International Journal of Science Education

ISSN: 0950-0693 (Print) 1464-5289 (Online) Journal homepage: https://www.tandfonline.com/loi/tsed20

Gender-fair assessment of young gifted students' scientific thinking skills

Y. J. Dori, A. Zohar, D. Fischer-Shachor, J. Kohan-Mass & M. Carmi

To cite this article: Y. J. Dori, A. Zohar, D. Fischer-Shachor, J. Kohan-Mass & M. Carmi (2018) Gender-fair assessment of young gifted students' scientific thinking skills, International Journal of Science Education, 40:6, 595-620, DOI: 10.1080/09500693.2018.1431419

To link to this article: https://doi.org/10.1080/09500693.2018.1431419

Published online: 28 Feb 2018.

INTERNATIONAL JOURNAL OF SCIENCE EDUCATION, 2018
VOL. 40, NO. 6, 595–620
https://doi.org/10.1080/09500693.2018.1431419

Gender-fair assessment of young gifted students' scientific thinking skills

Y. J. Dori[a,b], A. Zohar[c], D. Fischer-Shachor[a*], J. Kohan-Mass[c] and M. Carmi[a**]

[a]Faculty of Education in Science and Technology, Technion, Israel Institute of Technology, Haifa, Israel; [b]The Samuel Neaman Institute for National Policy Research, Haifa, Israel; [c]The Seymour Fox School of Education, The Hebrew University of Jerusalem, Jerusalem, Israel

ABSTRACT

This paper describes an Israeli national-level research examining the extent to which admissions of elementary school students to the gifted programmes based on standardised tests are gender-fair. In the research, the gifted students consisted of 275 boys, 128 girls, and additional 80 girls who were admitted to the gifted programme through affirmative action (AA). To assess these young students' scientific thinking skills, also referred to as science practices, open-ended questions of case-based questionnaires were developed. The investigated scientific thinking skills were question posing, explanation, graphing, inquiry, and metacognition. Analysis of the students' responses revealed that gifted girls who entered the programmes through AA performed at the same level as the other gifted students. We found significant differences between the three research groups in question posing and graphing skills. We suggest increasing gender-fairness by revising the standard national testing system to include case-based narratives followed by open-ended questions that assess gifted students' scientific thinking skills. This may diminish the gender inequity expressed by the different number of girls and boys accepted to the gifted programmes. We show that open-ended tools for analysing students' scientific thinking might better serve both research and practice by identifying gifted girls and boys equally well.

ARTICLE HISTORY
Received 3 June 2017
Accepted 18 January 2018

KEYWORDS
Gifted students; assessment; gender-fair; thinking skills; question posing; graphing skill; inquiry

Examination of women's status in the world indicates that in many domains, including academia and industry, women in the third Millennium are still under-represented in high positions (AIP, 2013). In the USA, women enrolment numbers in science, technology, engineering, and mathematics (STEM) undergraduate studies between 2000 and 2011 are lower than those of men (National Science Board, 2012), and the percentage of women who received a bachelor degree in computer sciences, mathematics, physics, and engineering declined (National Science Board, 2014). Gender differences vary from one country to

CONTACT Y. J. Dori yjdori@technion.ac.il
*Present address: The Unit of Teaching and General Studies ORT Braude, College of Engineering, Karmiel, Israel
**Present address: Department of Science Teaching, Weizmann Institute of Science, Rehovot, Israel

another and from one scientific domain to another, but in most countries, men outnumber women in physics and engineering studies, while women tend to choose biology, medicine, and environmental studies. Sadler, Gerhard, Hazari, and Tai (2012) showed that in the United States, among 6000 students, men were more interested in engineering, while women were more attracted to health and medicine during their high-school years. In Israel, the trend is similar: men outnumber women in science and engineering fields, such as physics, engineering, mathematics, statistics, and computer science, while in the paramedical fields, the gender balance is shifted towards women (Central Bureau of Statistics in Israel, 2011, 2013). Over the past decade, Israeli women have increased their participation in the workforce, making up a large percentage of employees (Eglash, 2010). Yet, despite the fact that a relatively large number of Israeli women hold graduate degrees, they still earn less, and are far less influential in the government, academia, and industry than their male counterparts are. Since women constitute half of the population, their representation in any science-, engineering-, and mathematics-related positions should get closer to this fraction.

Sadler et al. (2012) identified that the key factor that predicts interest in students' STEM career at the end of high school is their interest in learning STEM at the start of high school. Therefore, in order to foster young female and male students to choose STEM studies and eventually a STEM career, educators and science educators must make efforts to encourage elementary school-level students to develop interest and scientific thinking skills. This is especially true for young high achievers and even more so for gifted children. The effort to foster gender-fairness in education and workplace should begin in the early school grades. This is even more critical when examining and selecting young gifted students, since gifted girls are under-represented in many educational programmes for gifted children (Zorman & David, 2000).

Early learning experiences influence students' performance in STEM (Huang & Du, 2002). At the end of elementary school, children of both genders associate science with males, and this has implications for their future education and career choices as young adults (Hughes, 2002). In schools where there is a higher number of girls or single-gender classes, female students are less likely to choose female-dominated subjects and more likely to choose technical or scientific subjects (Lavy & Schlosser, 2011; Schneeweis and Zweimüller, 2012). In particular, studies on girls and boys in physics classes showed that having more or less girls in the classroom is not effective if it is not supported by a girl-friendly curriculum, a supportive teacher, and a welcoming learning environment (Haussler & Hoffmann, 2002; Hazari, Sonnert, Sadler, & Shanahan, 2010). According to Kahle (2004), the gender gap in science varies by discipline and is more apparent in physics than in biology. Moreover, she claimed that educators have not addressed gender issues properly, and a lack of gender-fair teaching approaches might leave girls behind in terms of their scientific literacy. Other researchers, who investigated gender differences in science performance with dependent variables such as hands-on activity and discipline, showed that fifth grade girls performed better than or the same as boys. This was the case not only in biology hands-on assignment, but also in physical sciences (Pine et al., 2006).

Baker and Leary (1995, also in 2003), who interviewed elementary and high-school girls, found that they liked both physical and biological topics, but were more interested in science topics that were relevant to their lives. They reported that these girls' strongest commitment to a scientific career related to the enthusiasm of a parent or a grandparent who had been engaged in science as a profession or through mutual scientific experiences with these family members.

Culture in the context of science education revolves around issues of equity for students from low-income, racial, and ethnic minority communities. Seiler (2013) commented on this view of culture as masking the role of systemic inequity in certain communities. Banks and Banks (1995) noted that it is important to create equal educational opportunities for students from diverse cultural groups. Our study attempts to narrow the currently existing gap in the success rate of girls and boys in admission to gifted programmes in Israel.

Committed to equity-based education and striving to improve this situation, the Israeli Ministry of Education (IMoE) has carried out a policy of affirmative action (AA) in the acceptance process for gifted girls in some of the dedicated classes and pullout programmes (Dori, Zohar, Fischer-Shachor, Kohan-Mass, & Carmi, 2009; Fischer-Shachor, Carmi, & Dori, 2010; Rosemarin, 2001; Tal & Medijensky, 2005). This paper describes a national-level research aimed at examining the extent to which admissions of elementary school students to the gifted programmes are gender-fair from two perspectives, namely (a) standardised tests and (b) overall scientific thinking skills and specific scientific thinking skills. Recently, The Framework for K-12 Science Education (National Research Council [NRC], 2012) has coined the term 'scientific practices', referring to scientific thinking skills as it emphasises the integration of these skills with scientific knowledge. These practices include asking questions, planning investigations, analysing and interpreting data, and constructing explanations. Throughout this paper, we refer to scientific practices as scientific thinking skills. Specifically, we based our gender-fair assessment on the following five scientific thinking skills: question posing, explanation, graphing, inquiry, and metacognition.

This research is pertinent to the domain of knowledge on gifted students, gender equity, and gender-fair assessment of students' scientific thinking skills, with the goal of making AA unnecessary. In the domain of assessment, current national standardised tests quite often suffer from lack of gender-fairness. Situated in the context of social and cultural inclusion, this problem is especially consequential when used for identifying elementary school students' eligibility for participating in gifted programmes. Our study aims to develop a new approach for assessing students' scientific thinking skills, supported by an adequate instrument, to improve gender-fairness in the process of identifying gifted students. Since gifted girls as well as boys are the potential next-generation leaders, especially in science and engineering domains, it is highly important to see to it that girls and boys are represented more or less equally in gifted programmes.

Theoretical background

Educators, researchers, and philosophers have discussed the question of nature versus nurture for many years. This debate has had a widespread influence on the framework of research in child development. Throughout history, researchers have been trying to shift the debate between the theories 'Developmental Systems' (nurture) and 'Golden Chromosome' (nature) (Renzulli, 2005).

Giftedness

According to the traditional definition, gifted children are those who have high intelligence, as measured by intelligence quotient (IQ) tests (Terman & Oden, 1959), and have a high degree of academic and professional success. Subsequent reviews and reports confirmed this traditional definition (Gallagher, 2004; US Office of Education, Marland, 1972). Even today, in many states in the USA, as well as in other countries, such as the UK, gifted students have been identified almost solely by their 90th or 95th percentile in IQ test scores (Kornilov, Tan, Elliott, Sternberg, & Grigorenko, 2012; McClain & Pfeiffer, 2012; Nisbett, 2009; Rosemarin, 2001). In Israel, gifted students are identified as being in the 98.5th percentile.

Current thinking challenges the traditional, quantitative definition of giftedness, and therefore, researchers keep updating the definition, seeking for a more accurate set of characteristics of gifted children. Researchers now differentiate between the traditional definition of giftedness and a modern, multidimensional definition, which characterises intelligence via cognitive, qualitative, psychological, and social aspects (Gardner, 1982, 1983; Nevo & Chawarski, 1997; Renzulli, 1978; Tannenbaum, 1983; Sternberg, 1984).

Renzulli (1978) defined giftedness as a behaviour that would reflect interaction between above-average ability, task commitment, and creativity. In order to explain the current definition of giftedness, Renzulli (1999) used the three-ring circle illustration (Baldwin, 2005), where the three important factors are *talent*, *creativity*, and *aptitude*, and the overlapping areas indicate giftedness. Later, Reis and Renzulli (2009a) concluded that the most dangerous myths regarding giftedness are that the group of gifted and talented students is homogenous and that giftedness remains over time and is a subject to an individual's characteristic. Furthermore, a few studies suggest that giftedness is not only inborn, but can also be learned (Reis & Renzulli, 2009a, 2009b).

Domain-specific giftedness

More recently, researchers have established that giftedness is not a general ability but rather discipline- or domain-specific, hence the importance of developing assessment in science education. According to VanTassel-Baska (2005), giftedness is the demonstration of 'general intelligence in a specific domain of human functioning at a level significantly beyond the norm such as to show promise for original contributions to a field of endeavor' (p. 359). Other researchers focused on studying giftedness in specific domains, such as mathematics, verbal abilities, and science. For example, Brody and Stanley (2005) described the Study of Mathematically Precocious Youth (SMPY), which initially included also scientific knowledge. The project emphasised searching gifted and talented students, establishing programmes to foster and study their abilities.

Gagné (2005) claimed that one can monitor intelligence by observing a chemist's scientific reasoning or analysing chess player games. VanTassel-Baska (2005) noted that many states adopted the Common Core State Standards (CCSS), enabling teachers to nurture gifted students to meet their learning needs. She also emphasised that learners' giftedness develops over time through the interaction with the educational environment. Therefore, assessment and comparison of gifted students' thinking skills should take place at the start of the programme and one year into the programme.

The need for a gender-fair assessment

Sternberg (2010) noted that the world of standardised testing of gifted students has remained the same for almost a century, since the works of Binet and Simon (1916, as cited by Sternberg, 2010). Although IQ tests are the most widely accepted predictors of giftedness, obviously there are other concepts of intelligence that this traditional assessment tool cannot measure (Baldwin 2005; Gardner, 1983; Sternberg, 1984). Therefore, many experts do not recommend the use of intelligence tests as the sole predictor of giftedness. Yet, identification of gifted students in Israel followed the traditional definition, basing it on national tests similar to IQ tests (Dori et al., 2009; Fischer-Shachor et al., 2010; Rosemarin, 2001; Silverman, 1986).

Criticism towards IQ tests continues as Ford, Harris, Tyson, and Trotman (2003) and others (Baldwin, 2005; Ford et al, 2003; Gallagher, 2005) have claimed that these tests are culturally biased, as they are more effective at identifying middle-class white students rather than minorities and lower socio-economic classes. More research is needed to ensure the development of adequate assessment methods to identify giftedness for the appropriate educational programmes and select a more diverse population of gifted students (Pierce et al., 2006; Neumeister, Adams, Pierce, Cassady, & Dixon, 2007). Our research focus in this study with respect to diversity was gender equity.

O'Neil (1992) suggested performance assessment as an alternative to the quantitative tools for identifying giftedness. Performance assessments help to get a complete picture of the student's abilities and thinking skills while measuring students as they attempt to solve novel problems, work collaboratively in groups, and synthesise knowledge across disciplines. Kornilov et al. (2012) noted that experts are not always able to identify children with high intellectual abilities using a narrow set of assessments tools, raising the need for an assessment approach that extends the scope of measures of giftedness while also being gender-fair.

Affirmative action for gifted students

Affirmative action (AA) is a policy that articulates a legal norm, which governments pursue for providing equal opportunities for minorities or disadvantaged groups due to their status relative to the majority group (Public Law, 1964, Parodi, 2003; Seelke, 2008). The term *affirmative action* appeared officially for the first time in Executive Order 10925, signed by President John F. Kennedy in 1961 and used with respect to discrimination in employment to ensure that employees be treated equally, regardless of their race, belief, colour, or national origin (Report to the President, 1995). Literature on AA suggests that normative or ethical difficulties rise from implementation of AA policies, the main one being violation of the right to equality, based on the freedom of competition and personal right to excellence. Furthermore, AA policies raise resistance for two reasons: first, the extent to which it is legitimate to deviate from equal allocation of resources, and second, justification of application of such action in favour of a specific group, such as female students.

The Bollinger decisions provided guidelines regarding how institutions can apply AA, including the legality of admission quotas implemented to increase diversity on campuses (Solórzano & Yosso, 2002). Researchers such as Berry (2004) raised questions related to the effects of AA, whether the consequences of AA outweigh its benefits, and whether AA provides gifted girls with an opportunity to gain from specially designed enrichment programmes. In this paper, we describe a study, in which we developed a new assessment approach and a supporting instrument to make the testing process more gender-fair and then measured the effectiveness of this instrument for future screening. AA started as an attempt to reduce race and class inequality in the society, it later transitioned to issues of equality in higher education and workplaces (Harrison, Kravitz, Mayer, Leslie, & Lev-Arey, 2006; Sander, 2004). However, the debate regarding AA is beyond the scope of this paper, as we are focusing on gender equity and on the effort to increase gifted girls' participation in gifted programmes. We argue that if assessment will be more gender-fair, the need for AA will decrease significantly.

Willingham and Cole (1997, p. vii) posed the following two questions: 'Do we know enough about gender differences and similarities to know where our concerns should lie?' and 'If we know enough, what should we do with the knowledge in order to design fair assessments for the future?' In a more recent publication, the same authors (Willingham & Cole, 2013) pose more specific questions, including whether multiple-choice questions favour males and why do girls have better grades than boys in school, while men score higher than women on tests. Most importantly, do the answers to these questions depend on what is assessed and how, and if so, how do we make sure that the tests are designed and used fairly?

Real equity manifests itself when the outcomes indicate that success is distributed across segments of the population roughly according to their relative share. Specifically, since females and males are in equal numbers, we deem a gender-fair allocation of resources, e.g. for gifted programmes, to be equally distributed between boys and girls. For many years, the percentage of gifted girls, as identified by the Israeli national IQ-based testing, was less than 30% (Dori et al., 2009). Guided by the pressing need to change this situation and identify gifted students' diverse capabilities that comply with the multidimensional definition of giftedness, the IMoE has introduced AA. This act aimed at increasing the percentage of girls in various enrichment programmes for the gifted. However, policy makers were concerned that the gifted female students admitted through AA might need to struggle more, achieve lower grades, and experience higher attrition rates. Therefore, we deemed it valuable to investigate whether there were differences between the various groups of gifted students in scientific thinking skills before and during the gifted programmes.

We were motivated by the fact that some gifted learners demonstrate inclination for studying science early on during their elementary education. Our assumption was that if we succeed in identifying these gifted students and nurture their ability to perform well in science, it is likely that they would choose to materialise their domain-specific potential. Providing students with adequate learning opportunities that match their predispositions can improve their science learning outcomes. For gifted girls, this encouragement is more critical because they are less inclined to elect science due to societal pressures.

Given this state of affairs, we have designed, validated, and implemented a gender-fair assessment instrument for gifted students at the elementary school level. Our instrument is a case-based narrative followed by open-ended questions that emphasise higher order scientific thinking skills (Dori, 2003; Dori & Sasson, 2008; Dori et al., 2009; Zohar &

Dori, 2003). This questionnaire aligns with the emphases set forward by the NRC (2012) with respect to scientific practices, also known as scientific thinking skills. It includes authentic problems involving verbal, mathematical, and scientific aspects relevant to students' daily lives. This case-based questionnaire served as our main research tool. The research questions were as follows:

(1) To what extent is the admission of young students to the gifted programmes gender-fair as reflected by assessment via:
    (a) standardised tests
    (b) overall scientific thinking skills as reflected by students' responses to the case-based questionnaire?
(2) What differences, if any, exist between the three research groups of gifted students with respect to each one of the following scientific thinking skills: question posing, explanations, graphing, inquiry, and metacognition?

As we elaborate in the Participants sub-section (see below), the three research groups are gifted boys (GB), gifted girls (GG), and gifted girls who were admitted to the gifted programmes through affirmative action (GGAA).

Method

In Israel, educational settings in regular classes do not usually fit the gifted students' cognitive, affective, and social needs. Several educational alternatives for gifted students include specialised schools for those who excel in mathematics and science or art, advanced classes or pullout programmes for the gifted.

Settings

Identification of gifted children in Israel relies on cognitive testing, similar to IQ, conducted by the Jerusalem-based National Institute for Research in the Behavioral Sciences, Henrietta Szold Institute. Children in the second or third grade are tested for cognitive abilities in two stages. First, all students are tested in their schools. Then, Szold Institute administers additional tests to the students who had scored in the top 15th percentile in the first stage. Young students screened and determined to be in the top 1.5–2% can attend regional gifted programmes, where they participate in the pullout programmes or the dedicated classes for gifted students. Students in the pullout programme spend one day a week in a gifted class and the rest of the week in their organic (homeroom) class with other (non-gifted) students. In the dedicated class, all the students are gifted, and this is their organic class where they can study at an accelerated pace. Since these classes are situated in regular schools, the gifted students have the opportunity to socialise with peers from the regular classes. In this study, we examined both the pullout programme and dedicated classes for the gifted. The domains of the courses in both programmes included science, mathematics, art, and liberal arts. Towards the end of each school year, the gifted students prepared projects, which were authentic learning outcomes that served for assessing the students' performance. Appendix A lists gifted students' project topics by domain and type, and Appendix B contains two examples of projects, one in mathematics and the other in science.

Participants

Prior to the main research, 234 Israeli gifted students participated in the first round of the pilot study, of whom 159 were GB and 75 were GG. Additional 100 students participated in a second round of the pilot study. In the main research, the participants included 483 Israeli gifted students. The students were divided into three groups (see Table 1) based on the Szold tests results: GB ($N = 275$), GG ($N = 128$), and affirmative action gifted girls – GGAA ($N = 80$). We sampled the research participants from the gifted programmes supported by the IMoE in a way that ensured representation of a variety of geographical regions, both rural and urban, as well as various socio-economic sectors. We also made sure to include students from the two types of the gifted programme (pullout or dedicated class), as well as various age levels. About half of the students studied in the pullout and the other half in a dedicated class. Approximately half of the students were young (aged 8–10 years), from third and fourth grades, and they participated in the gifted programmes for the first year. The rest were older (aged 10–12 years), fifth and sixth graders, in their third year of participation in the gifted programme.

Research plan

Prior to the main research, we conducted a pilot study with 334 students, who did not participate in the main research. The objective of the pilot study was to examine and validate the research instrument – the case-based pre- and post-questionnaires. Based on Stemler and Tsai (2008), we decided to use inter-rater consensus estimates. The objective of the calculated inter-rater reliability score was to establish reliability for our newly developed scoring rubric, as specified below. We also performed content validity and predictive validity (Creswell, 2014; Heffner, 2014). The content validity involved a steering committee appointed by the Chief Scientist of the IMoE. The committee, which included science and mathematics experts from various universities in Israel and gifted students experts from the IMoE, validated both the questionnaire content and the rubric for assessing students' responses. Specifically, the committee examined the two questionnaire versions

Table 1. Gifted students who responded to the pre-questionnaire.

Student characteristic	Value	*N*	%
Age level	Grade 3	195	40.4
	Grade 4	66	13.6
	Grade 5	150	31.1
	Grade 6	72	14.9
Research group			
	Gifted girls (GG)	128	26.5
	Girls admitted to the gifted programme with affirmative action (GGAA)	80	16.6
	Missing data	4	0.9
	Total girls	208	43.1
	Gifted boys (GB)	275	56.9
Gifted programme	Pullout	279	57.8
	Dedicated class	204	42.2
Total		483	

administered in the two stages of the pilot for each age group. They also went over and approved the modifications we made to a few questions in both the pre- and post-questionnaires, as well as the scoring rubric. The predictive validity was performed by investigating the correlations between students' school achievements in science and mathematics and their scores in the pre- and post-questionnaires, as explained in the Findings section.

All the gifted students who participated in the research had taken the standardised tests in order to participate in the gifted programmes. The standardised national test scores, provided by the Szold Institute, served for answering the first research question. We then administered the case-based pre-questionnaire to both young (third and fourth graders) and older gifted students (fifth and sixth graders). Finally, a year and a half later, we administered the case-based post-questionnaire only to those students who had been third or fourth graders in the initial assessment (the young gifted students), since most of the older students already finished the elementary school and graduated from these gifted programmes.

Table 1 presents the distribution of the students who responded to the pre-questionnaire by age level, research group, and programme type. Thick lines separate the characteristics, such that the distribution within each characteristic sums up to 100%.

Table 2 presents the student population for the post-questionnaire.

Data collection and analysis

We used two data sources: (a) the Szold standardised national test and (b) case-based questionnaires for gifted students.

Identification of gifted students: the Szold standardised national test

The Szold Institute has designed and administered the standardised tests for identifying gifted students in Israel since 1984. All the students in second and third grades take a test, which consists of two stages described above. This test is a standardised IQ test, consisting of several parts (see Figure 1) including mathematics, reading comprehension, sentence completion, general knowledge, and geometrical patterns, with each part of the test timed individually. Those who score in the top 1.5 percentile are eligible to attend one of the gifted programmes in the following year.

Table 2. Young gifted student population in the post-questionnaire.

Student characteristic	Value	*N*	%
Age level	Grade 3	136	68.7
	Grade 4	62	31.3
Research			
Group	Gifted girls (GG)	49	24.7
	Girls admitted to the gifted programme with affirmative action (GGAA)	39	19.7
	Missing data for boys*	3	1.5
	Total girls	88	44.4
	Boys	110	55.6
Gifted programme	Pullout	74	37.3
	Dedicated class	123	62.7
Total		198	

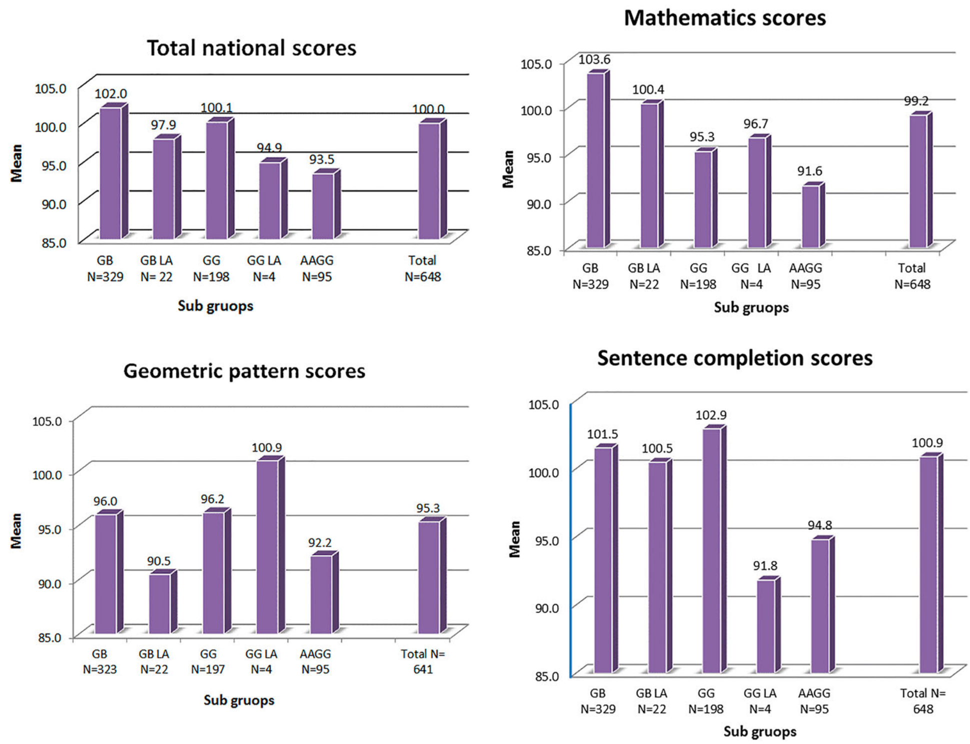

Figure 1. Students' scores of the total and three parts of the standardized national test by groups.

Case-based questionnaires for gifted students

In order to assess students' scientific thinking skills, we developed two different versions of pre- and post-questionnaires, one for young and another for older students. These case-based questionnaires, designed to assess the student's question posing, explanation, graphing, inquiry, and metacognitive skills, included short authentic stories – case-based narratives related to students' daily lives with follow-up assignments. The case-based questionnaires were adapted to young students from similar questionnaires for middle- and high-school students developed for and implemented in previous studies (Dori & Herscovitz, 1999; Dori & Sasson, 2008; Dori et al., 2009; Zohar & Dori, 2003; Zohar & Nemet, 2002).

The case-based questionnaires consisted of two parts. Part A included a case-based scenario that required understanding of a story and analysis of events. In this part, we examined three thinking skills: verbal – providing explanation, scientific – posing a complex question, and visual-mathematical – demonstrating graphing skills. Part B included another case-based scenario, which described an experiment and included assignments aiming to assess scientific inquiry and metacognitive skills.

We developed a rubric for each question. A team of four researchers, who are experts in science education, scored 100 students' responses to questions in the pilot stage. The inter-rater reliability was 85% in the first round, accounting for a total agreement of 72 responses, agreement by 3 of the 4 raters on 9 responses, and agreement of 2 by the 4 raters on 4 responses. Based on these results, we modified the questionnaires and validated them in a second round of the pilot study by administering them to other 100 gifted students. The students in the pilot had studied in similar gifted programmes but did not participate in the main study. Analysis of these students' responses also served to improve the questionnaires further. The inter-rater reliability in the second round of the pilot study was 90%. Below are examples from Part A and Part B of gifted students' responses in the main study with rubric samples for the various thinking skills.

The **Snack Time** case-based scenario (Case A) below is an example from Part A for young students, followed by an assignment designed to examine students' *question posing skill*.

Case A: Snack Time

> Many children like to have snacks, as dessert, between meals, on field trips or during recess. Snack options are unlimited – they can be sweet or salty, cubic or cylindrical, large or small, in personal or family-size packages, single or multi-colored, made of corn or potato, low-cost or expensive.
>
> Having read the **Snack Time** story, what made you curious? What else would you like to know about the subject discussed? Please write at least two questions.

Table 3 presents a sample of questions posed by students and the scoring rubric used to examine students' responses to Part A based on the following categories:

(1) *Complexity* refers to the extent of elaboration in the students' response. A low score was given for an answer that can be found within the story; a medium score was given when the posed question required knowledge beyond the text presented; and a high score was given when the question included several domains of knowledge or interest.
(2) *Phrasing* refers to clarity and style with which the student expressed response.

We assessed *explanation* using a students' assignment, in which we asked them to explain or justify a given statement. In the questionnaire administered for the young students, students were first required to suggest items and then justify their choice. In the older students' questionnaire, we asked students to demonstrate their explanation and

Table 3. Examples of two young students' responses to the question posing assignment in Part A.

Age level	Q. no.	Questions posed by the students	Complexity	Phrasing	Relation between questions[a]	Total score (before normalised)
Fourth grade	1	What kind of snack is most popular among children?	2	2	3	22
	2	Why do adults prefer spicy snacks?	3	2		
	3	Which snacks are considered healthier? Which snacks most suit diets?	3	2		
	4	What do people eat more – food or snacks? Does this vary by age?	3	2		
Third grade	1	Do snacks have a high percentage of fat?	2	2	3	13
	2	How are snacks given names?	2	2		
	3	Are snacks good for little kids?	1	1		

[a]This criterion refers to the extent to which the questions that the student provided build on top of each other.

explanation skills for items presented by researchers. The questionnaires for the older students were more difficult as they were required to justify a statement from a provided list, as opposed to a self-suggested item. An example of a question from the older students' questionnaire follows:

> Much information appears on the package of a popular snack. Beside each piece of information, write down why you think it is given.

An example of a sixth-grade student's high-level responses to the explanation assignment included explanation for (a) the 'Enriched with vitamins' information: 'To appeal to parents who believe vitamins are good for their children, so they will buy the product.' (b) The 'Store in a cool and dry place' instruction was: 'Important information for storing the snack in the most appropriate place to keep it fresh and fully enjoy it.', and (c) the 'Amount of energy per serving' information was: 'Important information, for the consumer to know the value of what they are eating.'

In the assignment designed to examine *graphing skills*, we provided students with data about the number of students participating in various after-school sports activities: swimming – three boys and five girls, basketball – four boys and four girls, and tennis – six boys and four girls. The assignment that followed was:

> Draw a graph, which describes the number of boys and girls combined who participated in the after school activities in your class.

The rubric for this skill included criteria relating to whether the student answered the question, processed the data correctly, drew an appropriate graph, and included scale and units in his/her graph.

After completing this assignment, we asked students to design an experiment: 'You have been tasked with a mission to discover whether eating snacks affects your health. Think what you have to do in order to carry out this task and write down all the activities according to their order.'

The best answer included elements of collecting data, analysing them or experimenting, and implied reference to drawing conclusions.

Below we provide an example of a high-level response written by an older child.

> A. go and purchase some snack, B. look at the back of the package and see the products [ingredients], C. after seeing the products we check them, D. check ourselves to see how we feel before eating the snack, E. eat the snack, F. try to check how we felt after eating the snack.

The question that followed that is presented below along with the response of the same student (who answered the previous question) called for demonstrating a *metacognitive skill*. 'Read again the previous question, which deals with the process of designing. Try to remember what you thought when you answered the question and write down: How do we plan?' The students responded: 'First of all we think, then we prepare a draft, then we write, and finally we draw conclusions.'

We asked: 'Why is it necessary to use a plan?' and the answer was: 'It is necessary to use a plan because if we will not use a plan we can get confused in one of the steps and then our entire experiment will be in vain.' We also asked the following question: 'What was easy and what was difficult in the planning process?' The student's response was: 'It was easy for me to think and come up with ideas. It was difficult for me to formulate all the ideas and write them down neatly according to the steps.'

The case-based scenario in Part B (Case B) aimed to investigate students' *inquiry and metacognitive* skills. Below is the inquiry-based narrative, followed by rubric exemplified by responses of an older student.

Case B: Washing Hands

Dan came home after participating in a sports activity. He was hungry and wanted to have a snack before dinner. He went to the pantry and helped himself to a snack.

"Did you wash your hands?" his mother asked.

"Yes," replied Dan and began to bite the snack.

"With soap?" mom asked.

"What does it matter?" Dan asked and then added: "I am just eating a small snack, not a whole meal … "

"Of course it matters! Before putting something in your mouth, a snack or a meal, you must wash your hands with soap and water. If you do not wash your hands with soap, all the bacteria remains on your skin" mom replied.

Dan thought his mother was wrong. He was sure it does not matter if you wash your hands with soap or just wash them with water. He decided to investigate who is right – his mother or him.

Formulate a research question that Dan could ask. In other words, what is the question that he wants to investigate?

A set of questions followed this case, asking students to formulate a research question and a hypothesis, to plan an experiment, to draw a conclusion from fictitious experimental results and to reflect on their thinking. We constructed a detailed rubric for each of these questions. For instance, for scoring the 'formulate a research question' task, we used two different criteria: The extent to which the formulated question related to conducting a 'scientific' inquiry, and the extent to which this question included a correct reference to two variables: testing the effect of the independent variable (with or without soap) on the dependent variable (growth of bacteria).

The scoring scale for the second criterion (correct reference to the two variables) was as follows:

(a) *Correct reference to two variables (5 points)*: testing the effect of the independent variable (with or without soap) on the dependent variable (growth of bacteria). An example of such a 5-points student's question is 'Does soap have an effect on the amount of bacteria on the skin?'
(b) *Partial reference to the two variables (3 points)*: correct reference to the dependent variable (growth of bacteria), but not to the independent variable. An example of such a 3-points question is 'When washing hands with water, does bacteria remain on the skin?'
(c) *Reference to only one variable (1 point)*: reference to dependent or independent variable, or reference to two variables, but in a partial or wrong manner. An example of such a 1-point question is 'Is soap needed or not?'

If there was no reference to the scenario, or if the student formulated a research question that could not be investigated in a 'scientific' way, as in 'Is Dan right?' the item receives a score of zero.

Statistical analysis

We performed statistical analysis in four stages. First, we analysed each scientific thinking skill separately. Second, we aggregated the five skills into two parts according to the two case-based narratives. Case A examined three skills: question posing, explanation, and graphing skills. Hence, we combined these three scores into a total average score for Part A – Score A. Case B examined two other skills: inquiry and metacognitive skills. Third, the scores for these two skills were combined into a total score for Part B – Score B. Scores for A and for B were then converted into standardised (Z) scores and the total normalised score (A + B) was calculated. Use of standardised Z scores was necessary, since the two parts of the questionnaire differed in their difficulty level and their level of fatigue, which increased after responding to Part A. Henceforth, the Z score represents the normalised score of the entire case-based questionnaire. Finally, we carried out predictive validity of the questionnaires by finding correlations between the students' questionnaire total scores and the students' school grades in science and mathematics.

Findings

We present the findings according to the order of the research questions.

Gender-fairness of admissions to the gifted programmes – standardized tests

Research question 1 part (a) concerned the extent to which admissions of young students to the gifted programmes based on the standardised tests are gender-fair. Examining the ratio between boys and girls admitted to the gifted programmes during the course of four years, we found out that prior to the introduction of AA, the boys-to-girls ratio in both the pullout and dedicated class programme types was about 2:1, namely, two-thirds of the children who score in the top 1.5 percentile were boys. More gifted girls than boys who were eligible to join one of the gifted programmes decided to forego and not to join. This caused a decrease in the percentage of GG in the gifted programmes. To remedy this, the IMoE increased the number of girls admitted with AA (GGAA) to restore the number of the total girls in the gifted programmes, making the GG and GGAA together at least one third of the students in these programmes.

To gain better understanding of the test outcomes, both total and by parts, we present Figure 1, which shows the distribution of the standardised test scores, as provided by Szold Institute. While in the paper we focus on three groups – GB, GG, and GGAA, only when we present the data provided by this institute, the outcomes of the standardised test include two additional groups: gifted boys who were accepted via late admission[1] (GBLA) and gifted girls who were accepted via late admission (GGLA). These students were not admitted originally to the programme based on their test scores, and were admitted later, following appeals to the IMoE. Notably, the number of GBLA ($N = 22$) is more than five times larger than the corresponding number of GGLA ($N = 4$), providing an indication of the lower drive or motivation of girls to participate in the gifted programmes relative to boys.

The four graphs include the standardised national test total scores and scores of three individual parts of the standardised test: mathematics, geometrical patters, and sentence completion. The total standardised national scores (Figure 1 top left) show that the boys ($N = 329$) received the highest scores ($\bar{x} = 102.2$), followed by gifted girls (N Next are the LA boys ($N = 22$), followed by the LA girls ($N = 4$), and .($\bar{x} = 100.1$) =198) then the GGAA ($N = 95$), who scored the lowest ($\bar{x} = 93.5$).

Analysis of variance (ANOVA) of the scores showed significant differences between the groups of gifted students in the total standardised national scores ($F_{(4,643)} = 58.63$, $p < .0001$) as well as scores in each of the test parts. Significant differences between the groups were found in mathematics ($F_{(4,643)} = 21.76$, $p < .0001$), general knowledge ($F_{(4,643)} = 21.04$, $p < .0001$), sentence completion ($F_{(4,643)} = 10.43$, $p < .001$), and geometrical patterns ($F_{(4,636)} = 6.78$, $p < .0001$). Interestingly, in the sentence completion part, presented in the bottom right of Figure 1, the GG scored the highest, and in the mathematics and the geometrical patterns parts, the differences in scores between GG and GGAA were the smallest (see top right and bottom left in Figure 1).

Gender-fairness of admissions to the gifted programmes – overall scientific thinking skills

Research question 1 part (b) called for determining the overall scientific thinking skills exhibited by each one of the three gifted student groups. These findings stem from analysis of gifted students' responses to the case-based questionnaires, which we designed for this study in order to answer research question 1 part (b) and research question 2.

We first provide descriptive statistics of Part A and Part B of the questionnaires. We follow this with an analysis that includes the normalised (Z) scores of the overall assessed scientific thinking skills, correlation, multiple regression test, and finally repeated measures.

Analysis of the overall scores of the gifted students' case-based questionnaires

Analysing the young gifted students' scores ($N = 258$) in Part A and Part B of the pre-questionnaire, we found out that in Part A, $\bar{x} = 63.12$; SD = 12.91, while in part B, $\bar{x} = 53.10$; SD = 16.77. We attribute this 10-point difference between the first and second parts to two factors. One factor was the higher difficulty level of both Case B and the two scientific thinking skills required – inquiry and metacognition. The other factor was the fatigue of students, who were required to respond to Part B after having performed Part A, which had called for understanding Case A, which was shorter and simpler, and applying three different thinking skills – question posing, explanation, and graphing. This difference in students' overall scientific thinking skills prompted us to use Z scores to analyse the overall differences between student groups. Distribution of the scores in the post-questionnaire was similar. Moreover, the correlation between Part A and Part B of the pre-questionnaire, as computed by the Pearson correlation test was high and significant ($N = 466$, $p < .01$, $r = 0.444$). The 466 gifted students who responded to the pre-questionnaire included both young and older elementary school students.

A similar high and significant correlation was found between the scores of the young students who answered Part A and Part B in the post-questionnaire (see Table 4).

Students' overall scientific thinking skills Z scores in the questionnaires and school Z scores

To establish the predictive validity of the scientific thinking skills questionnaires, we calculated the students' *Z* scores in the pre- and post-questionnaires and obtained the students' school grades in science and mathematics. We then calculated the average *Z* scores of these school grades. Performing Pearson correlation, we found that correlation between the scientific thinking skills *Z* scores in the pre-questionnaire and the average school *Z* scores is $r = 0.145$, $p < .05$. The regression test showed that the school *Z* scores predict the pre-questionnaire *Z* scores by 0.143 units ($p < .05$). The correlation between the post-questionnaire *Z* scores and school *Z* scores was $r = 0.331$, $p < 0.005$. The school *Z* scores predict the post-questionnaire *Z* scores by 0.361units ($p < .001$).

We also found that the pre-questionnaire *Z* scores predict by 0.229 units the post-questionnaire *Z* scores ($p < .05$).

Students' overall scientific thinking skills Z scores in the pre- and post-questionnaires

While the two age groups responded to the pre-questionnaire, only the young students took the post-questionnaire. Therefore, Table 5 presents the participants' pre- and post-questionnaire *Z* scores of the three research groups only for the young gifted students. Merging the two parts of the pre- and post-questionnaires, the number of young students decreased from 198 to 175 because we took in account only those students who responded to both parts of both the pre- and post-questionnaires.

ANOVA test of the young students' pre-questionnaire *Z* scores showed that the groups were homogeneous, with no significant differences between GG, GGAA, and GB. This finding clearly indicates that our case-based questionnaire and the assessment approach

Table 4. Average total scores of young gifted students in the two parts of the questionnaire and their correlation.

Post-questionnaire average total score	*N*	$\bar{x}$	SD	Pearson correlation
Total score A	198	66.47	12.68	0.49**
Total score B	198	54.75	19.27	

Note: **Correlation is significant at the 0.01 level (two-tailed).

Table 5. Analysis of *Z* scores of young gifted students who participated in both the pre- and the post-questionnaire, sorted by questionnaire and research group.

Questionnaire	Group	*N*	$\bar{x}$*Z* score	SD
Pre-questionnaire	GG	46	−0.10	1.18
	GGAA	33	0.4	0.97
	GB	96	−0.09	0.89
	Total	175	-0.00	1.00
Post-questionnaire	GG	46	0.23	0.88
	GGAA	33	0.11	1.14
	GB	96	−0.17	0.980
	Total	175	−0.01	1.00

is gender-fair: Unlike the standardised tests which resulted in significant differences between the three groups of gifted students, our assessment did not show any significant difference between the groups. Comparing the *Z* scores of students in the dedicated classes and those of students attending the pullout programme showed no significant difference either. Table 6, which presents multiple comparisons of the difference of *Z* scores between the research groups, shows that at the end of the one-year gifted programmes, there were significant differences neither between GGAA and GG nor between GG and GB. The only borderline difference was found between GGAA and GB.

ANOVA of older students' *Z* scores in the pre-questionnaire also showed homogeneity with no significant differences between the three research groups.

In further analysis, we found significant *correlation* between the pre- and post-questionnaire *Z* scores ($p < .01$, $r = 0.322$) of the young gifted students. We also performed a *repeated measures test* (a) *within* subjects, where the measured variable was time – pre- vs. post-questionnaire Z score, and (b) *between* subjects, where the variable was research group – GB, GG, GGAA. We found no significant difference *within* subjects, but we did find a significant interaction between time and research group ($F_{(2,172)} = 3.204$, $p < .05$). This finding indicates that the difference between the pre- and post-questionnaire *Z* scores was different among the three research groups.

A *Post Hoc* test (Hochburg GT2) with multiple comparisons of the *Z* scores showed no significant difference between GG and the other two research groups. However, there was a borderline significant difference between GB and GGAA, where the mean difference between the *Z* scores (net gain) was 0.39 (SE = 0.16, $p = .056$). The boys' *Z* scores in the questionnaire were lower than GGAA's scores. This finding implies that girls accepted through AA improved the most from the pre- to the post-questionnaire over the course of about a year and a half of participating in the gifted programme.

Differences between groups of gifted students in specific scientific thinking skills

The second research question inquired whether there are differences between groups of gifted students in five specific skills – question posing, explanation, graphing, inquiry, and metacognition – as assessed from their responses to our specially designed case-based questionnaires.

Analysing the *explanation*, *inquiry*, and *metacognitive* skills, we did not find any significant differences between the three groups of gifted student (GB, GG, and GGAA). This was true for both young and older students in the pre-questionnaire, and for the young students in the post-questionnaire. For example, in the explanation skill, the pre-questionnaire scores of the older students were 59.2 for GB, 62.2 for GG, and 67.7 for GGAA; the pre-questionnaire scores of the younger students were 55.0 for GB, 56.0 for GG, and 59.7

Table 6. Multiple comparisons of the difference of *Z* scores between the research groups.

(I) Affirmative action group	(J) Other affirmative action	Mean Difference Net gain, (I-J) *Z* scores	S.E.	*p* <
GG	GGAA	−0.19	0.18	n.s.
	GB	0.20	0.15	n.s.
GGAA	GB	0.39	0.16	.056 borderline significant

for GGAA; and the post-questionnaire scores of these younger students were 59.3 for GB, 61.1 for GG, and 60.1 for GGAA.

As for the question posing and graphing skills, we did find significant differences, as specified in Table 7. Since the difference patterns of the two skills were inverse of each other, we discuss each skill separately.

Analysing the assignments aimed at examining the *question posing* skill, we found no significant difference in the pre-questionnaire mean scores between the three groups of young students. However, as Table 7 shows, we found a significant difference between the three groups in both the post-questionnaire of the young students and the pre-questionnaire of the older students. The young students took the post-questionnaire a year and a half later, before they had completed their second year in one of the gifted programmes. The older students took the pre-questionnaire after spending a similar period of two years in one of the gifted programmes. For both young and older students who had spent about two years in a gifted programme, the scores of GG (girls accepted to the gifted program without affirmative action) were significantly higher ($p < .0001$) than those of GB (gifted boys). Yet, there was significant difference neither between GG and GGAA nor between GB and GGAA.

Turning to the analysis of the *graphing skills* scores, we found significant differences between the three groups in the pre-questionnaire only for the young students (see Table 7). Significant differences between the graphing scores of the three groups were found neither in the post-questionnaire of the young students nor in the pre-questionnaire of the older students. Probing further, we found no significant difference between the pre-questionnaire graphing scores of young GGAA and BG, but the scores of GG were significantly lower than those of both young GGAA ($p < .01$) and GB ($p < .005$).

Table 7. Analysis of gifted students' scores of question posing and graphing skills.

Questionnaire and skill	Research group		*N*	$\overline{X}$	SE	df	*F*	*p* <
Pre-question posing	Young	GB	150	56.4	1.7	258	0.56	n.s.
		GG	65	59.8	2.5			
		GGAA	46	57.1	3.2			
		Total	261	57.4	1.3			
Post-question posing	Young	GB	110	64.1	2.1	195	6.43	.0001
		GG	49	76.8	2.3			
		GGAA	39	68.0	3.4			
		Total	198	68.0	1.5			
Pre-question posing	Old	GB	125	70.2	1.7	219	3.1	.05
		GG	63	77.3	2.2			
		GGAA	34	72.0	3.0			
		Total	222	72.5	1.2			
Pre-graphing	Young	GB	150	70.5	1.47	258	6.8	.001
		GG	65	61.2	2.56			
		GGAA	46	72.4	2.70			
		Total	261	68.5	1.18			
Post-graphing	Young	GB	110	69.4	1.9	195	0.52	n.s.
		GG	49	67.8	2.4			
		GGAA	39	72.0	3.2			
		Total	198	69.5	1.4			
Pre-graphing	Old	GB	125	68.6	1.6	219	0.25	n.s.
		GG	63	68.1	2.5			
		GGAA	34	68.9	3.4			
		Total	222	68.5	1.3			

Discussion

Traditionally, boys have made up about two-thirds of children in gifted programmes in the Israeli school system. This lack of balance in gender of the students admitted to the gifted programmes triggered the AA research about gender-fairness of the testing and selection procedures. To narrow the gender gap and recruit diverse population in gifted programme, it is important to uncover evidence of leveraging positive impact on scientific thinking skills at a young age.

In an effort to foster gifted young girls to be selected to and attend gifted programmes, our study sought to improve the means by which elementary school students are tested and admitted to gifted programmes. Our approach for gender-fair assessment follows the performance assessment proposed by O'Neil (1992), who called for adopting a broader view of gifted students' abilities and thinking skills via diverse alternative assessment means. Our focus has been on evaluating the current assessment system in Israel to identify gifted children and to determine whether the assessment is gender-fair, in line with what Willingham and Cole (1997, 2013) had proposed for the general population.

The IMoE recognised the need to evaluate the effect of AA in general and affirmative action for gifted girls (GGAA) in particular (Dori et al., 2009; Fischer-Shachor et al., 2010). A key point of contention, raised by educational policy makers, was that gifted girls who are admitted to gifted programmes through GGAA are less likely to succeed and more likely to withdraw from the programmes. Berry (2004) also questioned whether the consequences of GGAA outweigh the benefits that such an action could provide. The practice of acceptance to a gifted programme was based on quantitative measures of the Szold standardised tests, which were equivalent to an IQ test (Renzulli, 1978; Sternberg, 2010). Since the introduction of AA, some of the young girls who had not passed the threshold were still admitted into one of the two gifted programmes through the implementation of AA, even if their respective scores in that test were somewhat lower than the acceptance threshold. The gifted children programmes helped narrow the gap between GB and GG on one hand and GGAA on the other, indicating that AA was justified academically and that it is likely that there had been lack of gender-fairness in the gifted programmes admission process. All three groups received the same 'treatment', as they attended the same gifted programmes. If upon entering the programmes the academic level of the GGAA students was significantly lower, they would not be able to attain the same level as the GB and GG students at the end of the programme.

Our assumption was that assessing the scientific thinking skills (a.k.a. scientific practices; NRC, 2012) can serve for identifying diverse abilities of gifted students that are in line with the modern, multidimensional definition of giftedness. In an effort to ensure that boys and girls are assessed fairly (Harrison, Kravitz, Mayer, Leslie & Lev-Arey, 2006; Sander, 2004), we designed case-based questionnaires, which evaluate a variety of higher order thinking skills. These skills called for diverse abilities of students: scientific – question posing and inquiry, verbal – providing explanation, visual-mathematical – graphing, and metacognition (Dori, 2003; Baldwin, 2005; Frasier, Garcia & Passow, 1995; Zohar & Dori, 2003).

The objective of including case-based narratives that are relevant to students' daily lives and open-ended questions in these questionnaires was to increase gender-fairness in the process of determining giftedness, by expanding the repertoire of tests. This is in line with

Baker and Leary (2003), who determined that girls become more engaged in science when it is relevant to their everyday life, and with Pine et al. (2006), who showed that girls perform in inquiry as well or better than boys do.

Stobart, Elwood, and Quinlan (1992) called for taking a 'hard look' at standardised examinations in which there are clear gender performance differences, most noticeably in English, mathematics, science, and modern languages. Other researchers, who examined cognitive patterns of sex differences in solving Graduate Record Examination (GRE) math problems, found differences favouring males for problems requiring spatially based solution strategies, but not when the strategies were verbal or similar to those in math textbooks. They found that the advantage boys have in standardised math tests can be minimised, equated, or maximised based on how the problems are presented and on the cognitive processes that need to be applied for solving these problems (Gallagher, Levin, & Cahalan, 2002). Along this line, our questionnaire includes open-ended questions and a variety of scientific thinking skills for promoting gender-fairness. Indeed, we found that the specific scientific thinking skills of girls were as good as or better than their boy peers. Moreover, the thinking skills of GGAA did not fall short of that of their GG peers.

Our findings indicated that GG and GGAA performed the same as or better than the GB in thinking skills tasks. These results demonstrate that effective assessment can and should be based on case-based questionnaires with open-ended questions that had been validated and shown to address the need for gender-fairness.

Limitation and further research

While our purpose was to develop a new assessment approach to improve gender-fairness, our sample consisted of students already identified by the existing instruments and a subset of girls who admitted by AA. We did not conduct the assessment with students who were tested but were not accepted to a gifted programme, since the IMoE did not approve conducting such study that encompassed these students to protect their privacy. Since we did not receive the IRB approval, we cannot determine if our case-based questionnaire would enable a more gender-fair assessment for other students. In previous studies (Dori, 2003; Dori & Herscovitz, 1999; Dori & Sasson, 2008; Zohar & Dori, 2003), we examined middle- and high-school students' scientific thinking skills using similar questionnaires. However, further research in the age group of young and older elementary school students is needed.

In response to our first research question regarding the gender-fairness of gifted programmes admissions, our findings showed that the current national standardised (IQ-like) tests in themselves lack gender-fairness, especially when administered to young students as the sole mean for identifying eligibility for gifted programmes.

Another limitation of this research is that AA is a broad and theoretical question well beyond the scope of this research. However, our focus on gender equity in the context of assessing scientific thinking skills might improve the rate of girls' participation in gifted programmes and consequently in science careers.

With respect to the second research question, regarding differences between groups of gifted students in specific skills, we found significant differences only in question posing and graphing skills. In question posing, there were no significant differences between the three research groups before the students started the programme but after more than a year, the girls significantly outperformed the boys. In the graphing skills, the significant differences that had existed between the three groups narrowed and became insignificant. Thus, we showed that gender differences are not just specific to particular scientific domains, such as physics versus biology (Baker, 2002; Kahle, 2004), but also that they vary from one scientific thinking skill to another.

Gifted students do not always outperform in all domains of study (Reis & Housand, 2007; Renzulli, 1978; Reis & Renzulli, 1999). Some of them exhibit various talents and skills in different domains and at different times. To ensure gender-fairness, we designed our assessment tool to analyse scientific thinking skills of gifted students, regardless of their gender. Comparing the pre-questionnaire scores to the post-questionnaire scores, we found significant improvement amongst both groups of girls – GG and GGAA, with GGAA improving the most. Analysing the young gifted students' overall scientific thinking skills in the pre- and post-questionnaires, we found a significant interaction between time and research group. There was no significant difference between GG and the other two research groups. However, there was a borderline significant difference between GB and GGAA. The boys' overall scientific thinking skills were lower than that of the GGAA, implying that girls accepted through AA improved the most from the pre- to the post-questionnaire during the year and a half of attending the gifted programme.

Research contribution

This research contributes to the body of knowledge in the domain of gender equity and gender-fair assessment for students in general and for gifted students in particular. In the domain of assessment, we propose a gender-fair research tool for analysing higher order thinking skills for elementary school gifted students. Practically, we have developed, validated and implemented a case-based questionnaire that can serve for assessing gifted and other young students in five thinking skills. We suggest administering this questionnaire or a similar one in addition to standardised tests, as the latter are typically focused on multiple-choice questions. Extending the instrument to include additional higher order thinking skills, such as synthesis and creativity, is a subject of future research, as is further generalisation of our findings to other countries. The development of different assessment tools for analysing students' scientific thinking skills is a continuous process. This study can potentially help educators and decision makers understand the challenges of fairly identifying gifted students, the role AA can play, and how specially designed assessments might help evaluate and reduce the gender gap. Our study shows that such measures might better serve both research and implementation by differentiating between gifted and non-gifted students and identifying GG and GB equally well.

Finally, yet importantly, the GG are the potential educators, scientists, and engineers who can undertake leadership roles to advance society in general and the status of women in particular. It is therefore of paramount importance to do our best to include as many girls as there are boys in the gifted programmes, not by implementing AA, which is justifiably often controversial, but through revising the testing system so it becomes more gender-fair.

Note

1. In our research, these late admission students (GBLA and GGLA) were merged into their corresponding GB and GG groups.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by Office of the Chief Scientist, Ministry of Education, Jerusalem, Israel [TRDF grant number # 200-5370].

References

AIP – American Institute of Physic. (2013). Women in Physics and Astronomy. Retrieved from http://www.aip.org/statistics/trends/gendertrends.html

Baker, D. (2002). Editorial: Where is gender and equity in science education? *Journal of Research in Science Teaching*, *39*, 659–663.

Baker, D., & Leary, R. (2003). Letting girls speak out about science. *Journal of Research in Science Teaching*, 40, (Supplement) S176–S200 (Previously was published in 1995).

Baldwin, A. Y. (2005). Identification concerns and promises for gifted students of diverse populations. *Theory Into Practice*, 105–114.

Banks, J. A., & Banks, C. (Eds.). (1995). *Handbook of research on multicultural education*. New York, NY: Macmillan.

Berry, R. M. (2004). Affirmative action in higher education: Costs, benefits, and implementation. *Journal of Public Budgeting, Accounting & Financial Management*, *16*(2), 257–276.

Binet, A., & Simon, T. (1916). *The development of intelligence in children: The Binet–Simon scale*. Publications of the Training School at Vineland New Jersey Department of Research No. 11. (E. S. Kite, Trans.).Baltimore, MD: Williams & Wilkins(Cited by Sternberg, 2010).

Brody, L. E. & Stanley, J. C. (2005). Youths who reason exceptionally well mathematically and/or verbally using the MVT: D^4 model to develop their talents. In R. J. Sternberg & J. E. Davidson (Eds.), *Conceptions of giftedness* (2nd ed., pp. 20–37). Cambridge: Cambridge University Press.

Central Bureau of Statistics. (2011). Retrieved from http://www1.cbs.gov.il/shnaton65/st12_19x.pdf

Central Bureau of Statistics. (2013). Israel in Figures 2013. Retrieved from http://emetnews.org/documents/israel-in-figures-2013.pdf

Creswell, J. W. (2014). *Research design qualitative, quantitative, and mixed methods approaches* (4th ed., p. 304). Thousand Oaks, CA: SAGE Publications.

Dori, Y. J. (2003). From nationwide standardized testing to school-based alternative embedded assessment in Israel: Students' performance in the "Matriculation 2000" Project. *Journal of Research in Science Teaching*, *40*(1), 34–52.

Dori, Y. J., & Herscovitz, O. (1999). Question posing capability as an alternative evaluation method: Analysis of an environmental case study. *Journal of Research in Science Teaching*, *36*(4), 411–430.

Dori, Y. J., & Sasson, I. (2008). Chemical understanding and graphing skills in an honors case-based computerized chemistry laboratory environment: The value of bidirectional visual and textual representations. *Journal of Research in Science Teaching*, *45*(2), 219–250.

Dori, Y. J., Zohar, A., Fischer-Shachor, D., Kohan-Mass, J., & Carmi, M. (2009). Report published by the Ministry of Education, Chief Scientist Office (200 pages in Hebrew), Jerusalem, Israel.

Eglash, R. (2010). Status of Israeli women improves little over decade Israel. Women's Network – IWN (2010). Retrieved from http://www.jpost.com/Israel/Article.aspx?id=170370

Fischer-Shachor, D., Carmi, M., & Dori, Y. J. (2010). Question posing and graphing skills of young gifted students: Gender and equal opportunities aspects. In G. Cakmakci & M. F. Taşar (Eds.),

Contemporary science education research: Learning and assessment (pp. 409–414). Ankara: Pegem Akademi.
Ford, D. Y., Harris, J. I., Tyson, C. A., & Trotman, M. F. (2003). Beyond deficit thinking: Providing access for gifted African American students. *Roeper Review*, *24*(2), 52–58.
Frasier, M. M., Garcia, J. H., & Passow, A. H. (1995). *A review of assessment issues in gifted education and their implications for identifying gifted minority students (RM95204)*. Storrs, CT: The National Research Center on the Gifted and Talented, University of Connecticut.
Gagné, F. (2005). From gifts to talents. The DMGT as a developmental model. In R. J. Sternberg & J. E. Davidson (Eds.), *Conceptions of giftedness* (2nd ed., pp. 98–119). Cambridge: Cambridge University Press.
Gallagher, J. J. (2004). No child left behind and gifted education. *Roeper Review*, *26*(3), 121–123.
Gallagher, J. J. (2005). The role of race in gifted education. *Roeper Review*, *27*(3), 135.
Gallagher, A., Levin, J., & Cahalan, C. (2002). *Cognitive patterns of gender differences on mathematics admissions tests*. *ETS Research Report Series*, *2002*(2), i-30.
Gardner, H. (1982). *Art, mind, and brain: A cognitive approach to creativity*. New York, NY: Basic Books.
Gardner, H. (1983) *Frames of mind: The theory of multiple intelligences*. New York, NY: Basic Books.
Harrison, D. A., Kravitz, D. A., Mayer, D. M., Leslie, L. M., & Lev-Arey, D. (2006). Understanding attitudes toward affirmative action programs in employment: Summary and meta-analysis of 35 years of research. *Journal of Applied Psychology*, *91*(5), 1013–1036.
Haussler, P., & Hoffmann, L. (2002). An intervention study to enhance girls' interest, self-concept and achievement in physics classes. *Journal of Research in Science Teaching*, *39*(9), 870–888.
Hazari, Z., Sonnert, G., Sadler, P. M., & Shanahan, M. C. (2010). Connecting high school physics experiences, outcome expectations, physics identity, and physics career choice: A gender study. *Journal of Research in Science Teaching*, *47*(8), 978–1003.
Heffner, C. L. (2014). Variables, validity, and reliability. Chapter 7 in research methods. Retrieved from https://allpsych.com/researchmethods/variablesvalidityreliability/
Huang, G., & Du, J. (2002). Computer use at home and at school: Does it relate to academic performance? *Journal of Women and Minorities in Science and Engineering*, *8*, 201–217.
Hughes, W. (2002). Gender attributions of science and academic attributes: An examination of undergraduate science, mathematics and technology majors. *Journal of Women and Minorities in Science and Engineering*, *8*, 53–65.
Kahle, J. B. (2004). Editorial: Will girls be left behind? Gender differences and accountability. *Journal of Research in Science Teaching*, *41*(10), 961–969.
Kornilov, S. A., Tan, M., Elliott, J. G., Sternberg, R.J., & Grigorenko, E.L. (2012). Gifted identification with Aurora: Widening the spotlight. *Journal of Psychoeducational Assessment*, *30*(1), 117–133.
Lavy, V., & Schlosser, A. (2011). Mechanisms and impacts of gender peer effects at school. *American Economic Journal: Applied Economics*, *3*(2), 1–33.
Marland, S. P., Jr. (1972). *Education of the gifted and talented: Report to the Congress of the United States by the U.S. Commissioner of Education and background papers submitted to the U.S. Office of Education* (2 vols). Washington, DC: U.S. Government Printing Office.
Mcclain, M.-C., & Pfeiffer, S. (2012). Identification of gifted students in the United States today: A look at state definitions, policies, and practices. *Journal of Applied School Psychology*, *28*(1), 59–88.
National Science Board. (2012). *Science and engineering indicators 2012*. Arlington, VA: National Science Foundation(NSB 12-01). Retrieved from http://www.nsf.gov/statistics/seind12/pdf/seind12.pdf
National Science Board. (2014). *Science and engineering indicators 2014*. Chapter 2: Higher Education. Arlington VA: National Science Foundation (NSB 14-01). Retrieved from http://www.nsf.gov/statistics/seind14/
National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academy of Sciences.

Neumeister, K. L. S., Adams, C. M., Pierce, R. L., Cassady, J. C., & Dixon F. A. (2007). Fourth-grade teachers' perceptions of giftedness: Implications for identifying and serving diverse gifted students. *Journal for the Education of the Gifted*, *20*(30), 479–499.
Nevo, B., & Chawarski, M. C. (1997). Individual differences in practical intelligence. *Intelligence*, *25*, 83–92.
Nisbett, R.E. (2009). *Intelligence and how to get it: Why schools and cultures count*. New York, NY: Norton.
O'Neil, J. (1992). Putting performance assessment to the test. *Educational Leadership*, *49*(8), 14–19.
Parodi, C. T. (2003). *Actions to achieve health equity for ethnic/racial groups regional workshop adoption and implementation of affirmative action policies for afro-descendants in Latin America and the Caribbean.* Governance and Policy, Pan American Health Organization, Regional Office of the World Health Organization.
Pierce, R. L., Adams, C. M., Neumeister, K .L. S., Cassady, J. C., Dixon, F. A., & Cross, T. L. (2006). Development of an identification procedure for a large urban school corporation: Identifying culturally diverse and academically gifted elementary students. *Roeper Review*, *29*(2), 113–118.
Pine, J., Aschbacher, P., Roth, E., Jones, M., McPhee, C., Martin, C., … Foley, B. (2006). Fifth graders' science inquiry abilities: A comparative study of students in hands-on and textbook curricula. *Journal of Research in Science Teaching*, *43*(5), 467–484.
Public Law. (1964). Public Law 88-235 - Civil Rights Act of 1964.
Reis, S.M., & Housand, A. (2007). Characteristics of gifted and talented learners: Similarities and differences across domains. In F.A. Karnes & K.R. Stephens (Eds), *Achieving excellence* (pp. 62–81). New York, NY: Pearson, Merrill/Prentice Hall.
Reis, S. M, & Renzulli, J. S. (2009a). Myth 1: The gifted and talented constitute one single homogeneous group and giftedness is a way of being that stays in the person over time and experiences. *Gifted Child Quarterly*, *53*(4), 233–235.
Reis, S. M. & Renzulli, J. S. (2009b). Developing talents and gifted behaviors in children. In B. MacFarlane & T. Stambaugh (Eds.), *Leading change in gifted education* (pp. 107–118). Waco, TX: Prufrock Press.
Renzulli, J. S. (1978). What makes giftedness? Reexamining a definition. *Phi Delta Kappan, 60*(3), 180–184, 261.
Renzulli, J. S. (1999). What is this thing called giftedness, and how do we develop it? A twenty-five year perspective. *Journal for the Education of the Gifted*, *23*(1), 3–54.
Renzulli, J. S. (2005). Equity, excellence, and economy in a system for identifying students in gifted education programs: A guidebook (RM05208). The National Research Center on the Gifted and Talented, University of Connecticut. Retrieved from http://www.gifted.uconn.edu/nrcgt/renzulli2.html
Report to the President. (1995). Affirmative action: History and rationale. Clinton Administration's Affirmative Action Review. Retrieved from http://clinton2.nara.gov/WH/EOP/OP/html/aa/aa02.html
Rosemarin, S. (2001). The evaluation of a pullout program for gifted children in Israel. *Gifted Education International*, *15*(3), 316–325.
Sadler, P. M., Gerhard, S., Hazari, Z., & Tai, R. (2012). Stability and volatility of STEM career interest in high school: A gender study. *Science Education*, *96*, 411–427.
Sander, R. H. (2004). A systemic analysis of affirmative action in American law schools. *Stanford Law Review*, *57*, 367–483.
Schneeweis, N., & Zweimüller, M. (2012). Girls, girls, girls: Gender composition and female school choice. *Economics of Education Review*, *31*, 482–500.
Seelke, C. R. (2008). *Afro-Latinos in Latin America and considerations for U.S. policy (RL32713).* Washington, DC: Congressional Research Service. Retrieved from http://digitalcommons.ilr.cornell.edu/key_workplace/486/
Seiler, G. (2013). New metaphors about culture: Implications for research in science teacher preparation. *Journal of Research in Science Teaching*, *50*(1), 104–121.
Silverman, L. K. (1986). The IQ controversy – conceptions and misconceptions. *Roeper Review*, *8*, 136–140.

Solórzano, D. G., & Yosso, T. J. (2002). A critical race counter story of race, racism, and affirmative action. *Equity & Excellence in Education*, *35*(2), 155–168.
Stemler, S. E., & Tsai, J. (2008). Best practices in inter-rater reliability: Three common approaches. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 29–49). Los Angeles, CA: Sage.
Sternberg, R. J. (1984). What should intelligence tests test? Implications of a triarchic theory of intelligence for intelligence testing. *Educational Researcher*, *13*, 5–15.
Sternberg, R. J. (2010). Assessment of gifted students for identification purposes: New techniques for a new millennium. *Learning and Individual Differences*, *20*(4), 327–336.
Stobart, G., Elwood, J., & Quinlan, M. (1992). Gender bias in examinations: How equal are the opportunities? *British Educational Research Journal*, *18*(3), 261–276.
Tal, T., & Medijensky, S. (2005). A model of alternative embedded assessment in a pullout enrichment program for the gifted. *Gifted Education International*, *20*, 166–186.
Tannenbaum, A. (1983). *Gifted children: Psychological and educational perspectives*. New York, NY: Macmillan Publishing Co., Inc.
Terman, L. M., & Oden, M. H. (1959) *The gifted group at mid-life*. Stanford, CA: Stanford University Press.
VanTassel-Baska, J. (2005). Domain-specific giftedness applications in school and life. In R. J. Sternberg & J. E. Davidson (Eds.), *Conceptions of giftedness* (2nd ed., pp. 358–376). Cambridge: Cambridge University Press.
Willingham, W. W., & Cole, N. S. (Eds). (1997). *Gender and fair assessment*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
Willingham, W. W., & Cole, N. S. (Eds). (2013). *Gender and fair assessment*. New York, NY: Taylor & Francis Group.
Zohar, A., & Dori, Y. J. (2003). Higher order thinking skills and low achieving students - Are they mutually exclusive? *The Journal of the Learning Sciences*, *12*(2), 145–182.
Zohar, A., & Nemet, F. (2002). Fostering students' knowledge and argumentation skills through dilemmas in human genetics. *Journal of Research in Science Teaching*, *39*, 35–62.
Zorman, R., & David, H. (2000). *There is another way girls and women: Achievements and challenges*. Jerusalem: Achva Publications (in Hebrew).

Appendices

Appendix A. Gifted students' project topics by domain and type

Domain	Project topic	Project type
Mathematics	Fractions	Games
	Arithmetic	Games
Science	Nature – environment	Ecological house model
	Nature – seeds	Written project
	Physics – energy	Amusement park model
	Biology – photosynthesis	Experiment, presentation
	Biology – proteins	Presentation
	Biology – carbohydrates	Presentation

Appendix B. Gifted students' projects and examples

Example 1: Project #62 – Mathematics, 4th grade – multiplication game
The objective of this project was to learn multiplication, understand operations order, and memorize the multiplication table. Following the preparation of the games by the students, the games are presented in an exhibition and played by 3rd graders mentored by the 4th graders. The games are then stored for future use in the school's mathematics room. As Figure A1 shows, the game is made clearly and aesthetically. The focus is on multiplication and order of operations, and the game is dynamic, requiring movement along a path based on rolling dice, so students are engaged and learn joyfully.

Figure A1. Project #62: Mathematics, 4th grade – multiplication game.

Example 2: Project #23 – Environment, 6th grade – ecological house model
The project's objective was to raise students' awareness of ecological issues of reuse and recycling. The teacher suggested several ideas which students developed. In this project, they developed a model of an ecological house in which water is collected from the roof to the center of the house, which also serves to place solar panels.

Figure A2. Project #23: Environment, 6th grade – ecological house model.