Bias factors in mathematics achievement tests among Israeli students from the Former Soviet Union

Michal Levi-Keren

Abstract: This study explains mathematical difficulties of students who immigrated from the Former Soviet Union (FSU) vis-à-vis Israeli students, by identifying the existing bias factors in achievement tests. These factors are irrelevant to the mathematical knowledge being measured, and therefore threaten the test results. The bias factors were identified through a structured process that included analysis of mathematics achievement test results, as well as interviews with immigrant students and “culture experts.” The prominent bias factors among the immigrants were primarily specific test items unfamiliar to their way of thinking as well as insufficient attention to item instructions that appeared in a different typographic presentation. This paper discusses the theoretical and practical implications of these findings.

1. Introduction

Immigration is a global phenomenon and the academic integration of immigrant students currently preoccupies many education systems worldwide. Being an immigration country, Israel has absorbed...
a great number of immigrants from various countries since its foundation. In recent years, the number of immigrants from the Former Soviet Union (FSU) was the highest compared to immigrants from other countries of origin, such as France, United States, Canada, and Ethiopia (Statistical Abstract of Israel, no 65, 2014). The percentage of FSU immigrants out of the total immigrant children living in Israel in 2014 was around 34% (Berman, 2015).

Learning difficulties of immigrant students are mainly manifested in subjects that require proficiency in the language of instruction used in the destination country. Although mathematics involves numbers and symbols, this discipline is still considered language-dependent (Francis, Rivera, Lesaux, Kieffer, & Rivera, 2006) and researchers now recognize it as a “cultural product” (Presmeg, 2007). Research findings regarding mathematics achievements of FSU immigrant students in Israel portray a complex and multidimensional picture. Studies based on self-reports of students or their teachers show that the achievement of FSU immigrant students in mathematics is similar to those of Israeli-born students, and sometimes even higher (Geva-May, 1992; Tatar, Kfir, Sever, Adler, & Regev, 1994). Further, Levin, Shohamy, and Spolsky (2003) found that ninth-grade FSU students scored higher than their Israeli-born counterparts. Similarly, analysis of MEITZAV tests (Measures of School Efficiency and Growth tests) in the years 2007–2010 indicated that the FSU immigrants’ achievements were similar to those of Israeli-born students in Hebrew-speaking schools, but higher than those of Israeli-born children when students were matched by socioeconomic background. A similar picture also emerged following the analysis of PISA (Programme for International Student Assessment) tests from 2009. The achievements of FSU immigrant students of low and medium socioeconomic status (SES) were higher than those of the entire student population in Hebrew-speaking schools (Goldschmidt, Glickman, Rap, Ron-Kaplan, & Regbi, 2012).

In contrast, there is empirical evidence that contradicts the prevalent view that FSU immigrant children are particularly good in mathematics. For example, veteran teachers of mathematics pointed out that the Russian students’ excellence was myth rather than reality, as they were not more talented than local students (Eisikovits, 2012). Levin et al. (2003) found that the achievements of fifth- and eleventh-grade FSU immigrant students were significantly lower than their Israeli-born peers, particularly in problem-solving. Furthermore, and contrary to the expectations, the achievement of FSU immigrant students becomes equal to those of Israeli-born children only after five to seven or nine years in the country, according to the investigated grade levels (5th, 9th or 11th). Levin et al. (2003) also showed that self-assessment of mathematics capabilities of this population was significantly lower than that of Israeli-born children. Another study illustrated that the percentage of FSU immigrant students who reported encountering difficulty in mathematics was similar to that of adolescent groups from other countries of origin like Ethiopia (a country that has a relatively low level of literacy) (Kahan-Strawczynski, Levi, & Konstantinov, 2010). Konstantinov (2015) found that the MEITZAV scores in mathematics among FSU-born students in 2012/2013 were lower than those of the other students from the Jewish sector in the fifth and eighth grades. Furthermore, the cumulative rate of dropout from grades 9 to 12 among FSU students in 2010 until 2013 was around 17% compared to the 10% dropout rate found among the Jewish sector. In addition, the FSU students’ rate of participation in the “Immigrant Youth from Risk to Chance” program was higher in comparison to Ethiopian- and Israeli-born students (Kahan-Strawczynski, Vazan-Sikron, & Levi, 2008).

The literature dealing with worldwide immigration also indicates contradictory and inconsistent findings associated with immigrants’ academic achievements. Some studies show that the achievements of immigrant students in various subjects, including mathematics, are similar to the achievements of native-born children after many years, and are sometimes even higher (e.g. Bankston & Zhou, 2002). However, other studies indicate gaps in academic achievements of immigrant children, which persist for many years and are not bridged with time (Andon, Thompson, & Becker, 2014; Ercikan, Roth, Simon, Sandilands, & Lyons-Thomas, 2014). One of the reasons for the difficulty in generalizing these findings is the fact that differences in achievements were found between immigrants and native-born students as well as within different immigrant groups (Levin & Shohamy, 2008).
Until now, the explanation of the learning gaps between immigrant students and native-born students was grounded in a complex and elaborate setup of characteristics typical of immigrant students. Some of them are demographic (e.g., parents’ education and support) while others are culture-dependent, and are related to the challenge involved in acquiring a language, ways of thinking, values, and different learning habits. Studies exploring these characteristics did not yield consistent results to enable conclusive generalizations (see review in Levin & Shohamy, 2008). Other researchers focused on an additional variable set, which reflects the common assessment approaches for investigating immigrant students’ achievements. They underscored the unequal learning conditions, including the means of assessment themselves. These assessments are sometimes incompatible with the students’ learning needs (Gándara, Rumberger, Maxwell-Jolly, & Callahan, 2003), and do not take into consideration their linguistic, value-oriented, cognitive, and demographic characteristics (Pollitt & Ahmed, 2000), which differ from the majority groups of the native-born students. Moreover, the use of testing accommodations (such as extended time for test taking), developed in order to improve the validity of the students’ assessment, bypassing the foci of their difficulty, has not always been shown to be effective (Abedi, 2008).

The approach adopted in this study for explaining the academic gaps between the FSU immigrant students, as the dominant immigrant group today in the education system, and the Israeli-born students, is different than the one applied in the past. Specifically, this study explores an issue relating to test validity. Its starting point is that immigrant students do not demonstrate optimal performance in tests due to the combined effect of their unique demographic and cultural characteristics (one set of variables), and the prevalent assessment approaches for measuring their academic achievements (as a second set of variables). The assumption is that each of these two sets of variables, as well as their interaction, is not relevant to the construct being measured in the mathematics achievement test, namely mathematical knowledge. Consequently, these variable sets can be viewed as bias factors.

The purpose of the current study is to implement a structured process for investigating bias factors when examining immigrant students’ mathematical achievements, while making the required distinction between two factor types: irrelevant factors (bias factors) and factors relevant to the examined construct. The latter illustrate true performance differences between the compared groups and are referred to as impact factors (Camilli & Shepard, 1994). The theoretical importance of this study resides in the implementation of this distinction for the purpose of deepening and enhancing comprehension of the immigrant students’ difficulties in a wider context. This context takes into consideration characteristics of both their cultural background and of the assignments themselves as reflecting certain aspects of the assessment approach applicable to them. The importance of this study at the applied level resides in establishing an empirical infrastructure for developing testing accommodations that are optimally adapted to the examinees’ specific and unique characteristics.

2. Literature review
The theoretical review presented below will first discuss the concept of “bias factors” and their manifestation in the achievement tests. Following this is a detailed presentation of the sets of variables associated with the immigrant students’ characteristics, as well as the assessment approaches applicable to these students.

2.1. Bias factors and their manifestation in mathematics learning
Bias factors affect examinees’ scores and are irrelevant to the measured construct. These factors distort the meaning of test inferences and therefore are perceived as threatening the assessment validity (Messick, 1989). Bias factors in a test can be associated with several components (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCMR], 1999): the test itself (unsuitable sampling of contents, unclear test instructions and so on), the examinees, and the test situation. Each of these components and their
interaction might lead to an incorrect interpretation of the test results. A biased item is one which examinees from different groups, but with equal ability, have a different chance of getting the item right, due to reasons that are irrelevant to the test’s objective (Camilli & Shepard, 1994).

An investigation of bias factors starts first and foremost, by identifying items characterized by Differential Item Functioning (DIF). As a psychometric feature of the item, DIF refers to an item that functions differentially among students in different groups with identical abilities. That is, it is found to be exceptionally difficult or easy for members of a particular group. Henceforth, these items will be referred to as DIF items. The DIF analysis is perceived as an essential effort in promoting the validity and fairness of the test results for subgroups of examinees (Sireci & Rios, 2013). Nevertheless, as will be clarified later, a DIF characterization is a necessary yet insufficient condition for an item being considered biased (Camilli & Shepard, 1994).

The research literature exploring DIF items in quantitative tests focused on immigrant groups in the United States. These groups took ability and aptitude tests for the purpose of admission to academic studies (e.g. Scheuneman & Grima, 1997); international tests such as PISA (Ercikan et al., 2014), as well as mathematical achievement tests (Martiniello, 2008; Uiterwijk & Vallen, 2005). Potential bias factors identified in these studies were: assignment type (word problems), extent to which the item was concrete, its mode of representation, its linguistic complexity, etc. The few studies conducted in Israel (e.g. Allalouf & Abramzon, 2008) have focused primarily on verbal ability tests rather than mathematics.

In immigrant students’ mathematics learning, the linguistic and cultural sources of bias might be more prominent due to the special nature of this subject. Mathematics is not solely “competence of numerical manipulations” (Bullock, 1994). Rather, mathematical language comprises unique components (e.g. vocabulary which partly exists in the natural language but in another meaning); unique syntax; variety of semantic constructs presented in a situation of an item; and symbolic representations that might be different from those familiar to students in their country of origin (Perkins & Flores, 2002). In addition to the mathematical language, mathematics learning requires linguistic knowledge in the natural language. This is due to the fact that in various types of mathematical learning texts there is frequently a linguistic complexity, manifested by the vocabulary, the verb form (active vs. passive), subordinate and relative clauses, and others (Abedi, 2011). Since this linguistic complexity is not relevant to the investigated mathematical competences, its inclusion in the achievement tests might lead to biased test results. This is true primarily among population groups like immigrant students, who are not proficient in the language in which a test is administered, the mathematical language, or the testing register (e.g. the phrase *None of the above* that is used almost exclusively in multiple-choice tests (Solano-Flores, Barnett-Clarke, & Kachchaf, 2013).

### 2.2. The unique characteristics of immigrant students from the FSU

Research into the linguistic background of FSU immigrant students illustrates that they come with literacy competences in their mother tongue (Kozulin, 1998). Furthermore, they tend to preserve their language of origin and their cultural identity in the family and the community, as well as for cultural consumption (Ben-Rafael, Olshoain, & Geijst, 1998). A recent study (Niznik, 2011) revealed that most of them choose to maintain their cultural identity in its original form or along with the Israeli culture. Regarding proficiency in the Hebrew language, most of the adolescents who immigrated to Israel from the FSU between the years 2000–2002 ranked their proficiency in Hebrew as lower than average. They also reported that they did not always manage to read and write in Hebrew or understand the content of lessons conducted in Hebrew. In fact, most of them (60%) claimed that lack of proficiency in the academic Hebrew language is the reason for their learning difficulties (Niznik, 2008).

Examination of the cognitive and value-oriented characteristics of these students indicates that learning difficulties stem from the gap between the ways of thinking and behavior norms to which they have been accustomed and those prevalent among Israeli students. Specifically, FSU immigrant students’ tend to obey authority and respect teachers; object to the perception of
independence presented by the teachers; demonstrate a lack of initiative; and are accustomed to school discipline (Ben-Peretz, Eilam, & Yankelevitz, 2011; Niznik, 2008). Moreover, some researchers argue that children who grew up and were educated in a non-democratic regime, which dictated decisions about a wide range of life areas, did not have the opportunity to analyze alternatives out of a variety of options, compare between them, and choose among them (Berger, 1997). Israeli teachers also pointed out the difficulty in critical reflection encountered by these students (Tokatly, 1992).

Despite these drawbacks, the gap between the values and learning principles common and emphasized in the country of origin and those acceptable in the new country can be also advantageous to the immigrant students. For example, in the education system of the FSU, a strong emphasis is put on natural sciences and physics at the expense of humanities (Horowitz, 1986). In fact, in Israel, the FSU immigrants tend to study science and technology more than Israeli-born students (Chachashvili-Bolotin, 2011). Sfard and Prusak (2005) investigated how sociocultural aspects seem to affect mathematical learning processes. They found that immigrant students from the FSU perceive knowledge of mathematics as a universal value that is one of the most important ingredients of education. In contrast, native Israelis stressed the fact that matriculating in this subject with high grades would largely increase their chances for being accepted into a university.

Thus far, the literature review has demonstrated that the ability to cope with a culture-dependent subject such as mathematics is linked to linguistic, cognitive, as well as value-oriented factors. The unique characteristics of FSU immigrant students in these contexts, including their strong and weak points, are not appropriately addressed in the testing policy applicable to them. Consequently, the test results do not reliably reflect their knowledge and capabilities. The next part of the literature review will therefore focus on the testing policy implemented in the case of these students, attempting to comprehend its implications with reference to testing their academic achievements.

2.3. Assessment approaches in immigrant-absorbing countries

2.3.1. Learning conditions and means of assessment

The opportunity to learn is one of the most commonly used definitions of fairness, in addition to lack of bias, as outlined in the Standards for Educational and Psychological Testing (AERA, APA, & NCMR, 1999). Unequal learning and testing conditions can partly account for the academic gaps between immigrant students and native speakers (Gándara et al., 2003). For example, Florez (2012) recently reported the doubtful validity of the cut scores of the English proficiency examination used by Arizona’s education authority to determine which immigrant students should get English support services.

The instructional language also impacts the learning conditions of the immigrant students in English-speaking countries. In the context of mathematics, the students have to learn new ways for speaking about mathematics while concurrently studying English as a second language (Brenner, 1998). The special testing register is, as mentioned, another factor which renders their performance more difficult, mainly when the majority culture dictates the content, contextual information, and formulation of the test items (Solano-Flores & Nelson-Barber, 2001).

Regarding the means of assessment, Olshtain (2007) points out potential gaps between the thinking and cultural patterns of those engaged in educational assessment and the examinees’ thinking patterns. She argued that exam developers function according to pedagogical practices prevalent in the Western world. These practices are based on text, formal analysis, low context (namely, transferring messages by verbal methods), a school-oriented approach, and so on. These could render comprehension of the exam items more difficult for examinees whose cultural background is different.

The dominant approach in the Israeli education system with regard to immigrant students also leads to unequal learning conditions and learning opportunities. Niznik (2008) stipulates that in contrast to the 1990s, today, immigrant students do not receive special academic assistance at school.
Moreover, many teachers have not been trained to teach immigrant students and therefore do not address their needs and abilities (Levin et al., 2003). According to the immigrant students themselves (first and second generations), there are significant differences between them and the non-immigrant Jewish students, regarding the feeling that the school staff cares about them and that they have someone to turn to in case they had a problem or difficulty (Kahan-Strawczynski, Amiel, Levi, & Konstantinov, 2013). It also has been illustrated that when head teachers perceive the FSU immigrant population in a more positive light (in terms of education level, motivation to study, etc), the more they relate to them as to other students and the less the policy they design includes unique programs that could enhance their success in learning (Litvak, 2008).

Today’s exam policy in Israel reflects a failure to acknowledge the cultural load, which the immigrants carry in their new country. Israeli teachers point out that FSU immigrant students frequently do not succeed on exams about familiar material. This is due to the fact that the format of these items often differs from those illustrated in class, which more strongly impacts the immigrant students (Tokatly, 1992). In addition to these empirical findings, it is important to note that academic tests in Israel are only given in Hebrew to all the immigrant groups. Moreover, the textbooks on which the knowledge tests are based are also not written in the immigrants’ mother tongue (Niznik, 2008).

In view of the above, it is likely to assume that the learning and assessment materials developed at school are insufficiently adapted to the unique needs of this population. Providing adjusted learning materials during the teaching process can be done by instructional accommodations (Maryland Accommodations Manual, 2012). These accommodations mainly serve for narrowing real learning gaps. Testing accommodations, provided during the assessment process, aim to bridge gaps resulting from irrelevant gaps (gaps that do not relate to knowledge or ability in the subject).

2.3.2. Providing testing accommodations
Providing testing accommodations is one of the main approaches designed to reduce the extent to which bias factors affect exams given to groups with special needs, including immigrants (Tindal, Heath, Hollenbeck, Almond, & Harniss, 1998). In the context of assessing academic achievements, this concept relates to certain changes introduced in the exam materials or in the way it is administered. The objective is to remove the sources of difficulty unrelated to the content being tested (Hollenbeck, Tindal, & Almond, 1998), and offer students an opportunity to demonstrate their ability and real knowledge (Thurlow, Thompson, & Lazarus, 2006). The accommodations thus aim to compensate for the examinees’ difficulty without giving them an advantage over students who do not receive the accommodations. This is a prerequisite for determining the validity of the scores obtained following these accommodations (Abedi, 2008).

Great importance is attributed to testing accommodations in light of the constant increase of immigrant pupils in immigrant-absorbing countries. This is also due to the wish to include immigrant students in large-scale assessments and attain a more complete picture of their learning and educational accountability (Thurlow et al., 2006). The accommodations developed for English Language Learners (henceforth ELL) include: allocating extra time, reading the questions aloud, linguistic simplification, writing the answers directly on the test booklet, and others (Abedi, 2008). In Israel, there is a formal inclination to assist immigrant students in the national achievement tests. In the MEITZAV national tests, accommodations were provided for immigrant students who lived in the country between one and three years (Ministry of Education, 2007). For high school matriculation exams, different testing accommodations were provided to immigrants, including: additional time, reading the test aloud in Hebrew, and using a bilingual dictionary (Ministry of Education, 2014). In the college entrance exam (Psychometric Entrance Test), the examinees can be tested in one of six languages, including a combined English–Hebrew version (National Institute for Testing & Evaluation, 2016).

Many studies of the validity and impact of the various testing accommodations on the achievements of immigrant students have yielded inconclusive results. They have shown that one cannot
find accommodations that are equally effective for all students (Abedi, 2008; Pennock-Roman & Rivera, 2011). Pennock-Roman and Rivera (2011), who perform meta-analyses, as well as other researchers (Solano-Flores & Li, 2013), have reinforced the perception that accommodations should be selected according to the examinees’ individual needs, including proficiency level in their mother tongues and in the second language, instructional language, and the time allocated to students during the test. Due to the criticisms voiced against the accommodations, some researchers advocate a re-investigation of the present research paradigm associated with this topic, which does not benefit the linguistic capabilities of immigrants (Schissel, 2012).

In sum, investigation of the learning conditions afforded to immigrant students and the means of assessment applied for measuring their achievements today illustrates that these students are given tests, which apparently do not truly reflect their real capabilities. Furthermore, the testing accommodations provided in order to facilitate the various difficulties facing these students are usually given at the entire test level to all the students who need them. This is done regardless of the wide variety of linguistic and cultural backgrounds, types of test assignments, and the complex interaction between the students and the test. The need to make accommodation-based decisions at the subject-level is common to both ELLs and students with disabilities. However, less research has been published regarding the processes used to make these decisions for ELLs (Kopriva, 2008).

In Israel, as previously mentioned, some of the results indicate academic gaps between immigrant students and the Israeli-born students, in favor of the latter. Under these conditions, it is reasonable to assume that these results are biased and their validity is doubtful.

This study therefore aims to answer three key questions associated with the identification of bias factors in FSU-born students’ achievements in mathematics tests. These questions reflect a procedure customary in the research literature regarding bias factors in test items (Camilli & Shepard, 1994; Uiterwijk & Vallen, 2005).

1. Which mathematics items have DIF, i.e. which items distinguish between FSU immigrants and Israeli-born students, despite similar ability levels across the groups?
2. What are the potential sources of difficulty that characterize the particular items identified as having DIF disfavoring the FSU immigrant students?
3. What is the nature of the sources of difficulty: Do they refer to intergroup differences that are due to measurement artifacts (bias), or to valid intergroup differences (impact)?

3. Methodology

3.1. Research design
This study was conducted using a mixed methods approach. The first and second research questions were investigated using quantitative analysis. These analyses served as the starting point for the third question, which was examined through qualitative analysis. This paper focuses primarily on the findings of the third question.

3.2. Research tools and data collection method
Investigation of the first research question involved a secondary data analysis of findings obtained from a comprehensive study where FSU immigrant students were given mathematics achievement tests (Levin et al., 2003). The test included both open-ended and multiple-choice items. The items tapped different dimensions of mathematics literacy: (1) arithmetic procedures; (2) numeric reasoning; (3) mathematics communication; and (4) various problem types, including word problems. The four dimensions were assessed within the context of curricular mathematical topics (numbers and four rules of mathematics, geometry, and others). These items were consolidated into two parallel versions: 51 items were included in the first version and 58 items in the second version. Thus, for each mathematical topic and dimension of mathematics literacy, there were a similar number of
items in each version. The Cronbach's alpha reliabilities were .94 and .96 for the first and the second forms, respectively.

These tests permitted identification of the DIF items by means of a technique based on item difficulty referred to as delta plot (Angoff & Ford, 1973). The basic idea behind this technique is the comparison of proportions of correct responses for each item (item difficulty or p-value) within each group separately. The item difficulty levels are transformed into a normal deviate (z score). To avoid any negative values in the distribution, z scores are converted to delta scores with a mean of 13 and standard deviation of 4. A higher value indicates a more difficult item in relation to the other items. Then, the paired delta scores for the two groups are plotted together. Plots of delta values from different groups of the same ability will form an ellipse along a 45-degree line crossing the origin, which represents equal difficulty of the items. A line called the major or principal axis is fitted to the scatterplot. The technique does take into account group ability differences by using this major axis of the ellipse as the point from which distances are computed: If the departure of the data is greater than 1.5 units of z scores, the item is considered to show DIF (Muñiz, Hambleton, & Xing, 2001). In the present article, indices of the presence of DIF include at least one of the following measures: (a) the standardized distance of each item from the major axis of the ellipse (b) the standardized difference between the p-values for the two groups, and (c) the standardized difference between the delta values for the two groups.

Among the many advantages embodied in this technique is its ability to provide useful information for investigating cultural differences (Angoff, 1982). Additionally, it is recommended in situations where the sample size does not enable the use of more statistically sophisticated techniques such as those of Item Response Theory (IRT) or Chi square tests (Camilli & Shepard, 1994). Recent DIF studies using the delta plot support its practical usefulness and interest for practitioners (Facon, Magis, & Courbois, 2012; Michaelides, 2010; Van Herwegen, Farran, & Annaz, 2011).

The second research question was examined by characterization of the DIF test items that are advantageous to Israeli-born students as compared to those advantageous to the FSU immigrant students. The item characterization was done with the assistance of four content experts, two from the field of mathematics and two from Hebrew language. This was done based on the characteristics discussed in the research literature as potential sources of difficulty for ethnicity groups, such as linguistically difficult and complex items and those that are different from those that appear in the students’ textbooks (e.g. Abedi & Lord, 2001; Carlton & Harris, 1992; Eid, 2002; Scheuneman & Grima, 1997; Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2006; Uiterwijk & Vallen, 2005).

The third research question was based on information analysis obtained from individual interviews conducted with a sample of FSU immigrant students. The interviews dealt with the students’ performance on both non-DIF items, as well as DIF items identified in the first research question as advantageous to Israeli-born students. Both the DIF items and non-DIF items were part of a series of items relating to a particular question in the test. The decision to include non-DIF items stemmed from the wish to explore the examinees’ achievements while relating to the textual context where the problematic items appeared. The test booklet presented to the interviewees consisted of 19 items. The interviews were conducted according to the “Think Aloud” technique (Alderson, 2002), which requires that the examinees read the test items aloud as well as express their thoughts aloud. The think-aloud protocols provided additional information to the surface characteristics of the items that were identified by expert reviewers as sources of DIF (Ercikan et al., 2014).

For each of the item sources of difficulty identified in the interviews, a decision was made regarding the difficulty source’s relevance to the construct measured in the item. This was done in order to determine whether it constitutes a bias factor. Compared to what is usually done in studies of this area (Ercikan et al., 2014; Uiterwijk & Vallen, 2005; Wu & Ercikan, 2006), the decision about the bias factors was made at both the individual level as well as the group level. In addition, the decision was based on a high number of potential bias factors rather than on a single potential difficulty source per item.
3.3. The sample and sampling method

The study included two samples, one for the quantitative analysis, and the other for the qualitative one. These are described below:

The quantitative sample: This sample was a representative national sample selected by probability sampling (Levin et al., 2003). It was comprised of 530 Israeli-born students and 379 FSU-born students who immigrated to Israel during the 1990s, most before the age of six (about 67%). All the students were in the 5th grade (age 11), studying in Israeli state schools in the Jewish sector. About 84% of the immigrant students came from the European republics of the FSU. This proportion is in accordance with the high percentage (93%) of immigrants who arrived in Israel during the years 1990–2001 from these western republics (Central Bureau of Statistics, 2006). All the students participated in the achievement test, and for identifying the DIF items, a comparison was performed between the focal group—FSU-born immigrant students—and the reference group—Israeli-born students.

The qualitative sample: In order to conduct the interviews, a purposeful sample was selected. The sample included four 5th grade students studying in elementary schools in the central area of Israel (Tel Aviv district). The students who were interviewed came from the European republics (three from Russia and one from Belorus). Their length of residence in Israel ranged from two to six years. After the test objective was explained to them, the teachers selected those students with expressive language capabilities. In addition to the individual interviews conducted with the FSU immigrant students, in-depth interviews were conducted with a math teacher who served as a “culture expert.” The teacher immigrated to Israel from the FSU 15 years before the interviews. She also worked as a regional tutor in the field of mathematics and studies toward an MA degree at the Tel Aviv University. The interviews with the teacher served to complement the data obtained from the interviews of the students and to comprehend their world within their cultural context as is customary in qualitative research.

3.4. Data analysis

Each student was interviewed individually in a separate room within the school. The interviews lasted from 45 min to two hours based on the student’s individual pace. The interviews were audio-recorded and transcribed by the researcher. They were then analyzed by Atals.ti software (Muhr, 2004). The characteristics of the examinees and of the test, identified out of the interview protocols, were defined as “super-codes” in the software. An inter-rater reliability test was conducted for checking the accuracy of the characteristics’ categorization. The percent of agreement (79%) between the researcher and another researcher, an expert in the field of mathematics teaching, was acceptable (Graham, Milanowski, & Miller, 2012).

To determine the nature of the item as encompassing bias or impact factors at the individual level, a mapping sentence was used. A mapping sentence is a technique based on the facet theory conceived by Guttman (Guttman & Greenbaum, 1998). It consists of three components (facets): the population of respondents being researched, the content according to which the items are classified, and the range (R), contains the possible performance of the respondents for each of the items: a correct or incorrect answer.

All of the potential bias factors identified in the study were mapped as elements, mutually exclusive and exhaustive, in the content facets (see Appendix A). These elements were based on the integration of multiple sources of information: a large-scale psychometric analysis of DIF; the literature review; analysis of the item characteristics; think-aloud responses to these items, and the culture expert’s judgments. The combination of one element from each facet, which constitutes the performance profile of the examinee in the item, is called structuple (Guttman & Greenbaum, 1998).

Creating the theoretical framework of the mapping sentence and using it for diagnosing the bias factors of the item is designed to provide a practical tool that has not been suggested until now. It facilitates conceptualization of the bias factors in a more accurate and focused way: the mapping sentence relates to the item characteristics, which may include bias factors, rather than to the entire...
item as biased or not. Moreover, the sentence allows identification of the bias factors at the examinee’s level since the performance of different examinees could be characterized by a unique combination of potential sources of difficulty.

4. Findings
The findings are presented according to the three research questions that guided the study. Throughout the third research question, an analysis of bias factors at two levels—the individual and the group—is presented.

4.1. Items displaying DIF between FSU immigrants and Israeli-born students
The first research question explored whether items on a mathematics test would show DIF between FSU immigrants and Israeli-born students. Out of the 109 test items (in both versions), 19 DIF items (17.4%) were identified. Among all the test items, the number of items favoring (namely, embodying a relative advantage for) the immigrant group was quite similar to the number of items favoring the Israeli-born students: 8 items (7.3%) vs. 11 items (10.1%), respectively. As can be seen in Table 1, of those items favoring the Israeli-born students, a high percentage (about 64%) engage the student in various types of problem-solving.

The incidence of balance between the items that were relatively easy (whereby a relative advantage was shown) for the focal group examinees and the items that were relatively easy for the reference group examinees is a statistical issue, common in analyses of DIF (Linn, 1994).

4.2. Potential sources of difficulty
The second research question focused on determining potential sources of difficulty in the identified DIF items. Two content experts in mathematics and two experts in Hebrew language characterized all the items that were flagged as DIF in the first research question. This was done with reference to the following categories: linguistic characteristics (e.g. item length); cognitive (e.g. redundancy of data in the item); sociocultural (e.g. the fact that the item is associated with the “real world”); and structural (e.g. mathematical literacy dimension measured by the item). These categories emerged from the research literature as leading to difficulties among examinees of minority groups, as well as those actually found in items of the present test.

Among the 20 characteristics found in the research literature as difficult to minority groups, seven (35%) were found—as expected—significantly more frequently in DIF items favoring Israeli-born students, compared to items that favor immigrant students. These characteristics are:

Linguistic:

(1) Longer items, with a relatively large number of words (according to the mean items length)
(2) A higher frequency of Modals (e.g. can, could, have to, must, might, should)
(3) A higher frequency of everyday language, as the dominant language in the item.

| Dimensions of mathematics literacy | Items favoring the Israeli born students (%) | Items favoring the FSU immigrant students (%) |
|-----------------------------------|---------------------------------------------|---------------------------------------------|
| Arithmetic procedures             | 1 (9.1)                                    | 2 (25.0)                                    |
| Numeric reasoning                 |                                             | 1 (12.5)                                    |
| Mathematics communication         | 3 (27.3)                                   | 1 (12.5)                                    |
| Problem solving, including word problems | 7 (63.6)                                | 4 (50.0)                                    |
| Total number of items             | 11 (100.0)                                 | 8 (100.0)                                   |
Cognitive:

(4) A high frequency of data redundancy.

Sociocultural:

(5) A higher frequency of items relating to gender groups.
(6) A higher frequency of concrete items, relating to the “real world”.

Structural:

(7) A higher frequency of verbal problems, as one of the dimensions of math literacy.

Thus, for example, in the cognitive category, a significantly higher percentage of DIF items favoring the Israeli-born students (80%) showed data redundancy ($p < .05$) compared to those items that favored the FSU immigrant students (17%). However, it may be that the above-mentioned characteristics are relevant to the measured construct (mathematical knowledge), and additional analysis is therefore necessary to determine whether these characteristics constitute bias factors.

4.3. Distinction between the sources of difficulty

The third question was aimed at determining which sources of difficulty lead to bias and which reflect real performance differences. The bias factors in the items were identified through an analysis of the answers given by a sample of interviewed students and the words of the culture expert, whose explanations shed light on and clarified the procedure that the interviewees underwent. The sources of difficulty in these items were analyzed according to the stages involved in solving mathematical problems (Polya, 2009): reading the question and understanding it, planning the solution, finding the technical solution, and examining the solution. For each of these stages, the characteristics of the answers were analyzed and then classified into five categories: linguistic characteristics, cognitive characteristics, test-taking skills, sociocultural characteristics, and assignment characteristics (structural characteristics and characteristics related to the way the item was presented). These characteristics categories were developed while attempting to achieve a maximum match between them and the characteristics list documented in the literature review and validated by the content experts. Nevertheless, they are somewhat different than the categories specified in the previous section (4.2), since they are based on data obtained from another source (interviews) and the compared items were different. Thus, the different categorizations demonstrate the distinction between etic categories (categories of characteristics that were documented in the literature review) and emic categories (categories that emerged during the empirical work and reflect the view of the participants (Guba & Lincoln, 1994).

The findings presented below illustrate the way by which one can decide about the nature of the item difficulty source at two analysis levels: the individual level and the group level.

4.3.1. Bias factors at the individual level

The distinction between a characteristic which causes bias and a characteristic which reflects real performance differences was grounded in two criteria: (1) creating a difficulty: based on the analysis of the student’s answer, the item characteristics were classified as “creating a difficulty” or as “not creating a difficulty” during the performance, and (2) relevance: the item characteristic was classified a priori as “relevant” or “irrelevant” to the construct that the test is designed to measure. Thus, this classification was determined regardless of the examinee’s performance. For example, mastering the multiplication table was perceived as a characteristic that is relevant to the examined feature of the item. Conversely, all the linguistic difficulties (such as lack of morphological accuracy during reading), and defective testing competences (such as lack of matching between the answer given orally and the one given in writing) were perceived as irrelevant characteristics.
The list of characteristics that did and did not create a difficulty, according to their degree of relevance to the measured construct in DIF and non-DIF items, was organized in a concise way by means of a mapping sentence (see Appendix A). This framework, which was consolidated following the students’ interviews, facilitated diagnosis of the specific difficulties that characterize the performance of a particular examinee on a specific item.

The mapping sentence could be used for each one of the four interviewers, for each one of the 19 items that were investigated. However, due to the scope of the article, only one sentence will be illustrated below for one item for one of the interviewees. This student is Russian-born and immigrated to Israel at the age of eight. First, the conversation conducted with him regarding his performance in the item included in Figure 1 will be presented, and then the structuple (the profile) of his answers will be displayed.

The student reads the question narrative. He does not read the sentence of the question. He finds it hard to read the word “equal” (reads like “quail”), reads “received” in the feminine form instead of the masculine form.

Examiner: Let’s look at the questions. What is being asked in the problem?
Student: Reads the first sentence again. What does “in which” mean?
Examiner: In the bag.
Student: Continues reading the question. Has a problem with “her children”. Reads “was left” instead of “were left”. You have to multiply 3 by 39.
Examiner: But first try answering the questions. First they ask you “What is being asked in the problem?” Let’s read the options.
Student: How many candies did mother buy: 39. How many children did mother have: […]. How many candies did mother give to each child. How many candies were left in the mother’s hand. It says 3 candies.
Examiner: Yes, but what is being asked in the problem?
Student: How many candies did mother buy.
Examiner: OK.
Student: How many children did mother have.
Examiner: Yes, but you have to indicate the answer to what you are being asked. You don’t have to address each of the statements. You have to write what is being asked in the problem.
Student: Of what, what they taught me?
Examiner: No, in this question. What actually is being asked?
Student: To how many children did she give candies.
Examiner: OK. Do you have here such an option? Let’s see.
Student: 39 ... how many times 3 is 39.
Examiner: But try now to answer the question. You have to indicate one answer. You have here four alternatives. They ask: What is being asked? What is the answer you are expected to give?
Student: Says and indicates: How many children did mother have (alternative b).
Examiner: OK. What you understand from the question is that you are asked how many children does the mother have?
Student: Ask ... No. Ask: How many candies did mother give to each child. (changes the answer to alternative c).
The student’s structuple: The student was tested on item Y out of a cluster of items in which DIF was identified as favoring Israeli-born students (a1). At the text level, the student skipped the instructions written in brackets (“indicate one answer”) (d3). From a cognitive point of view, he did not demonstrate mastery of procedural knowledge since he applied incorrect operations on the data (f3). Since the item is not similar to a school item, his reference to the item was irrelevant (g3.1). His irrelevant reference was manifested in two ways: responding to each of the answer alternatives and providing the mathematical solution instead of replying: “What is being asked in the problem?”. Another characteristic of the examinee’s performance referred to test-taking skills: the student did not read the item in full (h1). The item was defined as a word problem (j1) written in a close-ended format (k2).

The student’s performance in this item had four characteristics. Three of them created a difficulty that was irrelevant to the measured mathematical ability: skipping the item instructions written in brackets; irrelevant reference to the item due to the fact that the item was not like a school item; and applying a defective test-taking skill. The fourth, and the single characteristic that entailed a relevant difficulty, was lack of mastery of procedural knowledge. Hence, the item is defined as being of mixed nature with a prominent inclination to generalization of bias factors.

4.3.2. Bias factors at the group level
Identifying the bias factors at the group level was done while differentiating between DIF items found as favoring the Israeli-born students and items that were not found with DIF, but that were included in the series with the DIF item. Using the Atals.ti software, 58 events (number of times that the difficulty appears in the student’s answer) were identified, 23 of them were detected in the DIF items and 35 in the non-DIF items. These events related to the item characteristics, some of which were relevant to the measured construct and some that were not. This distribution is presented in Figure 2 below.
Figure 2 shows the gap between the percentage of relevant and irrelevant characteristics among FSU immigrant students. This gap is much higher in favor of the latter with regard to the DIF items (a difference of about 48%) compared to the non-DIF items (a difference of about 26%). The obtained picture thus implies a tendency of bias in the items. The reason is that in the DIF items found as problematic to the FSU students, there was a higher frequency of characteristics that constitute an irrelevant source of difficulty. On the other hand, the frequency of the relevant characteristics was higher in the non-DIF items.

As expected, all the relevant sources of difficulty regarding the measured construct are essentially cognitive. Thus, for example, this category includes applying incorrect operations on the data, lack of mastery of the multiplication table, failure to understand the essence of symbols, etc. Among the sources of difficulty irrelevant to the construct, the most common category is Cognitive characteristics (47% of the total irrelevant characteristics). These difficulties were manifested by irrelevant reference to certain test items, misinterpretation of items, or difficulty comprehending them due to their lack of similarity to typical school items. These difficulties were demonstrated by items that had more than one answer, or in items with a different formulation, like the item presented in Figure 1. As demonstrated above, students gave the mathematical solution (indicate how many candies each child received) or responded to each of the answer choices. Other female students from the interviewee group reacted as follows: We do not deal with this type of questions so I am not so used to it; I am not used to this. I have never had this [type of question]. The culture expert confirmed that the students were indeed unaccustomed to this type of item. In her opinion, the word “problem” should be replaced by “question” because the word “problem” already causes a problem.

The second most frequent category of the irrelevant sources of difficulty is linguistic characteristics (35%). These were mainly manifested by skipping the item instructions while reading it (e.g. “indicate one answer”), which probably stemmed from the lack of sufficient reference to the typographical aspects of the text. That is, a text in brackets, written in another font or in a separate line (see example in the student’s structuple presented above). In another item, an additional student skipped the instructions written in a font that was not highlighted (you can write your answer as an arithmetic exercise, a drawing, signs or any ways that you deem right). The culture expert argued that the students do not relate to what is written in another style of writing. They don’t need this given in order to answer. Why should they read it?

In summary, the findings show that in the DIF items favoring Israeli-born students, a number of bias factors were identified among the FSU immigrant students. These factors were described at the group level, and by means of the mapping sentence, also at the individual level.

5. Discussion and summary
The main aim of this study was to identify the bias factors connected with the achievements of FSU immigrant students on a mathematics test. To accomplish this, the research procedure consisted of the following phases, each investigated within the framework of a specific research question: identifying items displaying DIF between FSU immigrant and Israeli-born students; detecting the potential sources of difficulty that characterize the DIF test items identified as favoring the Israeli-born students; and finally, investigating which of the sources of difficulty reflect bias and which reflect real performance differences.

The discussion chapter presents the main findings in accordance to the research questions as well as the conclusions drawn from the findings of the third research question which relates to the heart and core of this study: identifying bias factor in the test items. Implications, limitations, and recommendations for further studies are then discussed.

5.1. Summary of the main findings
Findings of the first research question focusing on the detection phase of the DIF indicated that about 1/5 of all the mathematics test items functioned in a differential way among Israeli-born
students compared to FSU immigrant students. Most of these items demonstrated DIF against FSU students (11 out of 19 items). Among these DIF items, a high percentage (about 64%) require solving problems of various types. This finding corresponds to those reported in previous studies, which showed that this specific domain is difficult among FSU students (Levin et al., 2003).

Within the framework of the second research question, the explanatory phase, the DIF test items identified in the first research question were characterized by a set of categories that related to linguistic, cognitive, sociocultural, and structural aspects. About 1/3 of the characteristics reported in the research literature as creating a difficulty in minority groups were significantly more frequent in the DIF items favoring the Israeli-born students. Nevertheless, since some of the characteristics are relevant to the measured construct (mathematical knowledge) and some are not, this is not sufficient to define the characteristics as bias factors.

Within the framework of the third research question, the student interviews enabled additional sources of difficulty to be added to those identified a priori based on the literature review. All the potential sources of difficulty were organized into the following categories: linguistic, cognitive, sociocultural, test-taking skills, and assignment characteristics. These categories were then presented within the framework of a mapping sentence, where every content facet referred to another category of item characteristics. This sentence enabled the setting of a different structuple for every examinee regarding the bias factors in a particular item.

The findings at the individual level support the notion that each examinee can have a specific structuple according to his/her unique areas of difficulty. The findings at the group level indicate that the frequency of bias factors is higher in DIF items that favor the Israeli-born students than in non-DIF items. “Cognitive characteristics” and “linguistic characteristics” were the most prominent among the categories of bias factors in the DIF items.

5.2. Conclusions
A considerable number of bias factors were found to be largely cognitive in nature. In particular, the findings show that the interviewees found it difficult to give the anticipated answers when the item demands were unfamiliar to them. For example: indicating “what is being asked in the problem” or responding to an item which has more than one answer. As a result, their answer patterns were characterized by a reference that was irrelevant to the item, misinterpretation of the item, or difficulty understanding it. Unfamiliar test formats in particular place new immigrants at a disadvantage (Suarez-Orozco & Suarez-Orozco, 2015). Findings showing that students encounter difficulties when coping with items that are not similar to those illustrated in class or in textbooks were found among FSU immigrant students in Israel (Tokatly, 1992), minority groups in the United States (Carlton & Harris, 1992), and immigrants in the U.S. who studied and were tested in English as a second language (Starks-Martin, 1996).

The impact of cultural factors on immigrant students’ methods of coping with items that require a high level of thinking and which allow multiple correct answers was demonstrated in a study showing that students from minority groups, more than the rest of the students, tended to concur with the statements: “there is only one way for solving a mathematical problem,” and “mathematics learning means mostly memorizing facts” (Lubienski, 2001). In Israel too, findings indicate that FSU immigrant students usually interpret mathematical language according to the meticulous syntax of the “formal” mathematical discourse that is included in textbooks (Sfard & Prusak, 2005).

The interviewees in this study were born in the Soviet Union and some of them even studied in the Soviet education system. The pedagogical approach to mathematics teaching prevalent in these students’ countries of origin was mainly based on practice and on the solution of many exercises aiming to reinforce their technical competence as learners (Amit, 2010). Moreover, the approach of the FSU immigrant mathematics teacher in Israel is based on practice and memorizing (Gilad & Millet, 2014). At the pragmatic level, immigrant teachers have an ingrained belief that the solution
of a problem is correct only if it arrives at a correct answer. In contrast, the Israeli system holds the constructivist view that problem-solving is a process, and that a student should be given credit for a partial answer, even if the final result is missing or incorrect (Amit, 2010). These differences regarding the pedagogical approaches may account for the difficulties the FSU students experienced when coping with questions necessitating a rather “different” thinking.

Furthermore, the findings highlighted that the interviewees encountered particular difficulty with linguistic bias factors. This manifested itself in the students’ skipping the item instructions whose comprehension was perceived as essential for solving mathematics word problems (Scheuneman & Grima, 1997). Perhaps these linguistic difficulties, connected to the process of building the meaning of the text, stemmed from the examinees’ insufficient attention to the typographical aspects of the text. This finding is in line with the argument that the text layout of the page is important, particularly for beginning readers (Alderson, 2002), and with the finding that the layout of the text in math word problems doesn’t facilitate the delimitation of the syntactic boundaries for ELLs who do not master the test language (Martiniello, 2008).

This finding might also be related to the test language, which has a unique register of its own. This register could increase comprehension difficulties of students whose mother tongue is different from the test language (Solano-Flores et al., 2013). The FSU immigrant students’ partial proficiency in academic language can be accounted for by the fact that at the time of the test, these students had been in the country no more than six years. Various researchers maintain that incomplete proficiency in academic language leads to difficulties in performing learning activities such as understanding textbooks, writing tests, doing homework, and expressing oneself in class (Levin et al., 2003; Niznik, 2008). This explanation is also supported by other findings that illustrate that academic achievements of immigrants around the world (Hao & Bonstead-Bruns, 1998), as well as of FSU immigrant students in Israel (Levin et al., 2003), are higher the longer they stay in the new country.

The cognitive and linguistic bias factors described above were found mainly in items defined as exploring various types of problem-solving. Thus, these findings illustrate that solving mathematical problems is based on an interaction between the text comprehension (linguistic knowledge), the situation comprehension (knowledge of the world), and mathematical comprehension (mathematical knowledge) (Staub, 1995). It is likely that the linguistic knowledge and world knowledge that the immigrant students were required to demonstrate differed from the knowledge known to them from their countries of origin. Hence, it induced the interference of bias factors, which prevented them from expressing the necessary mathematical knowledge in their answers. In a broader view, one can claim that the identified bias factors support the view that mathematical thinking is an activity that occurs in a sociocultural context (Scribner, 1997).

Based on the above, it appears that while developing assessment tools and testing practices, one has to relate to the issue of cultural validity as an expression of test validity. Cultural validity is attained when proper attention is paid to sociocultural elements, such as values, experiences, and teaching and learning styles. These elements shape the students’ thinking and the way in which they turn the items into something logical for them, and proceed to answer them (Solano-Flores & Nelson-Barber, 2001).

5.3. Implications
As mentioned, the effectiveness of testing accommodations is disappointing. Therefore, detailed information about sources of DIF given to test planners might serve as a basis for the development of testing accommodations tailored to the individual student. Testing accommodations can be developed in a further study, based on findings of this study and of other similar studies. These accommodations, which will engage first and foremost in identifying the bias factors in the test items for the same examinee group to which the accommodations are intended, will be most relevant to the immigrant students. For example, it seems that when a test for FSU immigrant students is developed, one should underline or highlight the item instructions included in the test. Emphasizing the
verbs in the item instructions by underlining them is a recommended testing accommodation (Gibson, Haeberli, Glover, & Witter, 2003). Another specific, direct linguistic support accommodation is “read aloud test items” in the examinee’s native language or in English. This accommodation is very common in class evaluation in mathematics among teachers who teach ELLs (Wolf, Kao, Rivera, & Chang, 2012) and can prevent students from unintentionally ignoring relevant text that is needed to provide the answer. Thus, the test developers should think how to neutralize the potential “mines” embodied in the test, without changing the instructional goals they expect to achieve.

In addition to the option of providing accommodations at the item level, they can be given at the examinee’s level through a mapping sentence like the one implemented in this study. In order to enhance the effectiveness of the testing accommodations, they should be determined in accordance with instruction accommodations practiced and applied in class (Thurlow et al., 2006). For example, based on the results of this study, it seems that it is necessary to foster the FSU immigrant students’ flexibility in coping with mathematical issues, including development of their ability to consider and understand non-standard solutions. This recommendation is grounded in the argument that numeric reasoning is one of the skills that should be valued in school mathematics education (National Council of Teachers of Mathematics, [NCTM], 2000).

Abedi’s (2014) recent study demonstrated how the use of computer technology could assist in the proper implementation of accommodations. A computer-based system can allow accommodation tools to be accessed flexibly to meet the specific needs of each individual student. All the relevant student background data can be entered into the computer for accommodation decision-making, so it can provide one set of accommodations for one student and a different set for another. The growing presence of computers in schools makes the tailored computer-based test delivery system a realistic option for many schools and their teachers (Russell, 2011).

5.4. Limitations and recommendations for further research

A number of methodological issues may potentially limit the application of this study’s results. Firstly, although the method used to identify the items with a differential functioning had certain advantages; it also has a number of disadvantages. The most prominent of these is that it may miss items that function differently across groups if the discriminating power of the item is low, and on the other hand to flag items for DIF, if the discriminating power of the item is high (Dorans & Holland, 1993). Secondly, the database that served for examining the third research question consisted of interviews of a limited number of immigrant students. These interviews exposed the interviewees’ thinking processes and ways of coping with the test items, and thus explained the quantitative findings that emerged when exploring the first research question. Nonetheless, we cannot ignore one of the common critiques of qualitative findings, namely that it is difficult to apply the findings to other groups and environments that differ from those of the current study (Firestone, 1993).

A third methodological issue relates to characteristics of the interviewee sample, which comprised immigrant students from only one group of origin. The fact that no interviews were conducted with another student group with a similar level of achievements and who are Israeli-born prevents us from determining to what extent the difficulties identified among the immigrants are unique to them and are not also common to the Israeli-born students. In light of this, it is advisable to investigate the differential functioning in larger samples while using more than one statistical technique (Sireci & Rios, 2013), as well as conducting individual interviews with Israeli-born students. Moreover, it would be beneficial to expand the investigation to immigrant groups from other countries of origin such as America, Europe, and South America, as well as other ethnic minority groups (e.g. Israeli-Arabs and children of foreign workers).

The difficulty and complexity of explaining DIF that was emphasized in the Standards document (AERA, APA, & NCMR, 1999), was found in this study, and is in line with other studies (Camilli & Shepard, 1994; Uiterwijk & Vallen, 2005). Extending and deepening the empirical effort for improving the ability to identify the sources is an important line of inquiry, as is finding additional methods that
were not applied in this study. For example, we can include a component that is likely to lead to a differential functioning into one of the test items, and compare the performance on this item with items comprising a comparable component that is not expected to cause a differential functioning (Uiterwijk & Vallen, 2005).

To conclude, this study highlights the distinction between DIF and bias. While DIF is a neutral concept defined by statistical means, “bias” is loaded with ethical, social, and political meanings, which might have adverse effects on the immigrant students. The power of the tests and of their ability to shape an individual’s life in many and varied areas is acknowledged (Shohamy, 2001). Hence, it is important to emphasize the need for another assessment perception based on systematic information gathered about the cultural variety of the immigrant students. This perception takes into consideration numerous factors—linguistic, cognitive, social, and environmental—which differentially impact the students’ conduct during a test (Kauffman, Conroy, Gardner, & Oswald, 2008). This assessment perception, which displays multicultural sensitivity, should be manifested using testing practices that transmit a message of respect, caring, and real interest in the students and their promotion. We believe that the structured and systematic technique, presented in this paper in order to explore bias factors and mediate their effect, could assist in setting up a new empirical framework. This framework is essential for developing a fair and equitable assessment culture appropriate to multicultural societies.

### Abbreviations

- **FSU**: Former Soviet Union
- **DIF**: Differential Item Functioning
- **ELL**: English Language Learners

### Acknowledgments

The author would like to thank Prof. Elana Shohamy and Prof. Zipi Libman for their invaluable comments on earlier drafts.

### Funding

The author received no direct funding for this research.

### Author details

Michal Levi-Keren  
E-mail: kerenm@post.tau.ac.il  
ORCID ID: http://orcid.org/0000-0001-6323-7214

1 Kibbutzim College of Education, Technology and the Arts, Tel Aviv 62507, Israel.

2 School of Education, Tel Aviv University, Tel Aviv 69978, Israel.

### Citation information

Cite this article as: Bias factors in mathematics achievement tests among Israeli students from the Former Soviet Union, Michal Levi-Keren, Cogent Education (2016), 3: 1171423.

### References

Abedi, J. (2008). Utilizing accommodations in assessment. In E. Shohamy, & N. Hornberger (Eds.), The encyclopedia of language and education (2nd ed., Vol. 77 (pp. 331–348). New York, NY: Springer Science and Business Media LLC.

Abedi, J. (2011). Language issues in the design of accessible items. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), Handbook of accessible achievement tests for all students (pp. 217–230). New York, NY: Springer.

Abedi, J. (2014). The use of computer technology in designing appropriate test accommodations for English Language Learners. Applied Measurement in Education, 27, 261–272. doi:10.1080/08957347.2014.964310

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. Applied Measurement in Education, 14, 219–234.

Alderson, J. C. (2002). Assessing reading. UK: Cambridge University Press.

Allalouf, A., & Abramzon, A. (2008). Constructing better second language assessments based on Differential Item Functioning analysis. Language Assessment Quarterly, 5, 120–141. doi:10.1080/15434300801934710

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Amit, M. (2010). Cultural conflicts in mathematics classrooms and resolution–The case of immigrants from the Former Soviet Union and Israeli “Old timers”. Montana Mathematics Enthusiast, 7, 263–274. Retrieved from http://scholarworks.umt.edu/cgi/viewcontent.cgi?article=1187&context=tme

Andon, A., Thompson, C. G., & Becker, B. J. (2014). A quantitative synthesis of the immigrant achievement gap across OECD countries. Large-scale Assessments in Education, 2, 1–20. doi:10.1186/s40536-014-0007-2

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 96–116). Baltimore, MD: The Johns Hopkins University Press.

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10, 95–105.

Bankston, C. L. III., & Zhou, M. (2002). Being well vs. doing well: Self-esteem and school performance among
immigrant and nonimmigrant racial and ethnic groups. The International Migration Review, 36, 389–415. Retrieved from search.proquest.com/docview/215274177?accountid=14765

Ben-Perez, M., Elam, B., & Yankelevitz, Y. (2011). Classroom management in multicultural classes in an immigrant country: The case of Israel. In C. M. Evertson & C. S. Weinstein (Eds.), Handbook of classroom management: Research, practice, and contemporary issues (pp. 1121–1140). Mahwah, NJ: Lawrence Erlbaum.

Ben-Rafael, A., Olshaim, O., & Gejst, I. (1998). Identity and language: The social insertion of Soviet Jews in Israel. In E. Leshem & J. T. Shuval (Eds.), Immigration to Israel: Sociological perspectives. Studies of Israel Society, Vol. VIII (pp. 333–356). New Brunswick, NJ: Transaction.

Berger, R. (1997). Adolescent immigrants in search of identity: Clingers, eroders, oscillators and integrators. Child and Adolescent Social Work Journal, 14, 263–275.

Berman, T. (2015). Immigrant children in Israel 2015. Jerusalem: The Israel National Council for the Child, Center for Research and Development (in Hebrew).

Brenner, M. E. (1998). Development of mathematical communication in problem solving groups by language minority students. Bilingual Research Journal, 22, 149–174. doi:10.1080/15235882.1998.10162720

Bullock, J. O. (1994). Literacy in the language of mathematics. The American Mathematical Monthly, 101, 735–743. Retrieved from http://www.jstor.org/stable/2974528

Carnilli, G., & Shepard, L. A. (1994). Methods for identifying biased test items. Vol. 4. Thousand Oaks, CA: Sage Publications.

Carlton, S. T., & Harris, A. M. (1992). DIF detection and aptitude test: Gender and majority/minority group comparisons (Research Report No. RR-92-64). Princeton, NJ: Educational Testing Service.

Central Bureau of Statistics. (2006). Immigrants, by period of immigration, country of birth and last country of residence. Retrieved from http://www.cbs.gov.il/reader/shtnaton/templ_shtnton_e.html?num_tab=&stks=06&Create=2006

Chachashvili-Bolotin, S. (2011). Educational achievements and study patterns of immigrants from the Former Soviet Union in Israeli secondary schools. Canadian Issues Winter/Hiver 2010, 97–103. Retrieved from http://search.proquest.com/docview/1009073980?accountid=14765

Doran, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), Differential Item Functioning (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum.

Eid, G. (2002). Gender, Ethnicity, and language influences on Differential Item Functioning in the SAT (A dissertation presented to the faculty of the College of Education). Ohio University, Athens, OH.

Eisikovits, R. A. (2012). Coping with high-achieving transnationalist immigrant students: The experience of Israeli teachers. Milgash: Journal of Social-Educational Studies, 35, 113–130 (in Hebrew).

Erickson, K., Roth, W., Simon, M., Sandlinds, D., & Lyons-Thomas, J. (2014). Inconsistencies in DIF detection for sub-groups in heterogeneous language groups. Applied Measurement in Education, 27, 273–285.

Facon, B., Magis, D., & Courbois, Y. (2012). On the difficulty of relational concepts among participants with Down syndrome. Research in Developmental Disabilities, 33, 60–68.

Firestone, W. A. (1993). Alternative arguments for generalizing from data as applied to qualitative research. Educational Researcher, 22, 16–23. doi:10.3102/0013189X022002016

Flores, J. D. (2012). Examining the validity of the Arizona English Language Learners assessment cut scores. Language Policy, 11, 33–45. doi:10.1007/s10993-011-9225-4

Francis, D., Rivera, M., Lesaux, N., Kieffer, M., & Rivera, H. (2006). Practical guidelines for the education of English language learners: Research–based recommendations for the use of accommodations in large-scale assessment. Portsmouth, NH: RMC Research Corporation, Center on Instruction. Retrieved from http://www.centeroninstruction.org/files/Ell3-Assessments.pdf

Gandara, P., Rumberger, R., Maxwell-Jolly, J., & Callahan, R. (2003). English learners in California schools: Unequal resources, unequal outcomes. Education Policy Analysis Archives, 11. Retrieved from http://www.usc.edu/dept/education/CMMR/FullText/ELLS_in_California_Schools.pdf

Geva-May, I. (1992). Organisational and pedagogical aspects in the absorption of immigrant students at school. Jerusalem: Ministry of Education and Culture, Pedagogical Secretariat, Unit of Assessment and Measurement (in Hebrew).

Gibson, D., Haeberli, F. B., Glover, T. A., & Witter, E. A. (2003, April). The use of recommended and provided testing accommodations. Paper presented at the annual convention of the National Association of School Psychologists, Toronto.

Gilad, E., & Milliet, S. (2014). A Multicultural view of mathematics teaching: Teachers at Israeli primary schools. International Journal of Learning, Teaching and Educational Research, 7, 23–43.

Goldschmidt, N., Glickman, H., Rap, Y., Ron-Kaplan, O., & Regbi, A. (2012). Achievements of FUS immigrant students in the Israeli education system. Jerusalem: Ministry of Education, National Authority for Measurement and Assessment in Education (in Hebrew).

Graham, M., Milanowski, A., & Miller, J. (2012). Measuring and promoting inter-rater agreement of teacher and principal performance ratings. Center for Educator Compensation Reform. Retrieved from http://files.eric.ed.gov/fulltext/ED532068.pdf

Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), Handbook of qualitative research (pp. 105–117). Thousand Oaks, CA: Sage.

Guttmann, R., & Greenbaum, C. W. (1998). Facet theory: Its development and current status. European Psychologist, 3, 13–36. doi:10.1027//1016-9040.3.1.13

Hao, L., & Bonstead-Bruns, M. (1998). Parent-child differences in educational expectations and the academic achievement of immigrant and native students. Sociology of Education, 71, 175–198. Retrieved from http://www.proquest.com/docview/216479735?accountid=14765

Hollenbeck, K., Tindal, G., & Almond, P. (1998). Teachers’ knowledge of accommodations as a validity issue in high–stakes testing. The Journal of Special Education, 32, 175–183. Retrieved from http://search.proquest.com/docview/194700849?accountid=14765

Horowitz, T. R. (1986). The two worlds of children in Israel and the USSR (The Soviet children in a state of dissonance). In T. R. Horowitz (Ed.), Between two worlds–children from the Soviet Union in Israel (pp. 221–232). Lonham, MD: University Press of America.

Kahan-Straucznyczy, P., Vazan-Siron, L., & Levi, D. (2008). From risk to opportunity–A project for immigrant youth: Findings of an evaluation study (Research Report No. 515-08). Jerusalem: Myers-JDC-Brookdale Institute (in Hebrew). Retrieved from http://brookdale.jdc.org.il/?CategoryID=192&ArticleID=26

Kahan-Straucznyczy, P., Levi, D., & Konstantinov, V. (2010). Immigrant youth in Israel: The current situation (Research Report No 561-10). Jerusalem: Myers-JDC-Brookdale Institute (in Hebrew). Retrieved from http://brookdale.jdc.org.il/?CategoryID=192&ArticleID=218

Kahan-Straucznyczy, P., Amiel, S., Levi, D., & Konstantinov, V. (2013). First and second generations of immigrant youth...
from Ethiopia and the former Soviet Union—similarities and differences. Jerusalem: Myers-JDC-Brookdale Institute: Engelberg Center for Children and Youth and Ministry of Immigrant Absorption. Retrieved from http://brookdale.jdc.org.il/_Uploads/PublicationsFiles/ImmigYouthEthioPSENSUMMARY.pdf

Kauffman, J. M., Conroy, M., Gardner, R., & Oswald, D. (2008). Cultural sensitivity in the application of behavior principles to education. Education & Treatment of Children, 31, 239–262. Retrieved from http://search.proquest.com/docview/202668787?accountid=14765

Konstantinov, V. (2015). Patterns of integration into Israeli society among immigrants from the Former Soviet Union over the past two decades (Research report No.674–15). Jerusalem: Myers-JDC-Brookdale Institute (in Hebrew). Retrieved from http://brookdale.jdc.org.il/_Uploads/PublicationsFiles/Summary-eng(10).pdf

Koprina, R. K. (2008). Improving testing for English Language Learners. New York, NY: Routledge.

Kozulin, A. (1998). Psychological tools, a sociocultural approach to education (pp. 100–129). Massachusetts: Harvard University Press.

Levin, T., & Shohamy, E. (2008). Achievement of immigrant students in mathematics and academic Hebrew in Israeli school: A large-scale evaluation study. Studies in Educational Evaluation, 34, 1–14. doi:10.1016/j.stueduc.2008.01.001

Levin, T., Shohamy, E., & Spolsky, B. (2003). Educational achievement of immigrant students. Research Report submitted to the Ministry of Education. Tel Aviv University, School of Education (in Hebrew).

Linn, R. L. (1994). The use of Differential Item Functioning statistics: A discussion of current practice and future implications. In P. W. Holland & H. Wainer (Eds.), Differential Item Functioning (pp. 349–366). Hillsdale, NJ: Lawrence Erlbaum.

Litvak, A. (2008). Everybody is an academician, everybody knows mathematics: Policy of absorbing FUS immigrant students in the education system, public ideas and perception of variety. (M.A. dissertation in Public Policy). Jerusalem: Hebrew University of Jerusalem, Faculty of Social Sciences, School of Public Policy (in Hebrew).

Lubinski, S. T. (2001, April). Class, ethnicity, culture and mathematics problem solving. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Martiniello, M. (2008). Language and the performance of English-Language Learners in math word problems. Harvard Educational Review, 78, 333–368.

Maryland Accommodations Manual. (2012). Selecting, administering, and evaluating the use of accommodations for instruction and assessment. Retrieved from http://www.marylandpublicschools.org/MSDE/testing/

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. Educational Researcher, 18, 5–11. doi:10.3102/0013189X018002005

Micheaudès, M. P. (2010). Sensitivity of equated aggregate scores to the treatment of misbehaving common items. Applied Psychological Measurement, 34, 365–369.

Ministry of Education. (2007). Managing director circular 2008/3(0)—Test accommodations for students with special needs in national tests (Meitaz/hundred terms/state-religious current tests) at elementary school and junior high school. Jerusalem: Author (in Hebrew).

Ministry of Education. (2014). Managing director circular 2015/2(0)—The national plan for meaningful learning: Rights of immigrant students and citizens returning to Israel who take the matriculation exams. Jerusalem: Author (in Hebrew).

Muhr, T. (2004). User’s Manual for ATLAS.ti 5.0. ATLAS.ti Scientific Software Development Gmbh, Berlin Computer software.

Muniz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in item translations. International Journal of Testing, 1, 115–135.

National Council of Teachers of Mathematics. (2000). Principles and standards for school mathematics. Reston, VA: Author. Retrieved from http://www.nctm.org/standards

National Institute for Testing and Evaluation. (2016). Psychometric entrance test—test languages. Retrieved from https://www.nire.org.il/index.php/en/tests/psychometric/test-languages.html

Niznik, M. (2008). How to be an alien: Cross-cultural transition of Russian-speaking youth in Israeli high schools. Israel Studies Forum, 23, 66–83.

Niznik, M. (2011). Cultural practices and preferences of ‘Russian’ youth in Israel. Israel Affairs, 17, 89–107. doi:10.1080/13537121.2011.522072

Okhtain, E. (2007). Multicultural considerations in evaluation. Paper presented at the 23rd Language Testing Symposium of the Academic Committee for research on Language Testing, Open University, Ra’anana, Israel.

Pennock-Roman, M., & Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. Educational Measurement: Issues and Practice, 30, 10–28. doi:10.1111/j.1745-3992.2011.00207.x

Perkins, J., & Flores, A. (2002). Mathematical notations and procedures of recent immigrant students. Mathematics Teaching in the Middle School, 7, 346–351. Retrieved from http://search.proquest.com/docview/2311703579?accountid=16765

Pollitt, A., Marriott, C., & Ahmed, A. (2000, May). Language, contextual and cultural constrains on examination performance. Paper presented at the 26th conference of the International Association for Educational Assessment, Jerusalem.

Polya, G. (2009). How to solve it: A new aspect of mathematical method. Bronx, NY: Ishi Press International.

Presmeg, N. (2007). The role of culture in teaching and learning mathematics. In F. K. Lester (Ed.), Second handbook of research on mathematics teaching and learning (pp. 435–458). Charlotte, NC: Information Age Publishing.

Russell, M. (2011). Computer tests sensitive to individual needs. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), Handbook of accessible achievement tests for all students (pp. 255–273). New York, NY: Springer.

Scheuneman, J. D., & Grima, A. (1997). Characteristics of quantitative word items associated with differential performance for female and black examinees. Applied Measurement in Education, 10, 299–319. doi:10.1080/1045988070110843

Schissel, J. L. (2012). The pedagogical practice of test accommodations with emergent bilinguals: Policy-enforced washback in two urban schools (Doctoral Dissertation, Paper AAI3551552). ProQuest. Retrieved from http://repository.upenn.edu/dissertations/AAI3551552

Scribner, S. (1997). Mind in action: A functional approach to thinking. In M. Cole, Y. Engstrom, & O. Vasquez (Eds.), Mind, culture, and activity: Seminal papers from the laboratory of Comparative Human Cognition (pp. 354–368). Cambridge: Cambridge University Press.

Shaffel, J., Belton-Kocher, E., Glasnapp, D., & Poggi, J. (2006). The impact of language characteristics in mathematics test items on the performance of English Language Learners and students with disabilities. Educational Assessment, 11, 105–126.

Shohamy, E. (2001). The power of tests: A critical perspective on the uses of language tests. Harlow: Pearson Education Limit.
Sireci, S. G., & Rios, J. A. (2013). Decisions that make a difference in detecting Differential Item Functioning. Educational research and evaluation: An international journal on theory and practice, 19, 170–187. doi:10.1080/13803611.2013.767621

Solano-Flore, G., & Li, M. (2013). Generalizability theory and the fair and valid assessment of linguistic minorities. Educational research and evaluation: An international journal on theory and practice, 19, 245–263. doi:10.1080/13803611.2013.767632

Solano-Flore, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. Journal of Research in Science Teaching, 38, 553–573.

Solano-Flore, G., Barnett-Clarke, C., & Kachchaf, R. R. (2013). Semiotic structure and meaning making: The performance of English Language Learners on mathematics tests. Educational Assessment, 18, 147–161. doi:10.1080/10627197.2013.814515

Starks-Martin, G. (1996, March). Using “think-alouds”, journals, and portfolios to assess hmong students' perceptions of their Study/Learning strategies. Paper presented at the annual meeting of the teachers of English to speakers of other languages, Chicago, IL.

Statistical Abstract of Israel, no 65. (2014). Immigrants, by period of immigration, country of birth and last country of residence. Retrieved from http://www.cbs.gov.il/reader/shnaton/templ_shnaton_e.html?num_tab=st04_04&CYear=2014

Staub, F. C., & Reusser, K. (1995). The role of presentational structures in understanding and solving mathematical word problems. In C. A. Weaver, S. Mannes, & C. R. Fletcher (Eds.), Discourse comprehension: Essays in honor of Walter Kintsch (pp. 285–305). Hillsdale, NJ: Lawrence Erlbaum.

Suarez-Orozco, M., & Suarez-Orozco, C. (2015). Children of immigration. Phi Delta Kappan, 97, 8–14.

Tatar, M., Kfir, D., Sever, R., Adler, C., & Regev, H. (1994). Integration of immigrant students into Israeli elementary and secondary schools: A pilot study. Jerusalem: The NCJW Institute for Innovation in Education, School of Education, The Hebrew University (in Hebrew).

Thurlow, M. L., Thompson, S. J., & Lazarus, S. S. (2006). Considerations for the administration of tests to special needs students: Accommodations, modifications, and more. In S. M. Downing & T. M. Haladj (Eds.), Handbook of test development (pp. 653–673). Mahwah, NJ: Lawrence Erlbaum.

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Hanniss, M. (1998). Accommodating students with disabilities on large-scale tests: An experimental study. Exceptional Children, 64, 439–450. Retrieved from http://search.proquest.com/docview/201089943?accountid=14765

Tokatly, R. (1992). Absorption of immigrant students (goals, difficulties, achievements). In M. Baizer & R. Perlmuter (Eds.), Know where we have come from: Socio-Historical background of PUS immigrants (pp. 145–155). Jerusalem: Ministry of Education and Culture, Authority of Immigration Absorption (in Hebrew).

Uiterwijk, H., & Vallen, T. (2005). Linguistic sources of item bias for second generation immigrants in Dutch tests. Language Testing, 22, 211–234. doi:10.1191/0265532205lt301oa

Van Herwegen, J., Farran, E., & Annaz, D. (2011). Item and error analysis on Raven’s coloured progressive matrices in Williams syndrome. Research in Developmental Disabilities, 32, 93–99.

Wolf, M. K., Hao, J. C., Rivera, N. M., & Chang, S. M. (2012). Accommodation practices for English Language Learners in states’ mathematics assessments. Teachers College Record, 114, 1–26.

Wu, A. D., & Erickan, K. (2006). Using multiple-variable matching to identify cultural sources of Differential Item Functioning. International Journal of Testing, 6, 287–300. doi:10.1207/s15327574ijt0603_5
Appendix A

Mapping sentence of the potential sources of difficulty in the performance of a single item (sources of difficulty which are irrelevant to the measured construct were highlighted)

Pupil x responds to item y from cluster requiring from a

linguistic point of view at the word level and at the sentence level it

Facet A
Item cluster
a1 DIF items advantageous to native born children
a2 Non-DIF items but appearing in the cluster of an item with DIF

Facet B
Word level
b1 decoding of content words and functioning words
b2 decoding of abstract nouns
b3 decoding of construct state:
  b3.1 associated with mathematical language
  b3.2 not associated with mathematical language
b4 morphological accuracy
b5 combinations of the above

Facet C
Sentence level
   c1 decoding and/or understanding of a relative clause
   c2 decoding and/or understanding of impersonal structure
   c3 combinations of the above

Facet D
Text level
d1 reference
d2 using mathematical language
d3 instructions appearing in the item
d4 words or expressions which can have more than one interpretation
d5 combinations of the above

Moreover, from a cognitive aspect it is associated with mathematical language which requires

Facet E
Mathematical language requirements
e1 proficiency in mathematical vocabulary
e2 ability to shift (translation) from one representation mode to another
e3 combinations of the above

as well as mastery of procedural or declarative knowledge which includes

Facet F
Proficiency in procedural knowledge or declarative knowledge
f1 proficiency in the table of multiplication
f2 applying correct operations on the data
f3 responding correctly according to the expected operation
f4 combinations of the above

with cognitive requirements originating in

Facet G
Special item characteristics
   g1 data redundancy
g2 information overload stemming from the need to perform several calculation steps
g3 lack of similarity to a school item, leading to:
   g3.1 irrelevant reference
g3.2 misinterpretation of the item
g3.3 problems of understanding
   g4 requirement for execution functions
g5 combinations of the above

Facet H
Test-taking skills
h1 complete reading of the item
h2 matching between the answer given orally and the one given in writing
h3 an answer set down in the item gives
h4 combinations of the above

In the area of test-taking skills it requires
In the area of **socio-cultural characteristics** it relates to

and in the area of **assignment characteristics** it is characterized as checking a mathematical literacy dimension defined as

When relating to characteristics associated with its representation it refers to

of the pupil’s performance.