Vol. 19(3), pp. 44-52, March 2024
DOI: 10.5897/ERR2023.4385
Article Number: E71A3C271977
ISSN: 1990-3839

http://www.academicjournals.org/ERR

ACADEMIC JOURNALS
expand your knowledge

Educational Research and Reviews

*Full Length Research Paper*

# Synthesizing validity and reliability evidence for the draw a scientist test

**Julia Brochey-Taylor[1]* and Joseph A. Taylor[2]**

[1]Curriculum and Instruction, Colorado Springs District 11 United States.
[2]Department of Leadership, Research, and Foundations at the University of Colorado, Colorado Springs United States.

Received 08 December, 2023; Accepted 21, March 2024

**The purpose of this synthesis study was to assess the reliability and validity of the Draw-A-Scientist Test (DAST) and its variations across multiple studies, aiming to understand limitations and propose modifications for future application within and beyond the science domain. Given the existence of multiple DAST versions, this study quantified the frequency of validity threats across various DAST variations. Literature review results indicated that despite its widespread use, the DAST and its variations consistently encounter challenges related to construct validity and external validity. Additionally, this synthesis identified literature limitations in testing concurrent validity, predictive validity, and inter-rater reliability when applicable.**

**Key words:** Scientist, assessment, stereotype.

## INTRODUCTION

Over the years, research has consistently indicated that students predominantly perceive scientists as male individuals who conform to media stereotypes (Chambers, 1983). These findings have spurred efforts among researchers and educators to dispel these stereotypes, thereby promoting greater participation of females and individuals from diverse backgrounds in the field of science. Since 1983, numerous researchers investigating scientist stereotypes have utilized the Draw-A-Scientist Test (DAST) to explore student perceptions of scientists. However, as the DAST has undergone multiple iterations over time, concerns regarding its applicability and reliability across various contexts have emerged. Consequently, the central inquiry of this literature review was: Regarding validity and reliability, what are the recurring themes in the strengths and limitations reported by authors concerning the DAST and its variants as a measure of stereotype thinking in K-12 science students?

This review aimed to synthesize evidence regarding the validity and reliability of the DAST and its different versions. Each reviewed study presented author-reported strengths and limitations. While the primary focus was not on delineating the specific findings of these studies, such as the frequency of stereotypical elements and locations in students' drawings, occasional observations were made when they provided insights into the tests' reliability and validity. Notably, gender, race, and socioeconomic variables were emphasized, given that interventions following DAST administration have predominantly targeted these aspects.

*Corresponding author. E-mail: juliabrochey@gmail.com.

Additionally, despite a previous systematic review of DAST validity evidence (Chang et al., 2020), which did not encompass commonly used DAST versions containing questionnaires and interviews, this study provided a comprehensive review of research incorporating various DAST iterations and compared reliability and validity outcomes with those identified in Chang et al.'s review.

## METHODS

Using the Boolean keyword search string "DAST" AND "methodology" AND "validity" AND "reliability," a comprehensive simultaneous search was conducted across more than 300 databases, including the Education Resources Information Center (ERIC), PsychINFO, and the Social Sciences Citation Index. This search yielded numerous studies, which were subsequently identified and screened to determine their relevance to the research question. To be considered for synthesis, a study had to have assessed the validity or reliability of the DAST or its variations. Specifically, the study could examine any of the following types of validity/reliability: concurrent validity, predictive validity, construct validity, external validity, or inter-rater reliability. Studies from all publication years, participant age groups, and settings (both formal and informal educational settings) were eligible for inclusion in the synthesis. For the purposes of this study, the following validity types were operationally defined as follows:

1. Concurrent validity: The degree to which scores on the DAST align with scores from other measures that also aims to evaluate perceptions of scientists.
2. Predictive validity: The degree to which scores on the DAST correlate with future outcomes, such as interest in pursuing careers in STEM fields.
3. Inter-rater reliability: The degree to which ratings from multiple assessors of the same drawing are consistent with one another.
4. Construct validity: The degree to which the DAST accurately measures what it claims to measure, namely conceptions of scientists.
5. External validity: The extent to which the findings of the DAST can be generalized across different samples and contexts.

The results of these studies were quantified and compared to those of a similar systematic review conducted by Chang et al. (2020), which focused solely on the validity and reliability of the DAST and MDAST. Notably, Chang et al. (2020) did not address predictive validity or external validity, whereas this study expanded upon their review by examining evidence in these validity categories.

## RESULTS

The results of this review are provided in 2 sections. In the first section, the study synopsis provides summaries of the 16 studies, focusing on describing DAST variations, study samples, and primarily author-reported reliability and validity threats. Subsequently, a section called Aggregated Results, which seeks to summarize the reliability and validity threats across the corpus of studies reviewed was provided.

### Study synopses

Chambers (1983), in implementing the DAST, sought to find out how early in childhood the stereotypical image of a scientist may appear and at what age these images may be embedded. To implement this test, teachers asked students to *draw a picture of a scientist*. As a control, 18.9% of the sample was also asked to draw a person. Over the span of 11 years, 4807 kindergarten through fifth graders from Canada, parts of the United States, and Australia, participated in the test. Results revealed that by second grade, the stereotype of the scientist had begun to develop and by fifth grade most of the stereotypical elements appeared in student drawings, suggesting an embedded stereotype and an overall increase in stereotypical associations as grade levels progressed. Chambers noted that strength of the DAST, because it is nonverbal, is that it could be given to students who were not yet capable of providing clearly written responses. Additionally, Chambers suggested that correlations could exist between the scientist characteristics found in children's drawings and various other psychological and social parameters. Chambers stopped short of testing such correlations in his study. As such, the study missed an opportunity to test the concurrent and predictive validity of DAST. Furthermore, according to Chambers, another limitation of the DAST, while easier to administer than many other measures, may be the challenges it raised for confident interpretation; this is viewed as a construct validity concern. This construct validity concern prompted Chambers to assert that the test may be better suited as a tool for exploratory hypothesis generation than for confirmatory hypothesis testing.

In their study, the Draw a Scientist Test: Interpreting the Data, Symington and Spurling (1990) reported a potential issue with the methodological structure of Chambers' aforementioned DAST. According to their report, a flaw was apparent in the wording of the prompt given to participants. In Chambers' 1983 test, respondents were asked to *draw a picture of a scientist.* This prompt, according to the report, did not suggest a purpose and therefore, perhaps, students drew what they deemed to be expected of them - a stereotypical image of a scientist. When wording was later changed, children's depictions of scientists also changed. The report included two figures, each with a side-by-side comparison of drawings. Children were asked to A. *Draw a picture of a scientist* and B. *Do a drawing which tells me what you know about scientists and their work*. In both figures, responses to prompt B included more detail. According to Symington and Spurling, this may suggest children's depictions are not of their own opinions but of their interpretations of the given prompts. This clearly illustrates a construct validity threat.

In the report, Some Methodological Issues with Draw a Scientist Tests among Young Children, Losh et al. (2008) also raised questions regarding the validity of the DAST. The authors suggested that more comparisons be given so that students have an opportunity to draw individuals from various occupations, such as teachers and veterinarians. The study enabled researchers to examine 616 drawings by 206 elementary students. To study

differences in drawings across students, researchers grouped the drawings by student characteristics such as gender, ethnicity, and grade. From each drawing, the following scientist characteristics were extracted: Gender of figure; clarity of figure gender; color of figure; whether the figure was human; the number of drawing details; whether the figure smiled; and figure attractiveness. In making comparisons, researchers controlled for developmental factors such as grade as it was noted that girls develop fine motor skills earlier. Finally, they adjusted the order in which prompts were given to control for 'order' effects that might influence drawings. In reviewing the results, the researchers found many limitations and warned that caution should be taken in interpreting the drawings; for example, student maturation and projection were found to influence drawings as was gender of adults in respondents' field of view. Further, the study was limited to a medium-sized school in the South, a threat to its external validity. At the same time, researchers suggested that the use of drawings may still be valid, particularly for students who are too young to write clearly or write at all.

Farland-Smith (2012) sought to field test and discussed the development and reliability of a modified DAST along with its rubric, with the latter designed to refine the test. More specifically, Farland-Smith wondered if a rubric designed for the DAST could be used to test reliability among teacher-raters. Previous changes had been made to the initial DAST, which in 1995 included a checklist; this became known as the DAST-C. Farland-Smith further modified the test, creating the Modified Draw a Scientist Test (M-DAST), by changing the prompt to include categories of location, appearance, and activity of scientist. The test was given to students in four sixth grade classes in a public school system in the Midwestern United States and was administered by two teachers, which limits the generalizability of the findings. Teachers taught the same grade and had been teachers for 8-10 years. They were trained to score student pictures using a rubric, with score categories of: Sensationalized, traditional, and broader than traditional. The report reflected an inter-rater reliability of 89% for the appearance category, 94% for location category, and 88% for activity category. Though inter-rater reliability was high, it was reported that interpretation issues could still arise via the rubric, thus construct validity was called into question. Nevertheless, it was deemed a valuable tool, as was the M-DAST itself, as a valid and reliable way to study student perception of scientists.

Reinisch et al. (2017) reported findings from their study that examined the validity of the DAST and similar instruments. The authors purported that in recent years, many DAST studies were administered to pre-service teachers rather than to school children; this is partly due to the influence teachers may have on students. Considering the previously discussed M-DAST and its rubric, the authors established the following research questions: To what extent is it possible to objectively categorize pre-service science teachers? What methodical challenges could be encountered while assessing and rating drawings of scientists? The study included two cohorts of pre-service science teachers (n = 79). The drawings were analyzed across three main categories of appearance, location, and activity. Written explanations of the drawings were analyzed. With respect to providing drawings and written explanations, respondents indicated after the study that they had challenges in taking the test; for example, science is too great a field to cover in one drawing and drawing ability was limited. This is, of course, worth considering when examining the content and construct validity of the test. Besides the issues mentioned above, researchers noted a potential problem with having participants draw only one scientist; if participants are asked to *draw a scientist*, of course he or she would appear to be solitary. It was suggested, therefore, that prompts be examined extensively. Though strength of this test continues to be its nonverbal application, researchers did not deem this to be necessary for populations that are able to write/discuss their drawings. Focus groups that utilize discussions, rather than drawings, were suggested as an alternative to the test. The authors did, however, report strong inter-rater reliability.

In consideration of previous data pertaining to the DAST, Walls (2022) took a deeper look into claims that children have drawn primarily white male figures; Walls did so through a lens of Critical Race Theory (CRT). The paper, A CRT Analysis of the Draw-a-Scientist Test: Are they really that white, sought to expose the apparent exclusion of children of color and to answer: To what extent is colorblind ideology present in peer reviewed DAST research? How does CRT reveal other forms of inequity within DAST?

Utilizing 28 DAST studies, Walls gathered the following information: Publishing journal; number of participants per study; grade level; racial identification; gender; instruments used to collect data. Combining this with a focus on participants' drawn images of scientists, location of schools/classrooms, teacher prep programs/university, and the most frequently cited studies, enabled Walls to conduct a content analysis. Because Walls found there to be a lack of attention to race within the text of the studies, he reported there to be colorblind racism in research practices that employ the DAST. Wells found that 57% of the 28 studies did not report race. Walls was further critical of prior research, noting the confounding of race and ethnicity, which was used interchangeably throughout prior research. Regarding the second research question, Walls found former studies to be fraught with unsubstantiated claims, such as reports that most figures were white, though many researchers admittedly could not identify race in student drawings. Ultimately, Walls added his findings to the methodological issues that DAST researchers have previously reported.

Meyer et al. (2019) added to the current body of research by examining more closely career choices and their relationship to stereotypes. In this study, the test was given to 445 first-year university students. Considering previous claims that the perception of scientists impacts choice of career, researchers purported an expected difference in DAST examples created by students in majors outside of science. Results were consistent with findings in earlier research and, as was hypothesized, students from the arts and social sciences drew stereotypes more frequently. There were multiple limitations within the study. First, students were able to draw only one representation of a scientist. Additionally, only drawings were assessed and there were no verbal or written elaborations given. Further, gender could not always be assessed by those scoring the tests. Moreover, drawing skills varied among participants. Yet another threat was that the study only included participants from one university and only two thirds of participants were able to complete drawings.

Toma et al. (2022) sought to empirically test assertions within the DAST and DAST-C body of research. The study focused on the following key questions: How deeply DAST and DAST-C protocols predict interest in the field of science; how likely it is that those drawing stereotypical images display less interest in science as a career; to what extent does drawing male scientists predict/impact interest in a science career? To study this, Bogdan analyzed responses from 1799 students from seven schools. Bogdan used results from both the DAST and the DAST-C in the study. Analyses to test predictions and relationships utilized multiple regression Results failed to show that DAST and Dast-C scoring protocols could predict significant variance in interest in science as a career. Similarly, findings did not support the notion that females who drew male scientists had less interest in science as a career. Given the results, researchers questioned the validity of the DAST and DAST-C protocols. Still, a sample limitation of this study was that participants were all living in Columbia and, researchers noted, individuals in Columbia tend to place greater value on careers in science.

Finson (2003) reported results from a study that questioned the applicability of the DAST-C to different racial groups with the hypothesis that the difference in student perceptions would not be significant and that the test, therefore, could be administered to all racial groups to yield valid results. Recall that the DAST-C is an expanded version of the DAST. To test this, Finson studied 191 eighth graders in the Midwest. Scores were compared using ANOVA procedures. Results indicated no significant difference between the groups of participants ($F = 0.22$; $p = 0.80$). A potential threat was that the results were not necessarily generalizable among populations. Further, as mentioned in previous studies, the drawings relied on a certain amount of interpretation, rendering results potentially invalid. Threats could also lie in the drawings; Finson noted that some students may not have drawn the image that actually existed in their minds. Lastly, the study could not control for exposure to media stereotypes. It is our view that the author overlooked the inclusion of several other racial groups and that the sample size could have been increased.

Hillman et al. (2014) examined whether the presence of STEM Fellows made a difference in stereotypical drawings of students in grades K-12 and to determine validity of the DAST in comparison with a newer survey, containing six questions. The combination of DAST with a stereotypes survey was used in order to ascertain what stereotypes were still held by students. To do this, researchers gave 485 students pre and post surveys (62 elementary students, 208 middle-school students, and 215 high school students). This 2-year study took place in rural and suburban areas of Maine. Nine STEM Fellows served one classroom for the entire academic year. The fellow at the elementary school level rotated schools during science instruction. Fellows were in classrooms for 10 h per week with a purpose of infusing graduate research into curriculum. Results revealed there to be a significant, though low, positive correlation between the surveys' results indicating stereotypes and DAST results indicating the same. Due to this, researchers noted that DAST should be given in conjunction with interviews and or survey questions.

Bozzato et al. (2020) analyzed drawings of a convenience sample of 686 elementary school children. The participants were enrolled in urban, public schools in northern Italy and were of mid-socioeconomic status. Using the DAST-C and the DAST-Rubric, which are multidimensional, students were asked to draw scientists. Tests were administered during the school day with the help of the classroom teacher and research team members. These research team members were trained to code the tests; using Cohen's index, the inter-rater reliability was reported to be 0.87. Overall results were consistent with previous studies throughout the US and elsewhere, that is 61% of the children drew male scientists. Female scientists were included in the remaining 31.6%, drawn mostly by females. Of the drawings, 83.5% included symbols such as test tubes. Although the DAST-C and the rubric were employed to enhance understanding of student perceptions, authors determined that these instruments were limited to their detection of perception only, whereas tests that use interviews in conjunction with tests can give a broader, deeper picture. Another limitation was that the data were analyzed via a non-randomized sample and were therefore not representative of a larger population of Italian children.

Laubach et al. (2012) implemented the DAST to explore differences in perceptions between gender, grade, and tradition within Native American culture. Participants included high school students who attended boarding schools outside of the reservation (n=133). Students who

spoke their native language in the home were categorized as practicing cultural traditions. Researchers conducted a content analysis of the submissions. DAST-C protocols were followed by lead researchers who scored student drawings. Classroom teachers and science faculty additionally and individually scored a subset of the drawings. An inter-rater reliability of 0.92–0.94 was determined for the subset. A statistically significant finding was the difference between groups who practiced native tradition and those who did not ($p < 0.05$). That is, Native American students who practiced native traditions, when compared to those who did not, drew fewer stereotypical elements as instruments in their drawings, such as lab coats and test tubes. Researchers purported that the results could suggest that while those students who practiced traditions were more able to flex between traditional knowledge and Western science, most students did not appear to view themselves as being scientists. As with other studies, it appeared students drew primarily white figures in their drawings. There were a few limitations within the study. First, students were given limited materials, including pencils and pens; native language was used as an imperfect indicator of student connection to cultural practice; participants were only a small representation of the many Native American cultures that exist; and students drew only one scientist.

Quilez-Cervero et al. (2021) examined elementary students' perceptions of scientists in light of COVID-19. Participants were 128 early primary grade children, 58 girls and 70 boys, ranging in ages 6 to 8 years. Students attended a school in a small town in Spain. Instruments used in the study included an illustration drawn by students, the workplace of the scientist depicted in the study, the workplace activity, student interviews to obtain further information about the drawings, an analysis rubric per the M-DAST, and a table for classification of drawings based on the four categories of gender, age, representation of clinical researchers, and representation of COVID-19 in the drawing. Ultimately, the study revealed that most drawings depicted a broader view of traditional images. Still, there was a significant percentage of boys whose drawings depicted stereotypical images of the scientist. As in previous studies, when females were included in drawings, they were primarily drawn by females. Females also drew males and females working side-by-side. It was noted in the study that students drew primarily young scientists. More females included representations of COVID-19. Limitations included sample size, location of the facility within a small town in one region of Spain, and the period of time in which the study was conducted which was post-pandemic, post confinement. Further, the prompt may have not been clear to some students. Indeed, the prompt appeared somewhat complex for students below grade three. Another concern arose from researchers' observations that drawings depicted younger individuals more frequently than older ones. However, drawing an older person may necessitate more advanced fine motor skills; moreover, perceptions of youth versus age are subjective. Lastly, there was no reference to inter-rater reliability.

Lamminpää and Vesterinen (2020) used alternative prompts in order to measure student perception of the DASC (The Draw-A-Science-Comic test). The DASC is a study that invites participants to draw a sequential comic depicting science and scientists. Authors believed that the formerly used prompts, which included the word *comic*, may have influenced student drawings, perhaps misrepresenting their images of a scientist. Students were asked to respond to the following prompts: *Draw a comic about how you think science is made* (n=73), *draw a story about how you think science is made* (n=68), *draw a set of pictures about how you think science is made* (n=39). Lamminpää and Vesterinen were interested in answering the following: In what ways do the alternative prompts affect storytelling and appearance of scientist/activity/ attitude within depictions? How does the age of students affect storytelling, the age, and overall representation of scientists? In what way are danger elements depicted, particularly sequentially? How frequently do the danger elements occur in each science field? Results revealed that the DASC does not offer an advantage to interpreting student drawings compared to the traditional DAST or M-DAST when considering scientist appearance and setting. However, researchers found that the use of the comic prompts and picture prompts enabled students to provide more information than the story prompts or the traditional DAST. One limitation was the lack of focus on different branches of science; researchers noted that including different branches in prompts could help with assessing student stereotypes of each. Another limitation was the lack of interviews to accompany each drawing so that some interpretation was necessary and therefore debatable. Finally, students were given tests prior to attending camps for which they had registered; this anticipation could have influenced student drawings.

Fung (2002) reported results from administration of the DAST. The DAST was administered to Chinese students in Hong Kong -675 elementary and secondary students - in an effort to examine students' perceptions of scientists. It specifically made comparisons across grade and gender. The study was additionally used as a means by which to compare perceptions of scientists in Hong Kong with those perceptions held elsewhere. The prompts in this study were different from Chambers' DAST. Students were directed to *draw scientists as they see them* rather than to, simply, *draw a scientist*. Another difference was that students were invited to draw two scientists if desired. Coding was done by a research assistant and the author. Coding discrepancies occurred in less than 10% of the drawings and were corrected. Results were consistent with those found in Taiwan and the West. As is shown in many studies of its kind, the scientists were drawn as being

primarily male with an increase in stereotype as grades progressed. While some of the drawings did include subtext to further represent images, Fung recommended that studies in the future include interviews so that researchers can develop a better understanding of student perceptions. This lack of depth was a limitation of the test, though due to its ease of use, it was described as being useful.

In Bernard and Dudek (2017) study, the prompt "What is secondary-school students' image of people conducting scientific research?" was utilized to investigate participants' perceptions of science and scientists. This phrasing was chosen in light of limitations observed in other iterations of the DAST. The researchers deliberately refrained from using the term "scientist," opting instead for a broader terminology. Additionally, the study incorporated a questionnaire and allocated space for descriptions within the test format. The results of this indirect DAST (InDAST) were compared to the DAST. Questionnaires were examined both qualitatively and quantitatively. Data was organized and coded by one trained coder. Additionally, 40 randomly selected questionnaires were coded again by the primary coding person and then again by an independent researcher. With regard to inter-rater consistency, correlation coefficients between the primary coder and independent coder were 0.85($p<.001$) for DAST and 0.80 ($p<.001$) for InDAST. Overall, drawings echoed results found in previous tests; however, the InDAST depictions featured fewer older people and fewer wild hairstyles. It additionally featured more people working in groups. Further results indicated that the majority of drawings featured mainly men, followed by gender-neutral images, and then men who were alone. The least common depictions were of women. An important limitation was one that comes up frequently in DAST research, that is that drawings are still in need of interpretation and do not give insight as to the reasons why drawings appear as they do, for example. It was noted again in this research that interviews could be an important addition to the test.

Dickson and McMinn (2022) reported the results of a DAST that was given to students in grades 3 and 7 who attended government or private international schools in Abu Dhabi. Nearly 100% of government school students were citizens of the UAE while 50 to 80% of the private international schools were national citizens. The majority of the 234 participants were UAE citizens. Researchers could not determine a clear number pertaining to student nationality, however. Core content classroom teachers were given instructions regarding test administration. Students were given paper and colored pencils within science classrooms where they were prompted to *draw a picture of the scientist doing science.* Students were additionally prompted to *draw a person*, which was used as a control method. Coding was added to drawing paper according to school, grade, and gender. Researchers were interested in following the original DAST protocols (Chambers, 1983). Drawings were rated according to specific stereotypical images that were included in drawings (a yes/no or 1/10 scale). Additionally, a blind sample of 20 drawings were coded and compared by researchers. Interpretive differences were discussed until differences reached near 0. A descriptive analysis was used regarding characteristics that were included in the drawings. Researchers quantified additional characteristics as they arose such as representations of danger and mythical creatures. Researchers were primarily interested in analyzing drawings for gender and representations of national clothing. Findings were again indicative of stereotypical thinking as more males were drawn than females. Females tended to draw female scientists more frequently than did males. Researchers noted progress in closing the gender gap since Chambers' 1983 DAST. Regarding national clothing, students did not include this feature. Like other DAST research, the strengths of this DAST included ease of testing and low cost. Limitations included an unclear percentage of student nationality and lack of depth in information derived from drawings. Researchers suggested triangulating findings in the future such as including interviews and focus groups.

Chang et al. (2020) conducted a systematic review that evaluated 76 studies. Via this research, authors discovered four current justifications for the use of the DAST that utilize participants' drawings: Drawing tests can be used as an alternative method for students lacking written and or language ability and who may have socioeconomic and affective reasons; drawings can expose aspects that are not readily measured in other tests; tests reveal multiple characteristics; tests are formative assessments that can investigate students' ideas, thereby reforming education accordingly. Within the study, researchers also examined trends in drawings. Author research questions included: What are the main characteristics of the research studies? What evidence shows justification for the use of drawing as an assessment method? What are the primary findings from the studies? How did the studies test validity and reliability? To gather data, researchers used two databases, which were chosen due to the number of high-quality papers within. Only drawings that included coding of results were used; those that used supplemental interviews and the like were not used. There were 46 empirical studies analyzed and 30 additional studies were included, using a snowball sampling method. Three decades of research, overall, were analyzed. Pertaining to the analysis of validity, merely seven papers (9%) directly included information regarding validity evidence while 61.3% of the studies claimed the instruments were sufficiently reliable. Thirty-three studies reported inter-rater reliability as a means by which to measure reliability. Twelve papers used Cohen's Kappa coefficients to confirm the reliability of the drawing assessments.

**Table 1**. Validity issues discussed in reviewed articles.

| Study | Concurrent validity tested | Predictive validity tested | Construct validity threat | External validity threat | Inter-rater reliability tested |
|---|---|---|---|---|---|
| Chambers (1983) | | | X | | |
| Symington and Spurling (1990) | | | X | | |
| Losh et al. (2008) | | | X | X | |
| Farland-Smith (2012) | | | X | X | X |
| Reinisch et al. (2017) | | | X | | X |
| Meyer et al. (2019) | | X | X | X | X |
| Toma et al. (2022) | X | X | | X | X |
| Finson (2003) | | | X | X | X |
| Hillman et al. (2014) | X | | | X | |
| Bozzato et al. (2021) | | | X | X | X |
| Laubach et al. (2012) | | | X | X | X |
| Quilez-Cervero C et al. (2021) | | | X | X | |
| Lamminpää and Vesterinen (2020) | | | X | | X |
| Fung (2002) | | | X | | X |
| Bernard and Dudek (2017) | | | X | X | X |
| Dickson and McMinn (2022) | | | X | | X |

A challenge to this study was that reviews were made only of those papers published in journals that ensured quality and consistency. Researchers noted, however, that publication bias was not a concern.

## Aggregated results

The review's primary research question necessitated a synthesis of the validity evidence discussed in relevant studies of the DAST. To achieve this, Table 1 presents tallies of the validity types tested or identified as problematic in each reviewed study. Furthermore, Figure 1 illustrates these tallies as percentages of studies referencing each type of validity test or issue.

As depicted in Figure 1, less than 20% of the studies assessed concurrent validity or predictive validity. However, inter-rater reliability was examined in nearly 70% of the studies. Regarding author-reported threats to validity, 88% of the studies identified a construct validity concern, while 63% cited an external validity issue. It is evident that these studies did not consistently evaluate all relevant types of validity, and when validity concerns were raised, they predominantly fell within the construct and external validity categories.

## DISCUSSION

The DAST has been trusted for some 50 years to analyze student perception of scientists and their work. Despite much iteration, however, DAST studies continue to report similar limitations, particularly in the areas of external validity and construct validity. Although these limitations exist, studies also report positive changes in perspective over time, perhaps through the numerous interventions implemented throughout time. Researchers additionally report the ease of administration of various versions of the DAST, the affordability of the test, and its ability to study perceptions of students who are not yet verbally able to communicate such perspectives. Simultaneously, researchers have implemented verbal methods to gain a deeper understanding of student perception.

Though this study sought to synthesize validity and reliability evidence for the DAST, it was limited by the availability of only 16 eligible studies. In contrast, Chang et al. (2020) examined 76 studies that more generally examined using student drawings as assessments. Nine percent of their studies discussed how validity was tested, 2 of which included expert validity; two others included content validity (that is, the extent to which the measure assesses the entire domain of interest). Triangulation was used in one paper while two tested concurrent validity. Sixty-one percent of the studies in the Chang et al. (2020) review reported adequate reliability. Thirty-three of these studies specifically tested inter-rater reliability.

## Conclusion

Based on the results of this study as well as the broader study (Chang et al., 2020), it cannot be confidently stated whether the DAST always functions as a valid and/or reliable measure of student perception of scientists. Therefore, further research is needed and should focus on studying all variations of the DAST. Additionally, it should

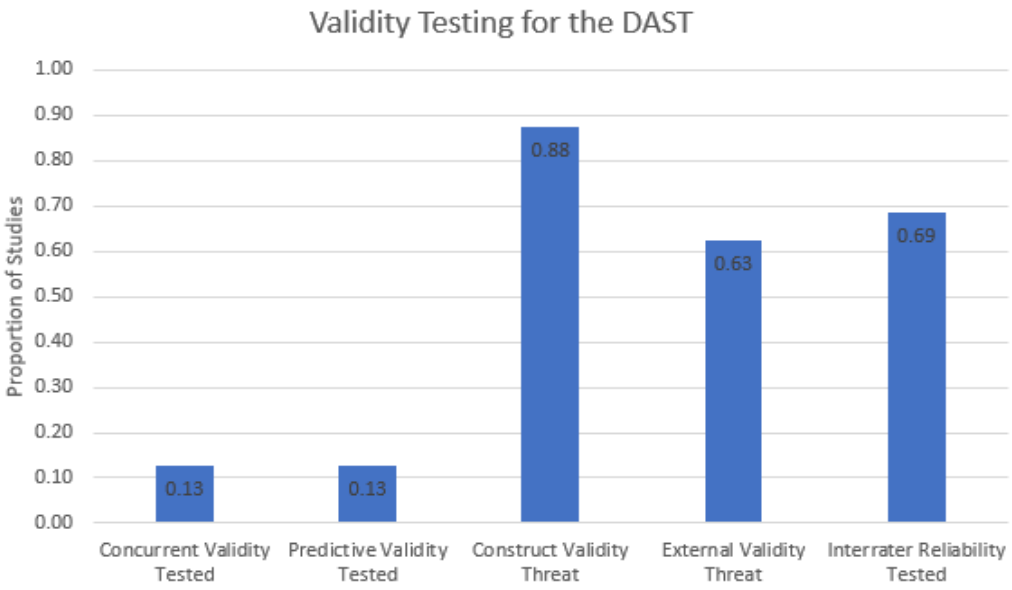


**Figure 1.** Percentage of studies testing or citing validity issues.

aim to document other types of validity regarding the DAST, most urgently predictive validity and concurrent validity. A notable contribution to consider for further research is the work of Walls (2022), which reported the lack of attention to race within DAST studies. A strength of this body of literature lies in documenting the utility of DAST for gathering student perceptions of scientists. Its weakness, as it relates to this review, included a lack of reporting pertaining to validity and reliability. These conclusions are echoed by those of Chang et al. (2020). As mentioned previously, however, due to its ease of use and other favorable attributes, the DAST could be worthy of further research and implementation within and outside of the science classroom.

## CONFLICT OF INTERESTS

The author has not declared any conflict of interests.

### REFERENCES

Bernard P, Dudek K (2017). Revisiting students' perceptions of research scientists –outcomes of an indirect draw-a-scientist test (InDAST). Journal of Baltic Science Education 16(4):562-575.

Bozzato P, Angelo FM, Longobardi C (2021). Gender stereotypes and grade level in the draw-a-scientist test in Italian school children. International Journal of Science Education 43(16):2640-2662.

Chambers D (1983). Stereotypic images of the scientist: The draw-a-scientist test. Science Education 67(2):255-265.

Chang H, Lin TJ, Lee MH, Wen-Yu LS, Lin TC, Tan AL, Tsai CC (2020). A systematic review of trends and findings in research employing drawing assessment in science education. Studies in Science Education 56(1):77-110.

Dickson M, McMinn M (2022). Children's perceptions of scientists and their work: The 'draw a scientist' test in the United Arab Emirates. Public Understanding of Science 31(8):1079-1094.

Farland-Smith D (2012). Development and field test of the modified draw-a-scientist test and the draw-a-scientist rubric. School Science and Mathematics 112(2):109-116.

Finson KD (2003). Applicability of the DAST-C to the images of scientists drawn by students of different racial groups. Journal of Elementary Science Education 15(1):15-26.

Fung YYH (2002). A comparative study of primary and secondary school students' images of scientists. Research in Science and Technological Education 20(2):199-213.

Hillman SJ, Bloodsworth KH, Tilburg CE, Zeeman SI, List HE (2014). K-12 students' perceptions of scientists: Finding a valid measurement and exploring whether exposure to scientists makes an impact. International Journal of Science Education 36(15):2580-2595.

Lamminpää J, Vesterinen VM (2020). Draw-a-science-comic: Alternative prompts and the presence of danger. LUMAT: International Journal on Math, Science and Technology Education 8(1):319-339.

Laubach TA, Crofford GD, Marek EA (2012). Exploring Native American students' perceptions of scientists. International Journal of Science Education 34(11):1769-1794.

Losh S, Wilke R, Pop M (2008). Some methodological issues with 'draw a scientist tests' among young children. International Journal of Science Education 30(6):773-792.

Meyer C, Guenther L, Joubert M (2019). The draw-a-scientist test in an African context: comparing students' (stereotypical) images of scientists across university faculty. Research in Science and Technological Education 37(1):1-14.

Quilez-Cervero C, Diez-Ojeda M, López Gallego AA, Queiruga-Dios MA (2021). Has the stereotype of the scientist changed in early primary school–aged students due to COVID-19? Education Sciences 11(7):365.

Reinisch B, Krell M, Hergert S, Gogolin S, Krüger D (2017). Methodical challenges concerning the draw-a-scientist test: a critical view about the assessment and evaluation of learners' conceptions of scientists. International Journal of Science Education 39(14):1952-1975.

Symington D, Spurling H (1990). The 'draw a scientist test': Interpreting the data. Research in Science and Technological Education 8(1):75-77.

Toma R, Lucía Orozco-Gómez M, Carolina Molano NA, Lucía Obando-Correal N, Rocío Stella Suárez Román (2022). Testing assumptions of the draw-a-scientist-test (DAST): Do stereotyped views affect career aspirations? International Journal of Science Education 44(16):2423-

2441.
Walls L (2022). A critical race theory analysis of the draw-a-scientist test: are they really that white? Cultural Studies of Science Education 17:141-168.