A Meta-Analytic Evaluation of Naglieri Nonverbal Ability Test: Exploring Its Validity Evidence and Effectiveness in Equitably Identifying Gifted Students

Abstract

The Naglieri Nonverbal Ability Test (NNAT) was developed to more equitably identify students of color, as it advertises itself as a culture-fair measure. In this meta-analytic evaluation, we aimed to investigate (a) the generalizability of validity evidence of NNAT by checking its construct and criterion validity with other measures (part I) and (b) whether NNAT truly meets its goal to identify more culturally diverse students (part II). After reviewing 1,714 studies, a total of 29 studies met our criteria (59 effect sizes from 22 studies for part I and 7 effect sizes from 7 studies for part II). In part I, we investigated empirical evidence of validity of NNAT in relationship with different types of measures (overall effect size of *r* was 0.44); The results revealed that the correlation between NNAT and the achievement test results was 0.68, followed by the intelligence measures similar to NNAT (e.g., CogAT, OLSAT; r=0.31) and other alternative measures often used to identify gifted students (e.g., teacher-rating scale; r=0.20). The moderator analysis results showed high correlations between NNAT and other measures when Naglieri is an author of the study. In part II, although NNAT identified more students of color compared to other nonverbal tests (overall effect size of *RR* was 0.42), findings revealed that students of color remain underrepresented in gifted programs and services.

Keywords: Naglieri Nonverbal Ability Test (NNAT), Validity Check, Underrepresented Students, Identification

Introduction

The Naglieri Nonverbal Ability Test (NNAT; Naglieri, 2003, 2011, 2018) is a popular group administered intelligence tests in the United States (Gentry et al., 2020; Hodges et al., 2018). It was originally developed to address the need to identify underrepresented students for gifted services based on the premise that their under-identification was related to possible limited verbal and quantitative skills (Naglieri & Ford, 2003). NNAT does not require a child to read, write, or speak (Naglieri, 1997); rather, it requires examination of the relationships among the parts of the provided matrix, which is language-free (Naglieri & Ford, 2003).

In 2003, Naglieri and Ford conducted a large-scale study and suggested using NNAT might be beneficial in identifying students of diverse backgrounds, such as students who are learning to speak English (ELL). However, there have been arguments around usefulness of NNAT since 2005, when Lohman (2005) critically reviewed Naglieri and Ford's (2003) methods and sampling. Although Naglieri and Ford (2005) responded to Lohman's critics, limited validity evidence exists concerning whether NNAT is truly a good measure to use to identify underrepresented groups. Hodges et al. (2018) in their meta-analysis of identification methods found that nontraditional identification testing methods, including NNAT, do a better job at narrowing the proportional gap between well-represented and underrepresented populations after reviewing 85 effect sizes from 54 studies. They reported a risk ratio of nontraditional identification methods of 0.34 compared to 0.27 with traditional methods meaning that the probability of being identified as gifted within underserved populations is 34% of that for the well-represented group when using nontraditional methods whereas only 27% with traditional

identification methods. However, the authors suggested that the field still needs better identification methods to address inequity issues given these small risk ratios and differences between the types of identification measures. As such, debate exists concerning whether nontraditional methods are useful in identifying underrepresented populations for gifted programs and services.

Among many measures, we specifically selected NNAT for our study because it is widely used in schools, districts, and programs to identify gifted students with diverse backgrounds (Gentry et al., 2020; Hodges et al., 2018). To be specific, Gentry et al. (2020) reported NNAT was one of the top 5 recommended group intelligence tests after investigating the list of recommended tests for gifted identification across the state in the U.S. In addition, NNAT is completely based on the non-verbal tests whereas similar measures including non-verbal subtests (e.g., CogAT, WISC, OLSAT) also contain other subtests to measure individuals' verbal or quantitative ability. Lastly, NNAT has the most recent version (NNAT-III) of the assessment updated in 2018.

In this study, we used a meta-analytic technique to synthesize validity evidence of NNAT. In the first phase of this study, we examined the generalizability of construct and criterion validity of NNAT by synthesizing the reported correlations between NNAT and other criterion measures often used for identification (i.e., intelligence tests, academic achievement outcomes, other alternative measures used in identification of gifted students). Then, we examined whether the NNAT does help identify more ethnically diverse students (i.e. Black, Latinx, Native American) for gifted programs and services.

Definition of Giftedness and the Needs of Using Multiple Criteria in the Identification Process

The identification of students with gifts and talents has been widely studied in the field of gifted education; however, there is still no agreement on the definition of giftedness, resulting in different identification methods. The definition of giftedness is important as it affects the direction of the gifted programming and in deciding who gets served. According to the National Association for Gifted Children (NAGC, 2019), giftedness is defined as follows:

> "Students with gifts and talents perform - or have the capability to perform - at higher levels compared to others of the same age, experience, and environment in one or more domains. They require modification(s) to their educational experience(s) to learn and realize their potential. The student with gifts and talents:
>
> • Come from all racial, ethnic, and cultural populations, as well as all economic strata.
>
> • Require sufficient access to appropriate learning opportunities to realize their potential.
>
> • Can have learning and processing disorders that require specialized intervention and accommodation.
>
> • Need support and guidance to develop socially and emotionally as well as in their areas of talent. Require varied services based on their changing needs."

As such, the most recent definition of giftedness addressed by NAGC is inclusive, aiming to incorporate individuals with different backgrounds and needs. This aligns with the purpose of the development of NNAT as it was created to equitably identify students from diverse cultural backgrounds who were traditionally underserved.

According to McBee and Makel (2019), operational definitions exists in the field. For instance, they argued applying multiple criteria in the identification processes is not a formal or theoretical concept, but it serves as a general concept (McBee & Makel, 2019). The use of multiple criteria addresses different dimensions of giftedness such as achievement, motivation, and creativity in various domains (McBee et al., 2014; McBee & Makel, 2019; VanTassel-Baska, 2007). By considering different student characteristics and traits (McBee et al., 2014), the identification procedures allow more talented students to receive gifted programming and services (McBee et al., 2016; Peters & Gentry, 2012).

Researchers also pointed out that traditional identification methods may not be able to assess the exceptional abilities of students from different backgrounds, and they suggested using new and alternative methods to equitably identify students (Carman et al., 2018; Naglieri & Ford, 2003; 2005). NNAT is one of the intelligence measures developed to assess non-verbal reasoning and general problem-solving skills, which was a nontraditional way of identifying students compared to other intelligence tests measuring verbal and quantitative ability (Naglieri, 2003; 2011; 2018). NNAT, however, is still under the category of testing intelligence, and there are even more diverse measures (e.g., creativity test, teacher-rating scales) that are used alternatively or additionally to identify students who might be missed from general achievement test scores or standardized intelligence measures.

Underrepresentation Issues in Gifted Education

According to the NAGC (n.d.), giftedness occurs in every demographic group. Nevertheless, cultural biases against non-White and economically disadvantaged populations affect the representation of these students in gifted programs and services. The process of

identifying gifted and talented students from racially, culturally, economically, and linguistically diverse populations has long been a critical issue in the field of gifted education. Due to a heavy focus on traditional identification methods (e.g., using standardized measures such as achievement and intelligence tests), the proportion of students of color in gifted programs falls short of their percentage of the population (Gentry, 2009; Gentry et al., 2019; Naglieri & Ford, 2003, 2005; Peters & Gentry, 2012; Yoon & Gentry, 2009).

Various definitions of giftedness can be found in the field, and the definition adopted directly influences identification methods and determines the constituency of gifted education programs. The NAGC and the Council of State Directors of Programs for the Gifted (CSDPG) reported that states decide their own definition of giftedness, identification methods, and programming options (NAGC & CSDPG, 2015). According to Ford et al. (2016), tests that measure students' manifested achievements are criticized for ignoring the potential abilities of gifted and talented students from low-income families and who are Black, Latinx, or Indigenous. For instance, intelligence tests play a central role in the identification process in gifted education (Pfeiffer & Blei, 2008). However, whether or not the tests yield data for making valid inferences about student ability has often been questioned (Fletcher & Hattie, 2011; Lohman & Gambrell, 2012; Naglieri & Ford, 2005); especially, for students with diverse backgrounds. Moreover, sampling issues (e.g., predominantly White, Western children and adults) in the development of intelligence tests, lack of evidence for fairness among subpopulations, and the uncertainty of constructs measured by the tests has also been criticized (Lohman, 2005). Due to these issues, traditional identification methods may not be sufficient to equitably identify all students,

resulting in the underrepresentation of youth who are Black, Latinx, Indigenous, learning to speak English, or from low-income families.

Nonverbal Tests as Identification Measures: Non-traditional Ways of Identifying Students with Gifts and Talents

Because traditional intelligence tests may not accurately measure the intellectual ability of students from diverse ethnic, cultural, and socio-economic backgrounds (Carman & Taylor, 2010), nonverbal (non-traditional) intelligence tests have been purported to resolve this weakness of traditional tests. Most nonverbal tests are designed to eliminate bias resulting from language use and from cultural and socioeconomic differences among students (Ablard, & Brody, 1993; Carman et al., 2018; Matthews, 1988; Naglieri & Ford, 2003, 2005). In a study of 1,935 kindergarten students living in poverty, Kaya et al. (2017) found that their nonverbal IQ subscale scores were significantly greater than their verbal subscale scores. Furthermore, although intelligence tests can be administered in students' native languages, racial differences have still been found (Rossen et al., 2005). The findings indicated that students' backgrounds may affect their scores on the traditional language-based intelligence tests and supported the use of nonverbal tests for identifying gifted and talented learners from diverse economic, linguistic, and ethnic backgrounds (Lohman & Gambrell, 2012).

Nonverbal tests are constructed to avoid cultural bias or to be "culture fair" by using concrete objects or line drawings and require nonverbal responses, such as pointing or assembling parts of a puzzle (Lohman, 2005). They are designed to measure abilities to recognize analogies, classify, and form logical sequences using pictures and figures (Lassiter et al., 2001). Some researchers support the claim that nonverbal tests are free from cultural and

ethnic effects, therefore, not biased (Powers & Barkan, 1986). Brown and Day (2006) even argued the differences in outcomes among ethnicities and races may reflect stereotypes and do not originate from the test itself. For instance, the Cognitive Abilities Test (CogAT; Lohman, 2013) includes a nonverbal part for measuring reasoning, problem-solving skills, and fluid reasoning ability, and Raven's Progressive Matrices (RPM; Raven et al., 2003) is a nonverbal test that consists of 60 items measuring students' abstract reasoning. However, not enough evidence exists to conclude RPM would more accurately identify gifted students from all groups due, in part, to questions about the adequacy of its standardization (Mills et al., 1993).

Nonetheless, Naglieri and Ford (2003, 2005, 2015) have consistently argued that nonverbal measures are more appropriate for students from diverse backgrounds. To achieve educational equity, Naglieri and Ford (2015) contended, "it is essential to distinguish between students with high general ability on a nonverbal test, regardless of their verbal or quantitative skills, versus students who may be academically gifted" (p. 236). Adopting this perspective, some school administrators, who are concerned about equity issues, choose measures, such as CogAT (Lohman, 2013), Otis-Lennon School Ability (OLSAT; Otis & Lennon, 2003) test, NNAT (Naglieri, 2018), RPM (Raven et al., 2003), aiming to diminish the opportunity gap resulting from socio-cultural and economic differences (Lohman & Gambrell, 2012).

The Development and Controversies of NNAT

The NNAT (Naglieri, 2003, 2011, 2018) was developed and standardized as a nonverbal measure to assess the general and reasoning abilities of students from kindergarten through grade 12. According to Naglieri, the NNAT is ideal for use with a diverse student population, because it contains minimal use of language and verbal directions and does not require reading, writing,

or speaking; as a result can be considered culturally unbiased (Naglieri & Ford, 2003). As an alternative to language-based items, the NNAT uses geometric shapes and designs that allow for scoring unaffected by a child's primary language, education, and socio-economic background. Naglieri and Ford (2003) reported that White students and students from culturally diverse backgrounds perform similarly on the NNAT. They also claimed the NNAT might be a way to offset the effects of poverty among students and provide a solution to the underrepresentation in gifted programs of children who are Black or Latinx (indigenous people were not addressed in this study) due to non-significant differences among racial groups (Naglieri & Ford, 2005). Thus, NNAT has been administered as a part of the admission process for gifted and talented programs widely in the U.S (Hodges et al., 2018).

However, Lohman et al. (2008) found that non-ELL students still score greater than ELL students on the NNAT, leading them to conclude that "these differences are congruent with the conclusion that nonverbal tests do not see through the veneer of culture, education, or language development" (p. 290). This finding is similar to those from other studies (e.g., Carman & Taylor, 2010; Lohman, 2005), which also concluded the NNAT does not effectively identify students from diverse backgrounds. To be specific, Lohman (2005) argued the claim of the NNAT identifying equal proportions of high-scoring White, Black, and Latinx students is implausible and not supported by the data presented by its advocates. He also criticized the validation procedure carried by Naglieri and Ford (2003), because their sampling did not represent the demographics of the U.S. schools. Later, Naglieri and Ford (2005) refuted this claim in their study, "It was not our intention to provide samples that were representative, but rather to compare the three large groups of students who were similar in composition" (p. 33).

Similarly, Carman and Taylor (2010) argued the researchers should have controlled for multiple demographic variables, including income status for their results. As such, the research on the validity of the NNAT data raises questions about the use of this instrument for identifying gifted students with diverse backgrounds. Therefore, it is important to synthesize and evaluate existing findings from multiple empirical studies of NNAT to better understand the psychometric properties of this test in relation to its use for student identification for gifted programs.

Components and Psychometric Properties of NNAT

The first version of NNAT was an individually-administered, paper-pencil, assessment to measure general nonverbal ability in children (Naglieri, 2003). This version consisted of 72 items and required 25-30 minutes average testing time for students ages five through17. The NNAT-II provided updated normative data used in assessing the scores of students across the U.S (Naglieri, 2011). Although the second version of the test consisted of an entirely new set of items, no difference existed in the question types, scoring methods, score flow, the methods used to derive different types of scores, or the potential use and interpretation of the scores when compared with the NNAT-I. The age range was expanded to four through 18 years, the number of the items was reduced to 48, and online administration was available. The third edition of the NNAT, the NNAT-III, was launched in 2016. NNAT-III had new content based on recent normative data, and new administration options for online was offered as a main change (Naglieri, 2018). Like the NNAT-II, the NNAT-III consists of 48 items with 30 minutes online or paper/pencil administration for students ages four through18 years or grades Pre-K through 12.

More specifically, the NNAT-III contains four question types including pattern completion, reasoning by analogy, serial reasoning, and spatial visualization. It has four forms (A, B, C, and D) designed for students in kindergarten, grade 1, grade 2, and grades 3 and 4, respectively, and three forms (E, F, and G) for students in grades 5 and 6; 7 through 9; and 10 through 12 respectively. Each form consists of 48 items that vary in item difficulties and structures, with items presented in approximate order of difficulty. The NNAT-III provides multiple scores including raw scores, scaled scores, normative scores, the Naglieri Ability Index (NAI), percentile ranks, stanines, and Normal Curve Equivalents (NCEs), and each of these score types provides different information for users. According to the test manual (Naglieri, 2018), stratified random sampling was used for standardized purposes and the final normative data with full standardized sample contained a mean score of 100.2, with a standard deviation of 15.8, and a range from 40 to 160.

Reliability Evidence of NNAT

The test manuals reported relatively high-reliability evidence across all versions of the NNAT (Naglieri, 2003, 2011, 2018). It indicated the correlations between NAI scores on the NNAT-I and NNAT-II, and NNAT-III and NNAT-II, were 0.77 to 0.79 and 0.73 to 0.79, respectively. IRT-based reliability estimates were calculated for an online format (n=425) and paper format (n=480), and the results indicated that scores on the alternate forms were highly correlated, with an average correlation of 0.79 for both studies (i.e., Naglieri, 2003, 2011). According to the third manual (Naglieri, 2018), NNAT-III has high internal consistency estimates of the data across the grade levels. The internal consistency estimates for kindergarten through fourth grade data ranged from 0.80 to 0.88. For the data from students in grades 5

through 12, the manual reported alpha reliability estimates ranging from 0.81 to 0.89 and odds-even reliability estimates ranging from 0.82 to 0.90. Overall, the data from the technical manuals revealed that NNAT is consistent and reliable among its versions, formats, and grade levels.

Validity Evidence of NNAT

The NNAT manuals also provide some evidence of its validity. The first manual (Naglieri, 2003) reported correlations of NNAT-I score with different ability and achievement tests including RPM (Raven et al., 1998; r=0.78), Test of Nonverbal Intelligence-3 (TONI-3; Brown et al., 1997; r=0.63), Wechsler Intelligence Scale for Children-IV (WISC-IV; Wechsler, 2003; r=0.62), and Wechsler Individual Achievement Test-II (WIAT–II; The psychological Corporation, 2002; r=0.55). Based on these moderate to high correlation estimates, Naglieri (2003) commented that previous intelligence tests measured similar constructs with NNAT in terms of construct validity. Naglieri (2003) also included separate validity results for examinees, classified as different ethnic groups, gifted or talented, having an intellectual disability, having a learning disability, language or hearing impairment, or who spoke English as a second language. The second manual (Naglieri, 2011) reported correlations between NNAT-II and Wechsler Nonverbal Scale of Ability (WVN; Wechsler & Naglieri, 2006) subsets (range of 0.58-0.74), OLSAT-8 (Otis & Lennon, 2003; range of 0.53-0.69), and Stanford-10 (Pearson, n.d.; range of 0.51-0.74). Similarly, the third manual (Naglieri, 2018) reported an adjusted correlation between NNAT-III and the OLSAT-8 (Otis & Lennon, 2003) with a range of 0.17-0.64. The findings from the manuals indicate all versions of NNAT show moderate to high correlation with other measures of intelligence, providing evidence of strong criterion-related validity.

Purpose of the Study

Although NNAT was designed for identifying diverse populations, controversary exists concerning whether the NNAT supports its claim to be an equitable and culturally-fair identification measure for underserved populations (Carman & Taylor, 2010; Lohman, 2005; Lohman et al., 2008). Therefore, we conducted a meta-analytic evaluation to synthesize: (a) evidence for supporting construct and criterion validity by exploring the extent to which the NNAT is related to other measures, and (b) evidence for equity by exploring whether NNAT truly does identify equitable proportions of students of color. Specific research questions are as follows:

1. What is the relationship among NNAT and other measures used in identifying gifted students?
2. To what extent does NNAT equitably identify students from underrepresented populations for gifted programs and services compared to well-represented populations in terms of proportional identification rates?
3. To what extent do publication types, NNAT versions, authorship, and measurement type moderate the correlation between NNAT and other identification methods as well as the proportional representation of gifted students between well-represented and underrepresented groups?

Method

Meta-analytic investigation on validity

The concept of validity generalization (VG) was introduced in psychology in late 1970s as organizational psychologists found that validity evidence reported differs by sample and by situation even with the same instrument. For instance, Hunter and Schmidt (1977) used a meta-

analytic technique to demonstrate the generalizability of the construct-related validity evidence across situations. Since its introduction, the application of meta-analysis to psychometric investigation has been observed in a variety of research disciplines such as financial literacy (Fernandes et al., 2012), clinical psychology (Gerlsma et al., 1990), and vocational behavior (Van Rooy & Viswesvaran, 2004). As such, we used a unique type of meta-analysis to investigate the psychometric properties of the NNAT.

In terms of validity, there are four main types of validity: construct, content, face, and criterion validity (Sim & Arnell, 1993). We used the concept of construct and criterion validity in this study to explore the validity evidence of NNAT. Construct validity is about ensuring whether the method of measurement matches the construct (e.g., intelligence, giftedness) of what researchers aimed to measure (Sim & Arnell, 1993). In this study, intelligence measures served as convergent validity (i.e., between scales designed to measure the same construct), and standardized achievement test scores measures served as divergent validity (i.e., between scales designed to measure different constructs) were used to investigate construct validity (Canivez & Rains, 2002). Criterion validity evaluates how closely the results of the measurement corresponds to the results of a different measurement or test (Sim & Arnell, 1993). Alternative measures that are neither intelligence test nor achievement test results but frequently used in identifying gifted students in the field were categorized under the criterion-related validity.

Study Search Processes

The target population of the study included any empirical quantitative study that reported the correlation between the score produced with any version of NNAT and score(s) on a criterion measure. During August and September in 2019, four steps were used to search the literature for

studies: (a) Searches of multiple databases through Academic Search Premier (e.g., PsycInfo, ERIC, Education Full Text, Education Source, and ProQuest); (b) Searches of seven major gifted education journals (i.e., *Gifted Child Quarterly, Roeper Review, Journal for the Education of the Gifted, Gifted and Talented International, Gifted Education International, High Ability Studies, Journal of Advanced Academics*); (c) Inspection of NNAT technical manuals (Versions 1, 2, and 3); (d) Additional search from references of each report and with *Google Scholar* search engine. Since the study focuses on one single instrument, NNAT, keywords used in the searches were limited to "Naglieri Nonverbal Ability Test," "NNAT nonverbal," and "NNAT Intelligence".

Inclusion and Exclusion Criteria

The following criteria were used to include studies in or exclude studies from this meta-analysis: (a) No specific year was selected. Although the first manual was officially published and distributed in 2003-2004, Naglieri began publishing his works in 1980s. Therefore, the authors of the studies conducted before 2003 may already have had access to the NNAT measure (e.g., Martin (1996)); (b) Both published and unpublished empirical quantitative studies regardless of publication medium was considered for inclusion in this study (e.g., peer-reviewed journal articles, thesis/dissertations, technical manuals); (c) Language was restricted to studies published in English, including those reported outside of the U.S; (d) Empirical quantitative studies that included sufficient quantitative information to calculate target effect sizes (e.g., sample size, means and standard deviations, correlation coefficients) were selected (Cooper, 1998).

In addition to these criteria, the following criterion was applied to each part of the meta-analysis: For part I, we only considered studies that reported correlations or indirect information

so that we could calculate correlations between NNAT and a criterion measure or method that used for identifying gifted students. For instance, we did not include the studies in which the authors explored the relationship between NNAT and the measure of diagnosing one's ADHD or personality type. We included studies that used ability measures with construct validity evidence or standardized academic measures as criterion measures. However, studies that reported correlation of NNAT between teacher- or school-developed test results, school GPA, or unstandardized or locally normed measures were excluded. For part II, we only included studies that reported ethnicity information for both gifted and non-gifted students to calculate risk ratio, as described in the Effect Size Calculation section that follows. If the studies included the composition of the general population from which we could calculate gifted and non-gifted students' proportion, we included them as well. Racial categories were used to create an underrepresented group (i.e., Black, Latinx, Native American) and a well-represented group (i.e., White, Asian). Studies that had partial information about race, such as information about only White and Latinx students, were also included in the study.

If the study in question fit both parts of our study, we coded the identified study to both parts of the analyses. A total of four studies (i.e., Bracken & Brown, 2008; Edmonds, 2016; Giessman et al., 2013; Lewis et al., 2007) that reported correlation and risk ratio were used in both parts of this study.

The initial electronic search through *PsycInfo, ERIC, Education Full Text, Education Source, and ProQuest* yielded 1,714 studies related to NNAT measures. After reviewing the abstracts, methods, and results sections, 43 studies remained in our study pool for further evaluation. The manual search of seven gifted education journals yielded five more studies; four

technical manuals (i.e., NNAT-I, NNAT-II, NNAT-III.1, and NNAT-III.2) were included, and the *Google Scholar* search engine and reference checks added 12 more studies. A total of 64 studies were identified as potential articles for inclusion after the initial screening. After further evaluation of each individual study, 33 studies were eliminated from part I, and 51 studies were eliminated from part II due to insufficient quantitative information, leaving 31 studies for part I analyses and 13 studies for part II analyses. These quantitative studies were then thoroughly examined to determine whether we could extract the effect sizes that we needed. Finally, a total of 22 and seven studies for part I and part II, respectively, met the criteria for inclusion. Part I (k=22) included eight journal articles, 11 theses or dissertations, three technical manuals; and a total of 59 effect sizes. Part II (k=7) included six journal articles, one dissertation, and a total of seven effect sizes (see Figure 1 for selection pathway of the studies included in this research).

Coding of the Studies and Study Characteristics

For the two different parts of this meta-analytic evaluation, the research team developed two different coding sheets. The second and fifth authors were a team for the part I and the third and fourth authors work as a team for part II. If each team members disagreed with one another or were unsure about the data, the first author who was involved in the whole processes reviewed the issues to make a final decision. The sixth and seventh authors monitored and provided guidance to the team to conduct analyses. After the research team agreed upon finalized coding sheets, we included study characteristics (e.g., study year, author, sample size) for both parts of the study and added the measurement information for part I and ethnicity information for part II to calculate each effect size. Each study was coded with the name of the first author and the year of publication. The type of publication was coded as a journal article, dissertation, or technical

report to later investigate publication bias. In addition, studies were coded whether or not Naglieri, who developed NNAT, authored the study; thus, enabling us to determine whether authorship was related to findings. As the NNAT has been updated twice since the first edition, the research team coded studies by NNAT version used in the studies to create a potential moderator for the meta-analysis.

Correlation between Measures and Measurement Type

This part includes construct and criterion related validity as measured by correlations between the NNAT and other measures used in the study. Convergent construct validity was investigated by correlating NNAT with other intelligence tests such as CogAT (Lohman, 2013), TONI (Brown et al., 2010), RPM (Raven et al., 2003), and OLSAT (Otis & Lennon, 2003). The relationship between the NNAT and students' academic achievement (e.g., SAT) was involved as a measure of divergent construct validity. Finally, the NNAT was correlated with alternative measures such as teacher recommendation scores used in identifying students with gifts and talents to investigate criterion validity. Table 1 includes the names and types of the measurement used in this study.

Ethnicity Information of Total and Gifted Students

For part II of the study, gifted and non-gifted sample sizes were coded as frequency counts and percentages by ethnicity (i.e., Asian, Black, Latinx, Multiracial, Native American, and White). Some of the entries were left blank if the authors did not provide any information about certain ethnicities. We excluded students with multiracial background for the analysis in this study, because we grouped White and Asian students as well-represented and Black, Latinx, and Native American as underrepresented. Although each ethnic group has different

characteristics, we placed them into two groups, well-represented and underrepresented, based on the literature in the field (Gentry et al., 2019; Hodges et al., 2018; Yoon & Gentry, 2009).

Effect Size Calculation

Effect size shows the strength, magnitude, and direction of a relationship among variables (Berkeljon & Baldwin, 2014). In part I, we used Pearson's correlation coefficient (*r*) as an effect size, because we were interested in finding the relationship between the NNAT and the other measures. We did not include the formula how to calculate *r* in part I because the studies reported *r* value as a basic statistical information which we can directly retrieve. To perform the analyses, we transformed *r* to Fisher's *z* to standardize the value based on the pooled, weighted standard deviation, so the transformed sampling distribution would follow a normal distribution. After running the analysis, we transformed the summary values back to *r* for representation and interpretation (Borenstein et al., 2011).

In part II, we used risk ratio (*RR*), which is appropriate for dichotomous outcomes, to compare the rates of underrepresented and well-represented groups in gifted programs (Hodges et al., 2018). Black, Latinx, and Native American students were included in the underrepresented group, which served as the focal group, and the well-represented group included White and Asian students, which served as the reference group. In this study, the overall effect size of *RR* is defined as:

$$RiskRatio(RR)=\frac{Proportion\ of\ Underrepresented\ Students\ Identified\ in\ the\ Gifted\ Program}{Proportion\ of\ Well-represen\quad Students\ Identified\ in\ the\ Gifted\ Program}$$

To interpret, an *RR* of 0.5 indicates that the probability of being identified as a focal group is half of that for the reference group (Borenstein et al., 2011). This means $RR = 1$ indicates that the probability of being identified as gifted among underrepresented group is as same as that for the

well-represented group; we expected to have higher RR close to one if NNAT is truly effective in identifying underserved populations. After calculating *RR*, we then transformed *RR* into a log risk ratio (*LRR*) to standardize by maintaining a symmetric distribution of effect sizes (Borenstein et al., 2011). The standard error (*SE)* of the *LRR* was also recalculated with the formula of (Hodges et al., 2018):

$$SE_{LRRi}=\sqrt{\frac{1}{GiftedURi}-\frac{1}{URi}+\frac{1}{GiftedWRi}-\frac{1}{WRi}}$$

UR_i is the number of underrepresented students in study i, $GiftedUR_i$ is the number of underrepresented students who were identified as gifted, WR_i is the number of well-represented students in the same study, and $GiftedWR_i$ is the number of well-represented students who were identified as gifted.

Handling Multiple Effect Sizes within Studies

Several methods are suggested to handle dependent effect sizes within a study to reduce biased parameter estimates (e.g., averaging multiple effect sizes within a study, selecting one representative effect size, and using shift unit of analysis), and we averaged dependent effect sizes to obtain synthetic effect size representing the study (Cooper, 2010; Sutton et al., 2000). When the sample sizes were different but dependent, we used weights to compute the weighted average.

In our study, we averaged the correlation when the dependent sample was tested with the same categorical measures (e.g., CogAT and OLSAT, which are both intelligence tests) or if the data provided was only subtest results from the same measure (e.g., reading and math sub-scores were separately provided within Stanford 10, not as a total). However, if the study included multiple measures from different categories (i.e., intelligence tests, achievement tests, and

alternative measures), we randomly selected one from each study. For instance, Lewis et al. (2008) included intelligence test and achievement test results: Thus, we decided the study to be allocated randomly to any one category since the samples are dependent. Five studies in total (i.e., Lindsey (2013); Lewis et al. (2008); Humble (2018); Naglieri (2003); and Naglieri (2011)) had multiple results from different measurement categories, and they were randomly assigned to one category. Studies that used the same or different measures with different samples within the study were coded as having different effect sizes since they are independent.

Data Analyses

The random-effects model was chosen as a methodological framework for this meta-analysis, to reflect the expected variation among studies (Borenstein et al., 2011). Whereas fixed-effects analysis assumes the true effect size is the same in all studies, random-effects analysis reflects the variation among studies. To be specific, Borenstein et al. (2010) stated two conditions need to be met to use the fixed-effect model, "First, there is good reason to believe that all the studies are functionally identical. Second, our [the] goal is to compute the common effect size, which would not be generalized beyond the (narrowly defined) population included in the analysis" (p. 105). However, they argued that the fixed-effect assumption is often implausible in many systematic reviews (Borenstein et al., 2010). Thus, the random-effects model is often used for meta-analysis of social and clinical studies as it "explicitly accounts for the heterogeneity of studies through a statistical parameter representing the inter-study variation" (DerSimonian & Kacker, 2007, p. 105).

Heterogeneity and Moderator Analyses

A heterogeneity analysis examines the amount of variation in the retrieved effect sizes among studies beyond sampling errors. The Q statistics follow a chi-square distribution with the degrees of freedom of k-1. If the value exceeds a critical value, it indicates the hypothesis of homogeneity is rejected (Ellis, 2010). An I^2 statistic, which describes the percentage of variation across studies due to systematic heterogeneity rather than chance in total observed variation, and τ^2, which estimates the amount of the between-study variance in a random-effects meta-analysis, were also used to determine the heterogeneity of the effect sizes (Higgins & Thompson, 2002). The variation in effect sizes among subgroups can be grouped by certain factors as contextual moderators. We used meta-analysis of variance (meta-ANOVA) to identify potential moderators that separately explained the differences among defined subgroups. Three factors were used as moderators to explore the effect size variation: (a) publication type (i.e., journal article, thesis/dissertation, technical report), (b) NNAT version (I, II, III), (c) whether Naglieri was the author of the publication (yes or no). We also used the measurement type (i.e., intelligence tests, achievement tests, alternative measures) as an additional moderator for part I of the study.

Sensitivity Analysis

As studies with null or contradictory results are less likely to be published compared to those with significant results, publication bias needs to be explored (Cooper et al., 2009). We created a funnel plot for a visual inspection of potential publication bias. In the absence of bias and between-study heterogeneity, the scatter resembles a symmetrical inverted funnel, and the effect estimates from small sample size studies scatter more widely at the bottom with the spread narrowing among larger sample size studies (Sterne et al., 2011).

Results

Two separate meta-analytic evaluation were conducted using JASP software to investigate NNAT's construct and criterion validity, as well as whether it meets the goal of identifying diverse students. Both analyses followed four steps: Calculating effect sizes for each study, calculating an overall effect size, conducting homogeneity analysis, and analyzing moderators if appropriate.

Part I: Construct and Criterion Validity Generalization of the NNAT - Correlation Between the NNAT and Other Measures

Table 2 summarizes the characteristics of studies included in the meta-analytic evaluation and their effect sizes. The overall average correlation based on the random-effects model was 0.44 (k=59) with a *SE* of 0.03 and a 95% Confidence Interval (CI) from 0.39 to 0.49. This indicates that the overall correlation between the NNAT and other measures that are frequently used in the process of identifying gifted students have a moderate relationship (Cohen, 1988). However, as shown in Figure 2 with a graphic display of the variation in effect sizes, the homogeneity results indicated that significant variation exists among retrieved correlations, $Q(58)=209.39$, $p<.01$, $\tau^2=.05$, $I^2=97.29$, indicating that the heterogeneity of effect sizes is due to systematic between-study variance.

The assessment of publication bias for part I of the NNAT's validity was performed, and the associated funnel plot is provided in Figure 3. Although the majority of the studies were on the top with low *SE,* which means high precision appeared symmetrically around the mean effect line, several studies (approximately 10) were asymmetrically located in the left part of the funnel. This might indicate the presence of publication bias. A fail-safe analysis was conducted and revealed that 138,085 studies are needed for the null result to be accepted. Trim-and-fill

analysis also suggested that the combined effects became 0.52 [CI: 0.45, 0.58], indicating only 25% of variance is common. This implies that there may be a risk for publication bias with the current sampled studies, and we might underestimate the average correlation between the NNAT and other criterion measures. However, the underestimation due to publication bias seems to be small. For the moderator analysis, we had four moderators, including the NNAT version, authorship (whether Naglieri is author or not), publication type, and measurement type (see Table 3). In terms of the NNAT versions, NNAT I, NNAT II, and NNAT III have the correlations of 0.39 (CI [0.31, 0.47], *SE*=0.05), 0.52 (CI [0.29, 0.75], *SE*=0.08), and 0.59 (CI [-0.01, 1.10], *SE*=0.27), respectively. Further, the overall $Q(2)$ value was nonsignificant (3.38, p=0.18), meaning the version of the NNAT version did not influence effect size differences among studies. Thus, the version of NNAT was not a moderator. It is important to note, however, there was only one study using NNAT-III, which may have affected the results.

In examining the moderator effect of authorship, the Q value was statistically significant ($Q(1) = 33.65$, $p<.001$) indicating that authorship influenced effect sizes. When Naglieri was the author of the studies, the effect size was 0.58 (CI [0.50, 0.88], *SE*=0.07); whereas when others authored the studies, the effect size was 0.32 (CI [0.24, 0.39], *SE*=0.04). The correlation almost doubled when Naglieri was the author of the studies, which means that the strength of construct and criterion validity evidence differ based on which authors reported the results. In Naglieri and his co-authors tended to report stronger validity evidence than did other authors.

The effect sizes of measurement types differed; intelligence tests, academic achievement tests, and alternative measures were 0.31 (CI [0.25, 0.38], *SE*=0.04), 0.68 (CI [0.52, 0.83], *SE*=0.05), and 0.22 (CI [-0.02, 0.44], *SE*=0.09), respectively. The $Q(2)$ value of 65.08 was

statistically different from zero (p<.001), indicating measurement type was also a significant moderator. A small to moderate correlation was found between the NNAT and other intelligence tests; whereas, academic achievement test results were moderately to strongly correlated with the NNAT. The effect size between the NNAT and alternative measures was small, implying the alternative measures used in identifying gifted students measure different characteristics of giftedness from the NNAT, which supports the claim of multiple criteria in terms of the criterion validity evidence.

A moderator analysis for the publication types suggested a difference existed in correlation across the types of the publication (Q(2)=21.91, p<.001). The average effect sizes for thesis and dissertations, journal articles, technical reports were 0.38 (CI [0.24, 0.50], SE=0.08), 0.33 (CI [0.01, 0.64], SE=0.09), and 0.71 (CI [0.37, 1.00], SE=0.10), respectively. The effect size of the technical reports was more than double compared to those of the journal articles and the thesis/dissertations, demonstrating that the technical reports, which Naglieri authored, showed higher correlation results between the NNAT and other measures.

Part II: Purpose of the NNAT – Better Representation of Underserved Populations in Gifted Programs

With respect to the NNAT identifying underrepresented populations for gifted programs, the overall average effect size was 0.42 (SE=0.20; 95% CI is 0.28 to 0.63). This indicates that the probability of underserved populations being identified for the gifted program is about 42% compared to that of the probability of well-represented group (White and Asian students). Thus, the students in the focal group remain under-identified for gifted programs, even though the NNAT was applied. Figure 4 provides a graphic display of the effect sizes from the studies, and

the values are reported in Table 5 with lower and upper limits of the 95% CI for each effect size. The effect size and CI were recalculated by transforming the summary values from *LRR* back to *RR* for interpretation (Borenstein et al., 2011).

Homogeneity results indicated that the significant variation exists among retrieved *RR,* $Q(6)=19.72$, $p<.01$, $\tau^2=.06$, $I^2=87.31$, indicating that the heterogeneity of effect sizes are due to between-study variance. However, we decided not to run the moderator analysis because of the small number of the total effect sizes retrieved ($k=7$). The number of effect sizes allocated in each subgroup of the possible moderators (i.e., NNAT version, authorship, publication type) were too small to accurately measure the differences. For instance, only one study is a dissertation, whereas six other studies were from journal articles. Similarly, one study was from Naglieri, and the other six were from other authors. Additionally, two studies were based on NNAT version II; whereas, five studies were based on version I.

An assessment of publication bias for part of the NNAT's reliability check was performed (see Figure 5). Although the majority of the studies on the top with low SE appeared relatively symmetrical around the mean effect line, one study shows high SE, indicating low precision of the study. This might indicate the presence of publication bias, and the fail-safe analysis result revealed that 837 studies are additionally needed to accept the null result. The result from trim-and-fill analysis showed that after the trim-and-fill procedure, the combined effects became 0.42 [0.28, 0.62]. This implies that the original effect size of 0.44 was slightly overestimated. However, it is noteworthy the funnel plot may not reveal publication bias when the number of the studies is small. Although there is no fixed cut-off criteria, some scholars suggest fewer than 10 studies may result in low power to detect chance from real asymmetry

(Fagerland, 2015; Lau et al., 2006). However, we still included funnel plot because there is no consistent rule regarding whether to report it. For example, after systematically reviewing 47 meta-analytic papers in the medical field, Lau et al. (2006) found inconsistent reporting of funnel plots among small and larger samples.

Discussion

According to Naglieri (2003, 2011, 2018), the NNAT was developed to address underrepresentation of culturally and linguistically diverse students among students identified with gifts and talents. However, there have been arguments around its effectiveness in identifying underrepresented groups (Lohman, 2005; Lohman et al., 2008; Carman & Taylor, 2010). It was important to investigate the validity evidence of this instrument, because the NNAT is widely used in identifying students with gifts and talents in the U.S. (Hodges et al., 2018).

Overall, this meta-analytic evaluation synthesizes the evidence for the validity evidence of the NNAT. In part I, the analysis yielded an overall correlation effect size of 0.44, indicating a moderate relationship between the NNAT and other criterion measures. The effect size found in this study was smaller than those reported by Naglieri (2003, 2011) who reported correlations of about 0.70 between NNAT and other measures (e.g., intelligence and achievement tests).

As diverse types of measures were used to find the construct and criterion-related evidence for validity, we used the measurement types as a moderator, dividing them into three subcategories: intelligence tests, achievement tests, and alternative measures. Measurement types worked as a significant moderator in the present study. In terms of intelligence tests, it was interesting to find that the effect size was 0.31, which demonstrates weak correlation of the

NNAT with other intelligence tests. Further, this finding is the opposite of the results reported in the technical manuals (Naglieri, 2003, 2011, 2018). This implies the NNAT and other intelligence tests might measure different constructs. It also indicates the need for further investigation to establish more concrete evidence of the convergent construct validity of the NNAT, possibly providing clearer information about its subsamples test developers used to investigate the relationship between the NNAT and other measures.

Although the effect size between the NNAT and intelligence tests was small (0.31), the correlation between NNAT and students' academic achievement was stronger (r=0.68). This result shows that the NNAT better predicts students' academic achievement than their general intelligence. The data provided in the NNAT's technical manuals show that the correlation between the NNAT-I and WIAT-II was 0.55 (Naglieri, 2003), and the correlation between the NNAT-II and Stanford10 achievement test ranged from 0.51 to 0.74 (Naglieri, 2011) depending on students' age group. As such, the divergent construct validity in terms of achievement tests in relationship with the NNAT continuously showed moderate to strong correlations no matter the source of the correlation (e.g., technical reports, journal articles).

One interesting finding was the small effect size between the NNAT and alternative measures. Similar to the small effect sizes between the NNAT and other intelligence tests (r=0.31), other measures such as teacher recommendations had low correlation (r=0.20) with the NNAT. This implies that the alternative measures used in identifying gifted students do not overlap much with the NNAT, supporting the notion that alternative measures gauge different types of gifted characteristics and align with the multiple criteria claim in the field of gifted education. Many researchers do not believe that a single measure can identify giftedness because

of measurement errors and different student characteristics and traits (Callahan, 2012; McBee et al., 2014). In fact, one of the most important implications of this study could be encouraging educators to use multiple criteria and sources to make identification decision, which has been supported by many scholars and organizations in the field of gifted education (American Educational Research Association, American Psychological Association, & National Council on Measurement, 2014; McBee et al., 2014; NAGC, 2011; VanTassel-Baska, 2007). This suggestion aligns with McBee et al.'s (2014) argument that the "best practice in gifted and talented identification procedures involves making decisions on the basis of multiple measures" (p. 69). Using multiple criteria and multiple pathways in identification procedures will ensure that larger numbers of gifted students receive gifted programming and services (McBee et al., 2016). The NNAT, therefore, can be applied in the diagnostic process in conjunction with achievement test scores, alternative pathways, or even other intelligence tests. However, it is important to note due to our limited sample of seven studies under the category of alternative measures, our results are not definitive and should be interpreted with caution.

In terms of the moderators, these results also revealed that authorship of a study functioned as a moderator of the correlation found between the NNAT and other measures. That is, the effect size of studies of which Naglieri was the author was 0.58, almost twice that of studies published by other authors (r=0.32). These results indicate that Naglieri consistently reported larger effect sizes than did other researchers. This is in line with Lohman et al.'s (2008) assertion that there might have been validation and reliability problems in the development of the NNAT. Similarly, another conclusion from the study is that the publication type worked as a moderator. Although the effect sizes of thesis and dissertations and journal articles were similar,

the effect size of technical reports was more than double compared to effect sizes of the other types. That is, the technical reports written by Naglieri had larger effect sizes, indicating results were overestimated compared to those from other publication types. Unsurprisingly, this parallels the findings concerning the effect of Naglieri's authorship from the moderator analysis as well as Lohman's previous criticism of the NNAT (Lohman, 2005; Lohman et al., 2008).

With regard to checking the purpose of the NNAT, this study provided an effect size of 0.42 as an overall risk ratio, which indicated that the probability of identification of underrepresented students was 42% compared to well-represented groups. This demonstrates that students of color remain underserved in gifted programs and services, even when NNAT was used as an identification measure for those populations. As the purpose of developing the NNAT was to more equitably identify students of color, we expected an effect size closer to one as a risk ration of one indicates the probability of being identified as gifted among underrepresented group is as same as that for the well-represented group (Borenstein et al., 2011). The findings in this study provide evidence that the NNAT might not be the culturally-fair measure it claims to be, as it does not equitably identify students from underserved racial groups. However, it is important to note that given the small sample size (k=7) we retrieved and the studies included were all based on the prior version of NNAT (i.e., NNAT-I, NNAT-II), results for the NNAT-III might be different. However, as it is so new, those studies do not yet exist. In addition, the risk ratio effect size was still greater than what Hodges et al. (2018) reported in their meta-analysis in terms of combined effects of nonverbal intelligence tests (e.g., NNAT (Naglieri, 2018), CogAT (Lohman, 2013), RPM (Raven et al., 2003), Torrance Tests of Creative Thinking (TTCT; Torrance, 2006)). They found an overall effect size of 0.34 indicating the representation of

students of color in gifted programs and services was still low after using nonverbal intelligence tests. Compared to Hodges et al.'s (2018) findings, which included many different nonverbal intelligence measures, we conclude the NNAT might work slightly better as a measure for identifying students from diverse populations compared to the other nonverbal intelligence tests.

Although the NNAT itself shows slightly better results identifying students of color when compared to the combined effects of nonverbal intelligence tests, a lack of evidence still exists showing the NNAT has achieved its purpose of equitably identifying culturally diverse students with gifts and talents. For instance, even though we could not run the moderator analysis for part II of the study due to a limited number of studies, a close examination of the forest plot revealed that Naglieri and Ford's (2003) effect size of risk ratio was much larger than those of other studies. This again showed Naglieri's tendency to produce stronger effect sizes than those of other researchers. However, it is important to note that this observation is based on only one study (Naglieri & Ford, 2003).

In addition, it was unfortunate that we could not include any information from the technical manuals regarding the risk ratio, as they did not include enough evidence to calculate the effect size of the proportion of each race group. In the technical manuals of the NNAT (Naglieri, 2003, 2011), Black and Latinx samples were only compared with a matched White control group. Based on the *t*-test results, White students' scores were greater than Black and Hispanic students' scores. For example, NNAT-I technical manual reported a mean standard score for the Black sample of 92.7 (n=205, SD=13.3) and a mean standard score for the matched White sample of 101.1 (n=205, SD=11.2) with *Cohen's d* of 0.68 (t=7.13, p<.01). Similarly, the Latinx students' mean score (m=97.0, SD=12.2, n=163) was lower than the matched White

student sample (*m*=99.6, *SD*=8.3, *n*=163) with *Cohen's d* of 0.22 (*t*=2.07, *p*=0.04). Naglieri (2003) specified in his manual that "the difference between the NNAT-I scores between these populations were considerably less than is found with traditional IQ tests, suggesting that this nonverbal test has utility for fair assessment of these diverse populations" (p. 46). In this argument, he did not provide data concerning the magnitude of the differences for other measures. Although *Cohen's d* can be transformed to a risk ratio, the race proportions of students identified as gifted and the score differences have different characteristics. Therefore, this group difference information was not used in the meta-analytic evaluation. The NNAT-II manual (Naglieri, 2011) also provided mean differences among three race groups (Black, Latinx, and White); whereas, the NNAT-III manual (Naglieri, 2018) did not include any information regarding mean scores by race.

Given this, it was difficult for us to conclude that the NNAT is a cultural-fair test as reported by Naglieri. Earlier, Lohman (2005) indicated that sampling and method issues were found in the development of the NNAT, and he stated that the data did not match the U.S. school population, ethnic subgroups, and SES proportions.

Limitation and Implications for Future Research

This two-part, meta-analytic evaluation investigated the NNAT, and each part of the study has some limitations. In part I, variables were categorized as "intelligence tests," "academic achievement tests," or "alternative measures," as using multiple measures and multiple criteria have been recommended for identification for admission into gifted education programs (McBee et al., 2014), aligning with an inclusive definition of giftedness (NAGC, 2019). Excluding "alternative measures" from the part I of the study might have resulted in

stronger construct validity evidence, as only few studies were included under this category. Another limitation of this study involves multiple effect sizes from several studies (n=5), which comprise 75% of the total effect sizes analyzed for part I of this study. This could be considered as a drawback in terms of publication bias.

The limitations of part II of the study include the small number of empirical studies: Only seven studies and seven effect sizes met criteria for inclusion, and moderator analysis for part II was not done due to the small number of effect sizes. With more studies, findings of differential effects by authorship could have been investigated and perhaps corroborated the findings from part I. Furthermore, although we applied the random-effects model in part II based on its concept described in the method section, the findings (e.g., the estimation of between-study variance) are less precise and may have risk for bias due to small number of studies in the meta-analysis (Borenstein et al., 2010; Field, 2001). This requires a caution for the interpretation of the results even if the random-effects model is still appropriate (Borenstein et al., 2010). Because of this caveat, Borenstein et al. (2010) suggested three options researchers may want to choose, although they acknowledged each of them still has issues; (a) reporting separate effects rather than a summary effect, (b) conducting a fixed-effect analysis, and (c) performing a Bayesian meta-analysis so that the estimate of τ^2 is based on data from other sets of studies. Therefore, it is critical to evaluate the summary effect along with careful examination of individual effects reported in Figure 4 and Table 5. It is also important to replicate the analysis with larger number of effect sizes to enhance statistical conclusion validity in future studies. Grouping all underserved racial groups (Black, Latinx, and Native American students) together to dichotomously dividing well-represented and underrepresented groups was another limitation.

Investigating each racial group might have revealed different results as they have different characteristics, but sample sizes and consistent inclusion of different racial groups made this impossible.

The number of studies investigating whether the NNAT truly identified more of students of color was much smaller than expected. In addition, it was surprising to find published articles and technical manuals consistently did not contain sufficient quantitative information. Future researchers could explore the usefulness and effectiveness of the NNAT on identifying students from different racial groups and economic backgrounds. In this study we investigated well-represented and underrepresented racial groups, but we acknowledge other factors like income need to be examined. Additional research might address the use of multiple identification methods that include verbal and nonverbal intelligence tests compared to those using nonverbal intelligence tests alone. Moreover, as the NNAT targets a wide range of age group (Pre-K through grade 12), a possible moderator in future research might be students' age or their grade level.

The NNAT could benefit from further revisions. Although the NNAT identified slightly more students of color compared to other nonverbal intelligence tests found in Hodges et al.'s (2018) study, students of color remain underrepresented in gifted programs and services even when the NNAT is used for identification. As the largest effect sizes were reported from Naglieri's works including the NNAT technical reports, further investigation is needed. NNAT authors might reconsider their sampling and methods and re-investigate the reliability and validity evidence of this widely used instrument in the future. Given that other researchers may

use the NNAT to validate their own instruments, it is important the NNAT authors continuously update the results.

References

References marked with an asterisk are studies were included in this meta-analysis.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* American Educational Research Association.

*Arrambide, T. O. (2016). *The validity of three instruments in identifying academically gifted Hispanic kindergarten English language learners* [Doctoral dissertation, Sam Houston State University]. (Order No. 10182942). ProQuest Dissertations & Theses Global.

*Balboni, G., Naglieri, J. A., & Cubelli, R. (2010). Concurrent and predictive validity of the Raven Progressive Matrices and the Naglieri Nonverbal Ability Test. *Journal of Psychoeducational Assessment*, *28*(3), 222-235. https://doi.org/10.1177/0734282909343763

Berkeljon, A., & Baldwin, S. A. (2014). An introduction to meta-analysis for psychotherapy outcome research. In Lutz, W & Knox, S (Ed.), *Quantitative and qualitative methods in*

psychotherapy research (pp. 221-234). Routledge. https://doi.org/10.4324/9780203386071-14

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research synthesis methods, 1*(2), 97-111. https://doi.org/10.1002/jrsm.12

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. John Wiley & Sons.

*Botella, M., Fürst, G., Myszkowski, N., Storme, M., Pereira Da Costa, M., & Luminet, O. (2015). French validation of the overexcitability questionnaire 2: Psychometric properties and factorial structure. *Journal of Personality Assessment*, *97*(2), 209-220. https://doi.org/10.1080/00223891.2014.938750

*Bracken, B. A., & Brown, E. F. (2008). Early identification of high-ability students: Clinical assessment of behavior. *Journal for the Education of the Gifted*, *31*(4), 403-426. https://doi:10.4219/jeg-2008-794

Brown, R. P., & Day, E. A. (2006). The difference isn't black and white: Stereotype threat and the race gap on Raven's advanced progressive matrices. *Journal of Applied Psychology*, *91*(4), 979-985. https://doi.org/10.1037/0021-9010.91.4.979

Brown, S. W., Renzulli, J. S., Gubbins, E. J., Siegle, D., Zhang, W., & Chen, C. H. (2005). Assumptions underlying the identification of gifted and talented students. *Gifted Child Quarterly*, *49*(1), 68-79. https://doi:10.1177/001698620504900107

Brown, L., Sherbenou, R. J., & Johnsen, S. K. (1997). *Test of Nonverbal Intelligence* (3rd ed.). Pro_Ed.

Brown, L., Sherbenou R. J., & Johnsen, S. K. (2010). *Test of Nonverbal Intelligence* (4th ed.). Pearson.

*Brulles, D., Peters, S. J., & Saunders, R. (2012). School wide mathematics achievement within the gifted cluster grouping model. *Journal of Advanced Academics*, *23*(3), 200-216. https://doi:10.1177/1932202X12451439

Canivez, G. L., & Rains, J. D. (2002). Construct validity of the Adjustment Scales for Children and Adolescents and the Preschool and Kindergarten Behavior Scales: Convergent and divergent evidence. *Psychology in the Schools*, *39*(6), 621-633. https://doi.org/10.1002/pits.10063

*Carman, C. A., & Taylor, D. K. (2010). Socioeconomic status effects on using the Naglieri Nonveral Ability Test (NNAT) to identify the gifted/talented. *Gifted Child Quarterly*, *54*(2), 75-84. https://doi:10.1177/0016986209355976

Carman, C. A., Walther, C. A., & Bartsch, R. A. (2018). Using the cognitive abilities test (CogAT) 7 nonverbal battery to identify the gifted/talented: An investigation of demographic effects and norming plans. *Gifted Child Quarterly*, *62*(2), 193-209. https://doi.org/10.1177/0016986217752097

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates Publishers.

Cooper, H. (1998). *Synthesizing research: A guide for literature reviews* (3rd ed.). Sage.

Cooper, H. (2010). *Research synthesis and meta-analysis* (4th ed.). Sage.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation.

Davis, G. A., Rimm, S. B., & Siegle, D. B. (2013). *Education of the gifted and talented: Pearson new international edition.* Pearson.

DerSimonian, R., & Kacker, R. (2007). Random-effects model for meta-analysis of clinical trials: an update. *Contemporary clinical trials*, *28*(2), 105-114. https://doi.org/10.1016/j.cct.2006.04.004

*Edmonds, M. M. (2016). *Identifying gifted minorities using nonverbal tests: The cognitive abilities test, form 7, versus the Naglieri nonverbal ability test, second edition* (Order No. 3730349). [Doctoral dissertation, Regent University]. ProQuest Dissertations & Theses Global.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results.* Cambridge University Press.

Fagerland, M. W. (2015). Evidence-based medicine and systematic reviews. In P, Laake, H. B, Benestad, & B. R., Olsen (Eds.), *Research in Medical and Biological Sciences* (pp. 431-461). Elsevier Ltd. https://doi.org/10.1016/B978-0-12-799943-2.00012-4

Fernandes, D., Lynch Jr, G., & Netemeyer, R. (2012). A meta-analytic and psychometric investigation of the effect of financial literacy on downstream financial behaviors. *Advances in Consumer Research*, *40*, 1052.

Field, A. P. (2001). Meta-analysis of correlation coefficients: a Monte Carlo comparison of fixed-and random-effects methods. Psychological methods, 6(2), 161. https://doi.org/10.1037/1082-989X.6.2.161

Fletcher, R. B., & Hattie, J. (2011). *Intelligence and intelligence testing*. Routledge.

*Esquierdo, J. J. (2006). *Early identification of Hispanic English language learners for gifted and talented programs* (Order No. 3219154) [Doctoral dissertation, Texas A&M University]. ProQuest Dissertations & Theses Global.

Ford, D. Y., Wright, B. L., Washington, A., & Henfield, M. S. (2016). Access and equity denied: Key theories for school psychologists to consider when assessing Black and Hispanic students for gifted education. *School Psychology Forum: Research in Practice*, *10*(3), 265-277.

Gentry, M. (2009). Myth 11: A comprehensive continuum of gifted education and talent development services. *Gifted Child Quarterly*, *53*, 262-265. https://doi:10.1177/0016986209346937

Gentry, M., Gray, A., Whiting, G., Maeda, Y., & Pereira, N. (2019). *Gifted education in the United States: Laws, access, equity, and missingness across the country by locale, Title I school status, and race* [Executive summary]. Retrieved from https://www.education.purdue.edu/geri/new-publications/gifted-education-in-the-united-states/

Gentry, M., Desmet, O., Chowkase, A., & Lee, H. (2020, November 12-17). *Intelligence tests to identify students with gifts and talents: Still perpetuating inequity*. [Conference presentation]. 66th Annual Convention of the National Association for Gifted Children, Orlando, FL.

Gerlsma, C., Emmelkamp, P. M., & Arrindell, W. A. (1990). Anxiety, depression, and perception of early parenting: A meta-analysis. *Clinical Psychology Review*, *10*(3), 251-277. https://doi.org/10.1016/0272-7358(90)90062-F

*Giessman, J. A., Gambrell, J. L., & Stebbins, M. S. (2013). Minority performance on the Naglieri Nonverbal Ability Test, versus the Cognitive Abilities Test, form 6: One gifted program's experience. *Gifted Child Quarterly*, *57*(2), 101-109. https://doi:10.1177/0016986213477190

Higgins, J., Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539-1558. https://doi:10.1002/sim.1186

Hodges, J., Tay, J., Maeda, Y., & Gentry, M. (2018). A meta-analysis of gifted and talented identification practices. *Gifted Child Quarterly*, *62*(2), 147-174. https://doi:10.1177/0016986217752107

*Humble, S., Dixon, P., & Schagen, I. (2018). Assessing intellectual potential in Tanzanian children in poor areas of Dar es Salaam. *Assessment in Education: Principles, Policy & Practice*, *25*(4), 399-414. https://doi.org/10.1080/0969594X.2016.1194257

Kaya, F., Stough, L. M., & Juntune, J. (2017). Verbal and nonverbal intelligence scores within the context of poverty. *Gifted Education International*, *33*(3), 257-272. https://doi.org/10.1177/0261429416640332

Lassiter, K. S., Harrison, T. K., Matthews, T. D., & Bell, N. L. (2001). The validity of the comprehensive test of nonverbal intelligence as a measure of fluid intelligence. *Assessment*, *8*(1), 95-103. https://doi.org/10.1177/107319110100800109

Lau, J., Ioannidis, J. P., Terrin, N., Schmid, C. H., & Olkin, I. (2006). The case of the misleading funnel plot. *BMJ, 333*(7568), 597-600. https://doi.org/10.1136/bmj.333.7568.597

*Lewis, J. D., DeCamp-Fritson, S. S., Ramage, J. C., McFarland, M. A., & Archwamety, T. (2007). Selecting for ethnically diverse children who may be gifted using Raven's

Standard Progressive Matrices and Naglieri Nonverbal Abilities Test. *Multicultural Education*, *15*(1), 38-42.

*Lindsey, C. D. (2013). *An investigation of the combined assessments used as entrance criteria for a gifted English middle school program* (Order No. 3525542). [Doctoral dissertation, University of New Orleans]. ProQuest Dissertations & Theses Global.

Lohman, D. F. (2005). Review of Naglieri and Ford (2003): Does the Naglieri Nonverbal Ability Test identify equal proportions of high-scoring White, Black, and Hispanic students?. *Gifted Child Quarterly*, *49*(1), 19-28. https://doi:10.1177/001698620504900103

Lohman, D. F (2013). *Cognitive Ability Test form 7: Planning and implementation guide.* Riverside.

Lohman, D. F., & Gambrell, J. L. (2012). Using nonverbal tests to help identify academically talented children. *Journal of Psychoeducational Assessment*, *30*(1), 25-44. https://doi.org/10.1177/001698620504900103

Lohman, D. F., Gambrell, J., & Lakin, J. (2008). The commonality of extreme discrepancies in the ability profiles of academically gifted students. *Psychology Science*, *50*(2), 269.

*Lohman, D. F., Korb, K. A., & Lakin, J. M. (2008). Identifying academically gifted English-language learners using nonverbal tests: A comparison of the Raven, NNAT, and CogAT. *Gifted Child Quarterly*, *52*(4), 275-296. https://doi.org/10.1177/0016986208321808

*Mann, R. L. (2005). *The identification of gifted students with spatial strengths: An exploratory study* (Order No. 3180228). [Doctoral dissertation, University of Connecticut]. ProQuest Dissertations & Theses Global.

*Mann, E. L. (2008). Parental perceptions of mathematical talent. *Social Psychology of Education*, *11*(1), 43-57. https://doi.org/10.1007/s11218-007-9034-y

*Martin, A. (1996). *The performance of a multilingual South African sample on two measures of nonverbal assessment* (Order No. 9630930). [Doctoral dissertation, The Ohio State University]. Dissertations & Theses Global.

Matthews, D. J. (1988). Raven's Matrices in the identification of giftedness. *Roeper Review*, *10*(3), 159-162. https://doi.org/10.1080/02783198809553115

McBee, M. T., & Makel, M. C. (2019). The quantitative implications of definitions of giftedness. *AERA Open, 5*(1), 1-13. https://doi.org/10.1177/2332858419831007

McBee, M. T., Peters, S. J., & Miller, E. M. (2016). The impact of the nomination stage on gifted program identification: A comprehensive psychometric analysis. *Gifted Child Quarterly*, *60*(4), 258-278. https://doi:10.1177/0016986216656256

McBee, M. T., Peters, S. J., & Waterman, C. (2014). Combining scores in multiple-criteria assessment systems: The impact of combination rule. *Gifted Child Quarterly*, *58*(1)*,* 69-89. https://doi:10.1177/0016986213513794

Mills, C. J., Ablard, K. E., & Brody, L. E. (1993). The Raven's Progressive Matrices: Its usefulness for identifying gifted/talented students. *Roeper Review*, *15*(3), 183-186. Combining scores in multiple-criteria assessment systems: The impact of combination rule. https://doi: 10.1080/02783199309553500

*Naglieri, J. A. (2003). *Naglieri Nonverbal ability test: Individual administration manual.* Multi-Health Systems.

*Naglieri, J. A. (2011). *Naglieri Nonverbal ability test manual: Technical information and normative data* (2nd ed.). NCS Pearson.

*Naglieri, J. A. (2018). *Naglieri Nonverbal Ability Test Manual Levels A-D* (3rd ed.). Pearson.

*Naglieri, J. A., Booth, A. L., & Winsler, A. (2004). Comparison of Hispanic children with and without limited English proficiency on the Naglieri Nonverbal Ability Test. *Psychological assessment*, *16*(1), 81. https://doi.org/10.1037/1040-3590.16.1.81

*Naglieri, J. A., & Ford, D. Y. (2003). Addressing underrepresentation of gifted minority children using the Naglieri Nonverbal Ability Test (NNAT). *Gifted Child Quarterly*, *47*(2), 155-160. https://doi:10.1177/001698620304700206

Naglieri, J. A., & Ford, D. Y. (2005). Increasing minority children's participation in gifted classes using the NNAT: A response to Lohman. *Gifted Child Quarterly*, *49*(1), 29-36. https://doi:10.1177/001698620504900104

Naglieri, J. A., & Ford, D. Y. (2015). Misconceptions about the Naglieri Nonverbal Ability Test: A commentary of concerns and disagreements. *Roeper Review*, *37*(4), 234-240. https://doi.org/10.1080/02783193.2015.1077497

*Naglieri, J. A., & Ronning, M. E. (2000). The relationship between general ability using the Naglieri Nonverbal Ability Test (NNAT) and Stanford Achievement Test (SAT) reading achievement. *Journal of Psychoeducational Assessment*, *18*(3), 230-239. https://doi.org/10.1177/073428290001800303

National Association for Gifted Children. (n.d.*).* What is giftedness?. NAGC. http://www.nagc.org/resources-publications/resources/what-giftedness

National Association for Gifted Children (2011). Identifying and serving culturally and linguistically diverse gifted students. NAGC. http://www.nagc.org/sites/default/files/Position%20Statement/Identifying%20and%20Serving%20Culturally%20and%20Linguistically.pdf

National Association for Gifted Children. (2019). Position Statement. NAGC. https://www.nagc.org/sites/default/files/Position%20Statement/Definition%20of%20Giftedness%20%282019%29.pdf

National Association for Gifted Children & The Council of State Directors of Programs for the Gifted. (2015). 2014-2015 State of the states in gifted education: Policy and practice data. NAGC. https://www.nagc.org/sites/default/files/key%20reports/2014-2015%20State%20of%20the%20States%20(final).pdf

Otis, A. S., & Lennon, R. T. (2003). *Otis-Lennon School Ability Test* (8th ed.). Pearson.

Pearson (n.d.). *Stanford Achievement Test Series* (10th ed.). Author.

Peters, S. J., & Gentry, M. (2012). Group-specific norms and teacher-rating scales: Implications for underrepresentation. *Journal of Advanced Academics*, *23*(2), 125-144. https://doi.org/10.1177/1932202X12438717

Pfeiffer, S. I., & Blei, S. (2008). Gifted identification beyond the IQ test: Rating scales and other assessment procedures. In *Handbook of giftedness in children* (pp. 177-198). Springer. https://doi:10.1007/978-0-387-74401-8

Powers, S., Barkan, J. H., & Jones, P. B. (1986). Reliability of the standard progressive matrices test for Hispanic and Anglo-American children. *Perceptual and Motor Skills*, *62*(2), 348-350. https://doi.org/10.2466/pms.1986.62.2.348

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. Oxford Psychologists Press.

Raven, J., Raven, J. C., & Court, J. H. (2003). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. Harcourt Assessment.

*Rosado, J. I. (2009). *Validation of the Spanish version of the gifted rating scales* (Order No. 3340757). [Doctoral dissertation, Florida State University]. ProQuest Dissertations & Theses Global.

Rossen, E. A., Shearer, D. K., Penfield, R. D., & Kranzler, J. H. (2005). Validity of the comprehensive test of nonverbal intelligence (CTONI). *Journal of Psychoeducational Assessment*, *23*(2), 161-172. https://doi.org/10.1177/073428290502300205

*Runyon, L. (2010). *Identification of highly gifted 5- and 6-year-old children: Measures to predict academic achievement* [Master's Thesis, University of North Texas]. (Order No. 1485564). ProQuest Dissertations & Theses Global.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, *62*(5), 529. https://doi.org/10.1037/0021-9010.62.5.529

Sim, J., & Arnell, P. (1993). Measurement validity in physical therapy research. *Physical therapy*, *73*(2), 102-110. https://doi.org/10.1093/ptj/73.2.102

Sterne, J. A., Sutton, A. J., Ioannidis, J., Terrin, N., Jones, D. R., Lau, J., & Higgins, J. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *British Medical Journal*, *343*(7818), 302–307. https://doi.org/10.1136/bmj.d4002

Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A., Song, F. (2000). *Methods for meta-analysis in medical research*. Wiley.

The Psychological Corporation. (2002). *Wechsler Individual Achievement Test* (2nd ed.). Author.

Torrance, P. E. (2006). *Torrance Tests of Creative Thinking*. Scholastic Testing Service.

VanTassel-Baska, J. (2007). *Alternative assessments with gifted and talented students*. Prufrock Press.

Van Rooy, D. L., & Viswesvaran, C. (2004). Emotional intelligence: A meta-analytic investigation of predictive validity and nomological net. *Journal of vocational Behavior*, *65*(1), 71-95. https://doi.org/10.1016/S0001-8791(03)00076-9

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children* (4th ed.). The Psychological Corporation.

Wechsler, D. & Naglieri, J. A. (2006). *Wechsler Nonverbal Scale of Ability*. Pearson.

*Wills, A. J. (2013). *A study of the relationship between kindergarten nonverbal ability and third-grade reading achievement* (Order No. AAI3505945). [Doctoral dissertation, University of Missouri-St. Louis]. ProQuest Dissertations & Theses Global.

*Worthington, J. J. (2002). *A comparison of gifted identification methods using measures of achievement, ability, multiple intelligences, and teacher nominations* [Doctoral dissertation, University of San Francisco]. (Order No. 3048673). ProQuest Dissertations & Theses Global.

*Yang, J. (2016). *The influential factors of math achievement in mathematically promising English language learners* [Doctoral dissertation, St. John's University] (Order No. 3664191). ProQuest Dissertations & Theses Global.

Yoon, S. Y., & Gentry, M. (2009). Racial and ethnic representation in gifted programs: Current status of and implications for gifted Asian American students. *Gifted Child Quarterly*, *53*(2), 121-136. https://doi.org/10.1177/0016986208330564

Table 1
Part I Study Characteristics and Effect Sizes

Study	Sample (*N*)	NNAT Ver	Author: Naglieri	Publication Type	Measurement Used	Measurement Type	*r*	Fisher's *z*	*z* SE
Arrambide (2017)	121	2	No	Thesis/Dissertation	Teacher Recommendation Form	Alternative Methods	0.40	0.42	0.07
Balboni et al. (2010)	253	1	Yes	Journal Article	Math and reading comprehension tests	Academic Achievement	0.46	0.50	0.05
Botella et al. (2015)	474	1	No	Journal Article	Overexcitability Questionnaire 2 (Imagination)	Alternative Methods	0.04	0.04	0.05
Bracken et al. (2008)	456	1	No	Journal Article	Bracken Basic Concept Scale-Revised (BBCS-R)	Intelligent Test	0.34	0.35	0.05
Edmonds (2016)	11,680	2	No	Thesis/Dissertation	Virginia SOLs-math	Academic Achievement	0.51	0.56	0.07
Esquierdo (2006)	778	1	No	Thesis/Dissertation	Hispanic Bilingual Gifted Screening Instrument (HBGSI)	Alternative Methods	0.27	0.28	0.04
Giessman et al. (2013)	3,665	2	No	Journal Article	CogAT6	Intelligent Test	0.20	0.21	0.02
Giessman et al. (2013)	1,217	2	No	Journal Article	CogAT6	Intelligent Test	0.20	0.20	0.03
Giessman et al. (2013)	284	2	No	Journal Article	CogAT6	Intelligent Test	0.10	0.10	0.06
Giessman et al. (2013)	296	2	No	Journal Article	CogAT6	Intelligent Test	-0.05	-0.05	0.06
Giessman et al. (2013)	30	2	No	Journal Article	CogAT6	Intelligent Test	0.03	0.03	0.19

Giessman et al. (2013)	9	2	No	Journal Article	CogAT6	Intelligent Test	0.09	0.09	0.41
Giessman et al. (2013)	332	2	No	Journal Article	CogAT6	Intelligent Test	0.13	0.13	0.06
Humble et al. (2018)	1,857	2	No	Journal Article	GMADE 1 to 4 (Pearson) and the English reading test from the 'Single Word Reading Test' (National Foundation for Educational Research), Kiswahili tests	Academic Achievement	0.73	0.93	0.02
Lewis et al. (2007)	175	1	No	Journal Article	RPM	Intelligent Test	0.52	0.58	0.08
Lindsey (2013)	1,188	1	No	Thesis/Dissertation	CogAT	Intelligent Test	0.16	0.16	0.03
Lohman et al. (2008)	17	1	No	Journal Article	RPM	Intelligent Test	0.37	0.39	0.27
Lohman et al. (2008)	91	1	No	Journal Article	RPM	Intelligent Test	0.37	0.38	0.11
Lohman et al. (2008)	116	1	No	Journal Article	RPM	Intelligent Test	0.37	0.38	0.09
Lohman et al. (2008)	90	1	No	Journal Article	RPM	Intelligent Test	0.32	0.33	0.11
Lohman et al. (2008)	132	1	No	Journal Article	RPM	Intelligent Test	0.30	0.31	0.09
Lohman et al. (2008)	120	1	No	Journal Article	RPM	Intelligent Test	0.35	0.36	0.09

Lohman et al. (2008)	99	1	No	Journal Article	RPM	Intelligent Test	0.14	0.14	0.10
Lohman et al. (2008)	24	1	No	Journal Article	RPM	Intelligent Test	0.38	0.40	0.22
Lohman et al. (2008)	89	1	No	Journal Article	RPM	Intelligent Test	0.37	0.39	0.11
Lohman et al. (2008)	74	1	No	Journal Article	RPM	Intelligent Test	0.41	0.44	0.12
Lohman et al. (2008)	81	1	No	Journal Article	RPM	Intelligent Test	0.32	0.33	0.11
Lohman et al. (2008)	59	1	No	Journal Article	RPM	Intelligent Test	0.33	0.34	0.13
Lohman et al. (2008)	38	1	No	Journal Article	RPM	Intelligent Test	0.33	0.34	0.17
Lohman et al. (2008)	34	1	No	Journal Article	RPM	Intelligent Test	0.26	0.26	0.18
Mann (2005)	15	1	No	Thesis/Dissertation	WISC-IV block design subtest	Intelligent Test	0.30	0.31	0.29
Mann (2008)	19	1	No	Journal Article	Teacher Rating	Alternative Methods	0.09	0.09	0.25
Mann (2008)	7	1	No	Journal Article	Teacher Rating	Alternative Methods	-0.12	-0.12	0.50
Martin (1996)	400	1	No	Thesis/Dissertation	SAT	Academic Achievement	0.73	0.93	0.04
Naglieri & Ronning (2000)	22,620	1	Yes	Journal Article	SAT-9	Academic Achievement	0.52	0.58	0.17
Naglieri (2003)	39	1	Yes	Technical Manual	RPM	Intelligent Test	0.78	1.05	0.15
Naglieri (2003)	50	1	Yes	Technical Manual	TONI-3	Intelligent Test	0.63	0.74	0.02

Naglieri (2003)	150	1	Yes	Technical Manual	WISC-IV	Intelligent Test	0.34	0.35	0.07
Naglieri (2011)	106	2	Yes	Technical Manual	Stanford 10	Academic Achievement	0.70	0.87	0.04
Naglieri (2011)	281	2	Yes	Technical Manual	Stanford 10	Academic Achievement	0.67	0.80	0.04
Naglieri (2011)	307	2	Yes	Technical Manual	Stanford 10	Academic Achievement	0.62	0.72	0.06
Naglieri (2011)	150	2	Yes	Technical Manual	Stanford 10	Academic Achievement	0.65	0.78	0.05
Naglieri (2011)	179	2	Yes	Technical Manual	Stanford 10	Academic Achievement	0.63	0.74	0.07
Naglieri (2011)	118	2	Yes	Technical Manual	Stanford 10	Academic Achievement	0.59	0.68	0.05
Naglieri (2011)	175	2	Yes	Technical Manual	Stanford 10	Academic Achievement	0.61	0.71	0.06
Naglieri (2011)	139	2	Yes	Technical Manual	Stanford 10	Academic Achievement	0.59	0.67	0.06
Naglieri (2011)	162	2	Yes	Technical Manual	Stanford 10	Academic Achievement	0.66	0.79	0.03
Naglieri (2011)	595	2	Yes	Technical Manual	Stanford 10	Academic Achievement	0.65	0.78	0.56
Naglieri (2011)	165	2	Yes	Technical Manual	Stanford 10	Academic Achievement	0.72	0.91	0.05
Naglieri (2011)	175	2	Yes	Technical Manual	Stanford 10	Academic Achievement	0.70	0.85	0.05
Naglieri (2018)	366	3	Yes	Technical Manual	OLSAT 8	Intelligent Test	0.55	0.62	0.00
Naglieri et al. (2004)	148	1	Yes	Journal Article	SAT	Academic Achievement	0.76	1.00	0.05

Naglieri et al. (2004)	144	1	Yes	Journal Article	SAT	Academic Achievement	0.60	0.69	0.05
Rosado (2009)	421	1	No	Thesis/Dissertation	GRS-S Spanish	Alternative Methods	0.30	0.30	0.24
Runyon (2010)	16	1	No	Thesis/Dissertation	Stanford 10 Math	Academic Achievement	-0.30	-0.31	0.28
Wills (2013)	749	1	No	Thesis/Dissertation	MAP achievement scores in reading	Academic Achievement	0.50	0.55	0.04
Wills (2013)	795	1	No	Thesis/Dissertation	MAP achievement scores in reading	Academic Achievement	0.44	0.47	0.04
Worthington (2002)	231	1	No	Thesis/Dissertation	Teele Inventory of Multiple Intelligences (TIMI)	Alternative Methods	0.17	0.17	0.07
Yang (2016)	194	1	No	Thesis/Dissertation	Iowa Test of Basic Skills (ITBS)	Academic Achievement	0.44	0.47	0.07

Table 2

Pearson's Correlation Coefficient estimate (r) and its 95% CIs for Study Part I

		95% CI	
Citation	*r*	*LL*	*UL*
Arrambide (2017)	0.42	0.29	0.55
Balboni et al. (2010)	0.50	0.41	0.59
Botella et al. (2015)	0.04	-0.05	0.13
Bracken et al (2006)	0.35	0.26	0.45
Edmonds (2016)	0.56	0.43	0.69
Esquerdo (2006)	0.28	0.21	0.35
Giessman et al. (2013)	0.20	0.17	0.24
Giessman et al. (2013)	0.20	0.15	0.26
Giessman et al. (2013)	0.10	-0.02	0.22
Giessman et al. (2013)	-0.05	-0.16	0.06
Giessman et al. (2013)	0.02	-0.35	0.40
Giessman et al. (2013)	0.09	-0.71	0.89
Giessman et al. (2013)	0.13	0.02	0.24
Humble (2018)	0.93	0.88	0.97
Lewis et al. (2007)	0.58	0.43	0.72
Lindsey (2013)	0.16	0.10	0.22
Lohman et al. (2008)	0.39	-0.13	0.92
Lohman et al. (2008)	0.38	0.17	0.59
Lohman et al. (2008)	0.38	0.20	0.57
Lohman et al. (2008)	0.33	0.12	0.54
Lohman et al. (2008)	0.31	0.14	0.48
Lohman et al. (2008)	0.36	0.18	0.54
Lohman et al. (2008)	0.14	-0.06	0.34
Lohman et al. (2008)	0.40	-0.03	0.83
Lohman et al. (2008)	0.39	0.18	0.60
Lohman et al. (2008)	0.44	0.21	0.67
Lohman et al. (2008)	0.33	0.10	0.55
Lohman et al. (2008)	0.34	0.08	0.61
Lohman et al. (2008)	0.34	0.01	0.67
Lohman et al. (2008)	0.26	-0.09	0.61
Mann (2005)	0.31	-0.26	0.87
Mann (2008)	0.09	-0.40	0.58
Mann (2008)	-0.12	-1.10	0.86
Martin (1996)	0.92	0.86	0.99
Naglieri & Ronning (2000)	0.58	0.57	0.59
Naglieri (2003)	1.04	0.72	1.37
Naglieri (2003)	0.74	0.45	1.03
Naglieri (2003)	0.35	0.31	0.40
Naglieri (2011)	0.87	0.73	1.00

Naglieri (2011)	0.80	0.72	0.89
Naglieri (2011)	0.72	0.64	0.80
Naglieri (2011)	0.78	0.66	0.89
Naglieri (2011)	0.74	0.64	0.85
Naglieri (2011)	0.68	0.55	0.81
Naglieri (2011)	0.71	0.60	0.81
Naglieri (2011)	0.67	0.55	0.79
Naglieri (2011)	0.78	0.68	0.89
Naglieri (2011)	0.78	0.72	0.83
Naglieri (2011)	0.91	-0.19	2.01
Naglieri (2011)	0.85	0.74	0.95
Naglieri (2018)	0.62	0.52	0.72
Naglieri et al. (2004)	1.00	0.91	1.10
Naglieri et al. (2004)	0.68	0.59	0.78
Rosado (2009)	0.30	-0.17	0.78
Runyon (2010)	-0.30	-0.85	0.24
Wills (2013)	0.55	0.48	0.62
Wills (2013)	0.47	0.40	0.54
Worthington (2002)	0.17	0.04	0.30
Yang (2016)	0.47	0.33	0.61

Note. CI = confidence interval; *LL* = lower limit; *UL* = upper limit.

Table 3
Moderators of Effect Sizes for Study Part I

Moderator	k^a	Q	r^b	95 % CL	
				LL	*UL*
NNAT Version		3.38			
NNAT 1	36		0.39	0.31	0.47
NNAT 2	22		0.52	0.29	0.75
NNAT 3	1		0.59	-0.01	1.10
Author: Naglieri		33.65*			
No	39		0.32	0.24	0.39
Yes	20		0.58	0.50	0.88
Measurement Type		65.08*			
Intelligence Tests	29		0.31	0.25	0.38
Academic Achievements	23		0.68	0.52	0.83
Alternative Measures	7		0.20	-0.02	0.44
Publish Type		21.91*			
Thesis/Dissertation	12		0.38	0.24	0.50
Journal Article	31		0.33	0.01	0.64
Technical Report	16		0.71	0.37	1.00

Note. CI = confidence interval; *LL* = lower limit; *UL* = upper limit.
[a] Number of effect sizes included in the analysis. [b] Random-effects model.
* $p < .05$.

Table 4

Part II Study Characteristics and Effect Sizes

Study	Sample (*n*)	NNAT Version	Author: Naglieri	Publication Type	Gen/GT White %	Gen/GT Asian %	Gen/GT Black %	Gen/GT Latinx %	Gen/GT Native %	*RR*	*LRR*	*LRR* SE
Bracken & Brown (2008)	752	1	No	Journal Article	60.64/60.84	11.84/20.98	8.64/6.99	13.70/8.40	0.8/0.70	0.61	0.49	0.21
Brulles et al. (2012)	3,716	1	No	Journal Article	18.14/31.23	3.01/7.22	7.16/5.78	69.89/54.15	1.80/1.60	0.43	0.84	0.08
Carman & Taylor (2010)	2,072	1	No	Journal Article	61.10/67.02	9.80/15.49	6.40/3.54	22.20/13.62	0.50/0.33	0.52	0.65	0.12
Edmonds (2016)	12,669	2	No	Thesis/ Dissertation	33.0/48.62	8.10/13.42	20.40/11.60	32.70/19.62	-	0.38	0.97	0.06
Giessman et al. (2013)	4,035	2	No	Journal Article	64.0/71.58	5.0/18.52	20.50/2.30	5.0/1.70	1/0	0.12	2.12	0.25
Lewis et al. (2007)	175	1	No	Journal Article	58.29/83.33	-	-	41.71/16.67	-	0.28	1.27	0.76
Naglieri & Ford (2003)	18,995	1	Yes	Journal Article	74.44/77.39	-	15.07/13.25	10.48/9.40	-	0.85	0.16	0.05

Note. Gen/GT indicates General Population/Identified Population of Gifted and Talented Program

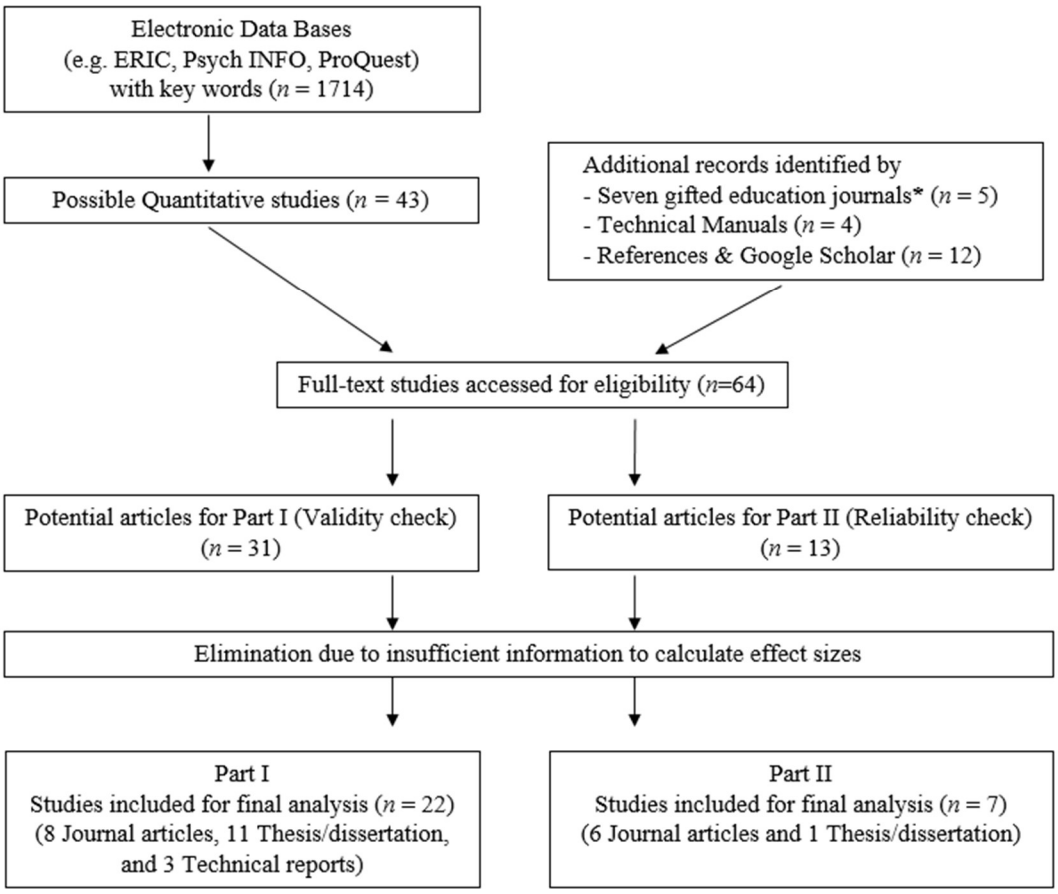
Table 5

Risk Ratio (RR) and its 95% CIs for Studies for Study Part II

		95% CI	
Citation	RR	*LL*	*UL*
Bracken & Brown (2008)	0.61	0.20	1.02
Brulles et al. (2012)	0.43	0.27	0.59
Carman & Taylor (2010)	0.52	0.28	0.76
Edmonds (2016)	0.32	0.26	0.50
Giessman et al. (2013)	0.12	-0.37	0.61
Lewis et al. (2007)	0.28	-1.21	1.77
Naglieri & Ford (2003)	0.85	0.75	0.95

Note. CI = confidence interval; *LL* = lower limit; *UL* = upper limit.

Figure 1
Literature Search Process



Note. *Original findings from the gifted education journals were six, however, we deleted one because the study has two versions – thesis/dissertation and journal article. The study was already included in the thesis/dissertation category from which we can retrieve richer information compared to the journal article.

Figure 2

Forest Plot of Study Part I

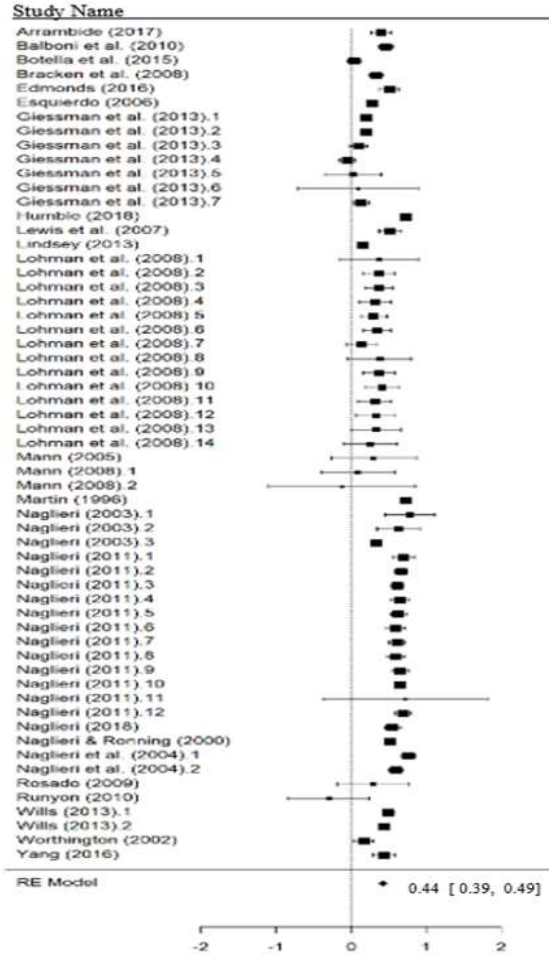


Figure 3

Funnel Plot for Publication Bias for Study Part I

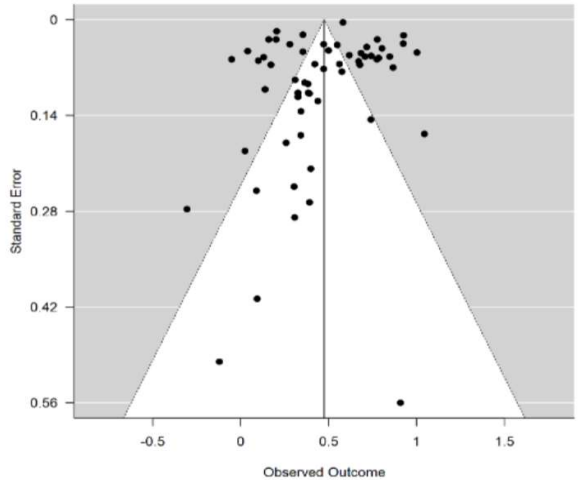


Figure 4
Forest Plot of Study Part II

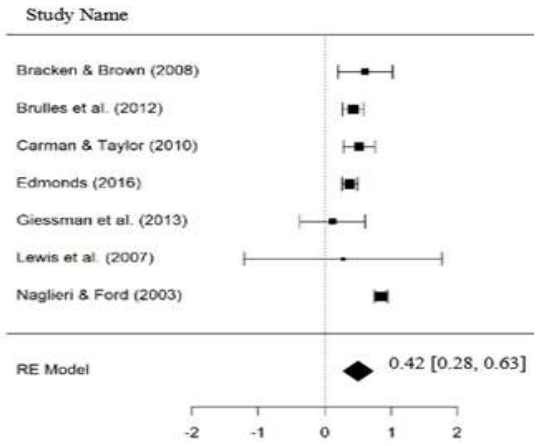


Figure 5
Funnel Plot for Publication Bias for Study Part II