We also estimated the semipartial correlation of early RAN on future reading controlling for PA ($r_{sp} = −0.25$), distilling decades of research that has studied RAN unique effect on reading beyond the contribution of PA.

Our meta-analysis adds uniquely to the literature assessing the links between RAN and reading by highlighting the relevance of assessing RAN in kindergarten or preschool, and the robustness of this relationship over time and across various RAN and reading measures. The only existing longitudinal meta-analysis between RAN and reading was limited in its coverage of the literature and theoretical scope, with no moderators assessed (Hjetland et al., 2017). Our database searching, in conjunction with a snowball search strategy, yielded many more included articles, resulting in a sample size nearly three times larger. This much larger sample was ascertained despite restricting our age range to kindergarten and preschool and restricting our language to English.

Though RAN has long been considered independent of PA (Bowers and Wolf, 1993; Wolf and Bowers, 1999), the shared variance between the two is considerable, and parsing their independent effects is essential to understanding their respective contributions to reading outcomes (Norton and Wolf, 2012; Vander Stappen and Reybroeck, 2018). We have therefore meta-analytically demonstrated for the first time the unique contribution of early RAN to later reading above early PA. This was ascertained by meta-analyzing semipartial correlations that were derived from correlation matrices. This analysis is the first step toward creating longitudinal meta-analytic path models of cognitive, pre-reading, and reading variables. We thus strongly advocate for researchers to share correlation matrices (and/or raw data), such as through Supplemental Material and platforms such as Open Science Framework.

Another major contribution of the present study is the analysis of a variety of potential practical and theoretical moderators of the relationship between early RAN and later reading. For practical moderators, our analyses show that number of total items, and how many unique items were included in each set did not moderate the RAN-reading relationship align with and extend previous concurrent findings from Araújo et al. (2015). Our study is the first, to our knowledge, to examine RAN tasks that were published and standardized versus researcher-created; these variations also did not significantly alter the predictive relation of RAN with reading. In sum, these results show that RAN’s relationship to reading is robust, regardless of how the measure is constructed. Whereas educators may not always have access to published, standardized measures, these data suggest that some RAN information is better than nothing.
For theoretical moderators, we found that RAN has a significantly stronger relation with reading when alphanumeric stimuli are used. This replicates and extending a previous concurrent meta-analysis across ages (Araújo et al., 2015), even despite the young age of the RAN assessments analyzed here. Partialling PA out slightly changed the moderating effect of alphanumeric stimuli from significant (Δr = 0.09; p = .02) to marginally significant (Δr = 0.07; p = .07), but these small changes do not meaningfully change our interpretation. In considering different reading measures as outcomes, we found only a significant difference for RAN better predicting real word than nonword reading from the primary moderator analyses. However, with early PA partialed out, RAN correlated similarly with nonword and real word reading (Δr = 0.04; p = .19). We also found differences in reading efficiency measures versus reading accuracy measures, only with PA partialed out. In contrast, Araújo and colleagues found differences between timed and untimed measures across orthographies and ages, without partialed out PA. We discuss the implications of these findings for theory and for practice, below.

**Insights to the Nature of the RAN-Reading Relationship**

Our primary moderator analyses show that alphanumeric RAN has a significantly stronger relationship with later reading than does non-alphanumeric RAN, as well as that nonword reading is significantly less related to RAN than real word reading. These results, taken together, support shared cognitive processes models, such that the more similar the processes that RAN and a given reading task tap, the more strongly that they will be correlated (Georgiou and Parrila, 2020). In the case of nonword reading, there is a heavy phonological decoding (letter-to-sound correspondence) component that RAN does not share, which is why partialed out PA reduces this effect. In other words, when PA was controlled for, RAN had no differentiable relationship to real word versus nonword reading. In the case of alphanumeric RAN, symbolic representation is required for both alphanumeric RAN and reading. Individual studies have found that alphanumeric RAN and non-alphanumeric RAN correlate equally well with later reading (e.g., Van den Bos et al., 2002) or that both load on the same latent factor (Papadopoulos et al., 2016). However, our meta-analysis in young children shows that alphanumeric RAN is stronger than non-alphanumeric RAN regardless of whether RAN was measured in preschool or kindergarten, and that age on its own had no effect on the RAN-reading relationship once the alphanumeric stimulus type effect was accounted for. This is strongly consistent with meta-analytic findings from Araújo et al. (2015). The effect is large in both the current and Araújo et al.’s meta-analyses, but not so large that it would be unexpected for an individual study to find similar correlation sizes between reading and alphanumeric and non-alphanumeric RAN.

Our results also show an interesting pattern for timed measures versus untimed measures, as well as reading efficiency versus reading accuracy. In the primary moderator analyses, neither timed versus untimed nor efficiency versus accuracy showed significant results. However, by partialed out the effect of PA, it is clear that RAN alone has a stronger relationship with reading efficiency than reading accuracy measures. Though the semipartial moderator analysis for timed versus untimed measures did not reach significance (Δr = 0.05; p = .11), there was a moderate change from the primary moderator analyses which show the same RAN-reading correlations for timed and untimed measures (Δr = 0.00; p = .90).

One potential reason the semipartial moderator analyses did not reach significance for timed versus untimed measures is that in the early years of reading development, accuracy-based and time-based measures are strongly correlated (e.g., Schatschneider et al., 2004). A difference emerges in intermediate and advanced readers once children build reading automaticity, but it is not present in beginning readers in either our sample or in the beginning and pre-readers included in the meta-analysis from Araújo et al. (2015). This may be particularly true for the English-speaking samples used here, as reading accuracy takes longer to transition to reading efficiency in opaque orthographies (Seymour et al., 2003). These findings are consistent with the idea that reading accuracy is not yet automatic in early grades in English (Chall, 1983; Samuels and Flor, 1997), and as a result, various reading measures may be more highly correlated early in schooling (i.e., less differentiable) than they are at later stages when most children have developed automaticity. More highly correlated reading measures in our earlier outcome timepoint (centered around 2nd grade) would likely result in weaker moderating effects when comparing different types of reading measures.

Consistent with other meta-analyses’ findings of no differences in relations with RAN between good versus poor readers, we found no difference between samples with a large proportion of children at-risk for dyslexia and those with very few at risk. This may indicate that children at-risk and children not at risk are using similar cognitive processes, even if these processes are impaired in children at risk. Although we are not fully able to explore the lower tail of RAN and reading performers, these results further support the idea that RAN is a continuous ability and dimensionally predicts of reading, rather than a dichotomous “present or absent” skill.

**Practical Insights for Using RAN as a Screener**

These results provide practical insights into using RAN for effective screening for later reading difficulties.
Importantly, RAN should always be assessed as part of a battery of screening measures, as RAN alone only predicts 14% of variance in future reading scores. No screening battery is perfectly accurate (with no false positives or negatives), but a nuanced understanding of a child’s profile will provide educators with the clearest path forward. Nonetheless, our results indicate that the relation between early RAN and later reading is remarkably consistent. The particular characteristics of the RAN measure, such as number of items and whether the task was from a published test, did not significantly alter the strength of the RAN-reading relationship. These facets of RAN as a predictor had not been assessed in previous meta-analyses, yet they provide concrete guidance for researchers and educators in planning RAN measures for screening. There was not a significant difference between RAN measures conducted in preschool versus kindergarten in terms of their relationship with later reading; there was a trend toward stronger predictive power, but the trend was reduced when controlling for alphanumeric RAN, which is often administered in later years. The advantage of earlier identification of potential reading difficulties, so that earlier intervention can be provided, suggests that it would be optimal to employ RAN tasks in screening in pre-school or pre-kindergarten, as soon as RAN can be assessed validly.

The stimulus type used in early RAN assessment is a relevant consideration, as alphanumeric RAN measures were more strongly related to later reading than were non-alphanumeric measures. An important caveat is that RAN tasks, by definition, depend on the child being able to name items with automaticity, and many articles noted that many children could not perform a RAN Letters task in kindergarten, as their letter name knowledge was not yet accurate and automatic (e.g., Catts et al., 1999). Thus, for children in kindergarten or preschool who do not yet know the names of letters or digits automatically, a RAN task using colors or objects would be a better choice; once letters or digits are known with automaticity, those are a better choice for later reading prediction. To what degree a speeded naming task is automatized in young children has long been debated (e.g., Vall et al., 2019; Wolf et al., 1986) and is not particularly testable in a meta-analysis. Nonetheless, our results clearly demonstrate that RAN, when measured at a young age, maintains its robust relationship with reading.

The question that frequently follows after RAN screening is “what RAN time or score is worrisome?” Unfortunately, research has not yet determined a single cutoff score for “dyslexia risk” or what is “good” versus “poor” RAN; in fact, this may not be possible given that RAN is both a continuous measure and one aspect of the constellation of reading-related abilities. At this point, using a published, standardized RAN measure that provides standard scores or percentiles provides the advantage that it may help educators and clinicians understand where a child’s RAN ability falls relative to their peers as an indicator of risk for dyslexia, even though our data showed that researcher-created measures equally predicted later RAN. It is important to note that administering a RAN task according to any standardized instructions and minimizing distractions so as to obtain the child’s best performance is crucial to obtaining a valid score.

Educators and clinicians should also recognize that an effective screening battery for dyslexia and reading difficulties must include RAN alongside other indicators such as phonological awareness (see Petscher et al., 2019, for recommendations). Even using the most evidence-based screening tools in combination with assessment of the child’s family or neuroimaging measures, there is still uncertainty about which children will develop reading difficulty (Norton et al., 2019; Zuk et al., 2020). As the field moves forward in understanding early indicators of reading difficulties, RAN will undoubtedly play a role, given its universal and robust relation with reading.

**Limitations**

There are several limitations of this study to consider. The primary limitation was that we restricted our sample to only English-speaking students. As English is an outlier orthography, many of our findings about the transition from accurate to efficient reading are not generalizable to more transparent orthographies. Specifically, the children in our study likely acquire reading efficiency later than those learning transparent orthographies, which would affect many of our analyses, such as RAN’s relationship to timed measures. We plan to address this shortcoming in future studies that include cross-language comparisons.

Another potential limitation of our study was our decision to not create composite measures of study quality. Instead, we chose to analyze study quality in terms of moderators, based on the concern over validity of using simple sums to describe study quality (Shamliyan et al., 2010; Whiting et al., 2005). Similarly, Hjetland et al. (2017) found no effect of study quality in an overlapping sample of papers, which aligns with our results that sample size, latent variables, and use of published/standardized tests do not predict variation in effect sizes. These variables functionally comprise study quality in longitudinal designs capturing the relationship between RAN and reading.

Another limitation is the limited statistical power for moderator analyses. Although we found no differences for unique RAN items or total RAN items, we had limited power to detect possible effects for a multitude of reasons. Araújo et al. (2015) noted similar difficulties, even with a larger corpus of sources and subjects. We offer the same caution in interpreting our moderator analysis results with low power.

Other limitations relate to the RAN tasks themselves. One limitation was the fact that there were incomplete...
descriptions of the measures in many studies, which was particularly common for researcher-created RAN tasks. Despite our effort to carefully review all available information in the published papers (and in many cases, request additional details from authors via email), many papers had incomplete descriptions of their RAN tasks, particularly relating to how many unique items and how many total items the task had. Furthermore, there was not much variability in the number of unique items, as many articles used Denckla & Rudel’s (1976) version or the updated RAN-RAS tests (Wolf and Denckla, 2005) each with 5 unique items per task, or the CTOPP that has 6 unique items. Despite the incomplete information from a number of studies, we believe we had sufficient power to detect these effects if they truly existed, as 288 (of 373) effect sizes were analyzed for the model that tested unique and total items as moderators.

The definition of at-risk in samples also varied greatly across studies and could limit interpretation of our results. For example, Cardoso-Martins & Pennington (2004) recruited a high-risk group from the children whose one of the parents has reading problems and a low-risk group from the children with no family history of reading problems. Hulme et al. (2015) also divided groups based on family history; however, they included another criterion of whether children have language impairment or not. In contrast, Heath & Hogben (2004) divided groups only based on poor and good phonological awareness abilities. Felton (1992) used teacher ratings of children’s expected reading ability. There is strong evidence for different subtypes or component skills in dyslexia even beyond the double deficit (e.g., O’Brien et al., 2012), and pooling these samples could miss whether the RAN-reading relationship changes with the etiology for a given subgroup. Furthermore, examining the lower end of the RAN distribution through the lens of dyslexia risk does not directly test nonlinearities in the relationship between RAN and reading. Nonetheless, the heterogeneity present in our coding reflects the real-world heterogeneity of risk definitions, and our categories were designed to reflect that.

A final limitation is that the studies selected for the semipartial analyses may have some bias. Specifically, the reporting of correlation matrices in supplementary or primary data has become somewhat standard practice for large studies. The results from primary moderator and semipartial moderator analyses appeared highly similar, but we cannot rule out that some bias may be present in selecting these studies for a semipartial correlational meta-analysis.

**Future Directions**

We chose to focus on only traditional RAN tasks at certain timepoints in the English language in order to maximize practical and policy impact. As a result, there are several clear directions for future research to expand upon our study by broadening the scope. Future studies may consider different designs, such as meta-analytic path modeling of the relationships among cognitive, pre-reading, and reading variables. Though the majority of studies and all published tests focus on RAN total time, aspects of RAN such as analyses of inter-item pause times as a predictor would be promising to investigate, as pause times have been shown to relate highly with reading fluency (Lervåg and Hulme, 2009).

Given that we focused on a single outcome timepoint in each study that was close to the end of Grade 2, another potential future direction would be to test how longitudinal RAN-reading relationships change within studies and more broadly over time. As we prioritized collecting only one time point per study, we were not able to analyze whether correlations from early RAN to later reading changed over time within a study, as is suggested by a number of authors (de Jong and van der Leij, 2002; Wagner et al., 1997). To our knowledge, correlated effects RVE models have not been used to analyze longitudinal, within-study data. Many of the papers collected for the present analysis would be ideal to use in testing whether RVE is suitable for longitudinally dependent effect sizes and provide further insight into how RAN relates to reading over time.

Another clear direction for future research is to include multiple languages, as well as individuals who speak multiple languages, to assess similarities and differences of RAN as a predictor reading ability (Gottardo et al., 2021). In the past, other authors had suggested that RAN is a better predictor in more transparent languages (see Georgiou et al., 2008). In their meta-analysis, Araújo et al. (2015) reported that orthographically opaque orthographies such as English have a stronger concurrent correlation between RAN and reading than do transparent orthographies, but we do not have meta-analytic evidence of this effect longitudinally. Cross-linguistic studies have provided evidence that kindergarten RAN may be a stronger longitudinal predictor in opaque orthographies than more transparent orthographies, but there are no significant differences across languages for RAN measured in grade 1 (Furnes and Samuelsson, 2011; Landerl et al., 2021). Other studies have found equally strong correlations in transparent orthographies such as Czech (Caravolas et al., 2013), and qualitative reviews have noted that the longitudinal, cross-linguistic effect is likely small (Landerl et al., 2021). Taken together, this further highlights the need for a larger systematic approach that is sensitive to the many between-study differences in cross-linguistic research, such as the selection of developmentally appropriate reading measures across languages (see Papadopoulos et al., 2021 for a review).

Finally, given that a major focus was the utility of using RAN as a screener, future research should endeavor
to provide concrete recommendations of what RAN performance indicates meaningful risk for reading difficulties and dyslexia. Few studies have provided clear formulas or cutoffs about which children are at greatest risk (Catts et al., 2001 is a notable exception). Even fewer studies have examined how best to provide intervention specific to children who have RAN difficulties that impact their reading, as it seems that training RAN itself is not effective in improving reading (de Jong and Vriealink, 2004; Kirby et al., 2010). Indeed, early measures of RAN may be an important, easy-to-collect early indicator of reading problems, akin to a “check engine light” that signals the need for further assessment and monitoring (Norton, 2020).

Acknowledgment
We thank Dr. Beth Tipton for analytic and statistical guidance. We thank Meakailyn Phillips, Anu Raife, Cadence Reed-Bippen, and Alice Wang for assistance with data screening and coding. We acknowledge Bailey Sone for assistance with data organization. REDCap is supported by a Clinical and Translational Science Award (CTSA) grant UL1TR001422 from the National Institutes of Health (NIH). Funding was provided by Northwestern University to ESN. Data and code are available on the Open Science Framework. Data have not been previously presented or published.

Conflict of Interest
There are no conflicts of interest to disclose.

NOTES
1 Note that here, we present all correlations as negative, despite factors like raw versus standard scores, indicating that faster RAN is associated with better reading, as this is usually the observed direction of the relation.

2 Bishop & League (2006) reported a positive correlation between RAN time and reading ability (it appears the authors used raw time measures of RAN). However, in an earlier report from the same sample, Bishop (2003) reported positive correlations using standard scores (in the expected direction of this relationship). The RAN-reading correlations from this paper should likely have been treated as negative in this case for Hjelmas analyses, as all other measures in this and other meta-analyses were negative.

3 Some articles were triple-screened during training, but all other articles were double-screened.

REFERENCES
* indicates paper included in meta-analysis; a full list of sources included in meta-analysis is available in supplemental material.


Submitted July 9, 2021
Final revision received March 7, 2022
Accepted March 8, 2022

SEAN MCWEENY is a doctoral candidate in the Department of Communication Sciences and Disorders at Northwestern University, Evanston, IL, USA. Email: SeanMcWeeny2022@u.northwestern.edu.

SOUJIN CHOI is a doctoral candidate in the Department of Communication Sciences and Disorders at Northwestern University, Evanston, IL, USA. Email: SouChoi2024@u.northwestern.edu.

JUNE CHOE was an undergraduate in the LEARN Lab at Northwestern University, Evanston, IL, USA, when this work was conducted. He is now a doctoral student in the Department of Linguistics at the University of Pennsylvania, Philadelphia, PA, USA. Email: jchoe001@gmail.com.

ALEXANDER LATOURRETTE was a doctoral student in the Department of Psychology at Northwestern University, Evanston, IL, USA, when this work was conducted. He is now a postdoctoral scholar in the Department of Psychology at the University of Pennsylvania, Philadelphia, PA, USA. Email: alexanderlatourrette2019@u.northwestern.edu.

MEGAN Y. ROBERTS is an associate professor in the Department of Communication Sciences and Disorders at Northwestern University, Evanston, IL, USA. Email: megan.y.roberts@northwestern.edu.

ELIZABETH S. NORTON (corresponding author) is the Jane Steiner Hoffman and Michael Hoffman assistant professor in the Department of Communication Sciences and Disorders at Northwestern University, Evanston, IL, USA. Email: enorton@northwestern.edu.

Supporting Information

Additional supporting information may be found in the online version of this article on the publisher's website:10.1002/rrq.467/suppinfo