JSSE

Journal of Social Science Education

2023, Vol. 22 (2)

Edited by:
Jan Löfström,
David Rosenlund &
Birgit Weber

Article

Measuring citizenship competences: Assessment of measurement invariance

Hoek, L.[a], **Zijlstra, B. J. H.**[a], **Munniksma, A.**[a], & **Dijkstra, A. B.**[ab]

[a] Department of Child Development and Education, University of Amsterdam, Amsterdam, The Netherlands
[b] Dutch Inspectorate of Education, Utrecht, The Netherlands

Keywords: Citizenship competences; Measurement invariance; Measurement equivalence; Citizenship education; Social outcomes.

Highlights:

- Standardised questionnaires are used to measure the outcomes of citizenship education.
- A prerequisite for cross-group comparisons based on these questionnaires is an assessment of measurement invariance.
- This study used data from 6035 students from 87 Dutch primary schools to examine the measurement invariance of citizenship knowledge, attitude, and skill across sex, socioeconomic position and migration background.
- The measurement invariance was sufficient in most cases.
- Periodic assessment of measurement invariance in instruments measuring citizenship competences is important due to the dynamic nature of the construct.

Purpose: Standardised questionnaires are used to measure the outcomes of citizenship education. These outcomes are often compared across groups to document different outcomes, for example, between boys and girls. A prerequisite for cross-group comparisons is an assessment of measurement invariance.

Methodology: This study used data from 6035 students from 87 Dutch primary schools to examine the measurement invariance of the Citizenship Competences Questionnaire (Ten Dam et al., 2011). Dutch schools use this questionnaire to gain insight into students' citizenship knowledge, attitudes, and skills. Measurement invariance was assessed across sex, socioeconomic position, and migration background.

Findings: Measurement invariance was sufficient in most cases, allowing for cross-group comparisons of associations between latent constructs and their indicators, and in some cases, for cross-group comparisons of the latent means. We conclude by emphasising that periodic assessment of measurement invariance in instruments measuring citizenship competences is important due to the dynamic nature of the construct.

Corresponding author: Lianne Hoek, MSc., Nieuwe Achtergracht 127, 1018WS Amsterdam, The Netherlands. E-Mail: l.h.m.hoek@uva.nl

Suggested citation:
Hoek, L., Zijlstra, B. J. H., Munniksma, A., & Dijkstra, A. B. (2023). Measuring citizenship competences: Assessment of measurement invariance. In: *Journal of Social Science Education 22(2).* https://doi.org/10.11576/jsse-5837

Declaration of conflicts of interests: No potential conflict of interest was reported by the authors.

1 INTRODUCTION

Like other educational domains, it is essential to gain insight into what students learn in citizenship education. This insight can be used to facilitate the learning process and evaluate the contents and delivery methods of the curriculum. There are several ways to obtain insight into student outcomes, including standardised measurement instruments (Daas, Ten Dam, & Dijkstra, 2016). Using standardised measurement instruments is beneficial in many ways, for example, because of their practical usefulness and the opportunities for securing the quality and validity of the measurement. This is why standardised questionnaires have been widely used to measure students' citizenship competences in terms of knowledge, attitude, and skill (Ireland, Kerr, Lopes, Nelson, & Cleaver, 2006; Schulz et al., 2018).

Multi-group comparisons based on standardised questionnaires have demonstrated that citizenship competences are related to students' background characteristics, such as sex, socioeconomic position (SEP), and migration background - both internationally (Ireland et al., 2006; Kerr et al., 2007) and in the Dutch school system in which our data were gathered (Dijkstra, Geijsel, Ledoux, Van der Veen, & Ten Dam, 2015; Geijsel, Ledoux, Reumerman, & Ten Dam, 2012). These findings are important for educational practice because they may lead to adaptations in the contents or delivery methods of citizenship education so that the learning objectives are met, and all students benefit optimally. However, a prerequisite for meaningful comparisons across different population groups (e.g., boys and girls) is that the construct to be measured is understood similarly in each group (Isac, Palmerio, & Van der Werf, 2019; Steinmetz, Schmidt, Tina-Booh, Wieczorek, & Schwartz, 2009). In other words: it is essential to know whether the difference in citizenship competences between, e.g., boys and girls, is a 'true' difference or a difference caused by boys and girls systematically understanding questions in a different way. To determine whether a questionnaire measures the same across groups, the measurement invariance should be assessed (Meredith, 1993).

Assessment of measurement invariance is important for the methodological quality of a measurement instrument (Meuleman et al., 2022). However, as the interpretation of a construct can change over time (Putnick & Bornstein, 2016), the establishment of measurement invariance is not a static given. This applies in particular to citizenship education, which is a dynamic construct that, like other societal phenomena, takes on meaning in and moves along with changes in society (Mattei & Broeks, 2018). This underscores the importance of periodic assessment of measurement invariance in student questionnaires, as is the case for measuring citizenship competences. Examples of such changes, amongst others, include increased cultural diversity (Eurostat, 2021; US Census Bureau, 2020) or social inequality (Organisation for Economic Co-operation and Development, 2017). Both changes, for instance, may have influenced the public debate and, resultingly, have led to different interpretations of constructs related to citizenship education.

For example, students with and without migration backgrounds may perceive debates about how to deal with socioeconomic or socio-cultural differences differently nowadays, in times of what is sometimes referred to as a shift from diversity to ‘super-diversity’ (Vertovec, 2007), as opposed to earlier in time, also influencing the response in measurement instruments regarding these topics. Hence, it is important to periodically assess the measurement invariance in instruments that measure citizenship competences, in case changes in society may have shifted the norms and values of citizens regarding value-sensitive themes (Munck, Barber, & Torney-Purta, 2018).

Whereas both conceptualisation and operationalisation are vital steps in developing measurement instruments, the assessment of measurement invariance relates explicitly to the operationalisation of a measurement instrument. It assumes that the conceptualisation of the instrument is carefully considered and, therefore, not associated with differences in student characteristics like sex, SEP, migration background, religiosity, age and others. The assumption of such a generic conceptualisation of citizenship is the basis of large-scale standardised measurement instruments for citizenship competences. National and international examples include instruments like the International Civic and Citizenship Education Study (ICCS) for international comparisons used in periodic cycles (Schulz et al., 2018), the Citizenship Education Longitudinal Study (CELS) for longitudinal research in the UK (Cleaver, Ireland, Kerr, & Lopes, 2006), or the Citizenship Competences Questionnaire (CCQ) for annual measurements used by schools in The Netherlands (Ten Dam, Geijsel, Reumerman, & Ledoux, 2011). In the conceptualisation of such instruments, important building blocks are core democratic values, trust, participation and involvement, and the support for institutions of a democratic society. Such instruments generally move away from conceptualisations that take a position regarding an ‘ideal’ balance between values like, for example, individual freedom and solidarity towards others – also known as contested or school-specific goals in citizenship education (Eidhof, Ten Dam, Dijkstra, & Van de Werfhorst, 2016). Rather, these conceptualisations adhere to consensus goals, referring to commonly shared values of reciprocity of ‘treating others as you want others to treat you’ – also known as *regula aurea* or the Golden Rule (cf. Etzioni, 1996; Wattles, 1996).

Despite its importance, research on measurement invariance in instruments for citizenship competences is still scarce. It is predominantly assessed in between-country research (Byrne & Watkins, 2003), such as the ICCS. The technical report of the ICCS (Schulz, Carstens, Losito, & Fraillon, 2016), for example, demonstrated that most of the citizenship constructs show measurement invariance up to a level where associations of latent constructs can be safely compared across countries, but not always to the level where comparison of means of these latent constructs is justified. In addition, Isac et al. (2019) assessed the measurement invariance of a specific part of the ICCS questionnaire that focuses on young people’s tolerant attitude towards immigrants (i.e., support for equal rights). The authors found that most items were measurement invariant to the level where average scale scores can be validly compared across European countries. However, the assessment of measurement invariance is also important in within-country comparisons (Steinmetz et al., 2009; Vandenberg & Lance, 2000). In this respect, some researchers mentioned that homogeneity of the population in terms of the measures being compared across groups is often implicitly assumed, regardless of whether it was assessed (Muthén, 1989; Steinmetz et al., 2009).

This study aims to examine the measurement invariance of a Dutch questionnaire for assessing citizenship competences across groups. By doing so, we assess whether this questionnaire that was designed over fifteen years ago is still[1] valid across different groups in society today, as societal changes may have affected how different groups interpret the questionnaire items. In this way, we contribute to the insight into the assessment of measurement invariance in within-country comparative research on citizenship competences, as this empirical knowledge base on measurement invariance in these questionnaires is still scarce. Indirectly, this study sheds light on whether the measurement of citizenship competences is robust amid changes in society and population. For educational practice, this study serves as a validity check of a measurement instrument that schools use to improve their citizenship education.

Following previous research showing relevant and robust differences in citizenship competences based on sex, SEP, and migration background (Geijsel et al., 2012; Ireland et al., 2006; Kerr et al., 2007), we focus on the assessment of measurement invariance across these three student-background characteristics. The findings of this study may underscore the results of these previous studies that are largely built on the same generic conceptualisation of citizenship competences, in case the measurement instrument appears largely measurement invariant or, alternatively, place caution on the findings of these studies in case measurement invariance could not be established. Moreover, various scholars have advocated investigating measurement invariance over sex, SEP, and migration background (Kline, 2015; Wray-Lake, Metzger, & Syvertsen, 2017) or have stressed the importance of examining similarities and differences in citizenship competences based on these groups (Cleaver et al., 2006). The instrument used for testing measurement invariance is the CCQ, a large-scale standardised measurement instrument based on a generic conceptualisation of citizenship competence. The CCQ is used primarily for annual measurements of students' citizenship competences at Dutch schools. The questionnaire has been used for over a decade – yielding a rich dataset of consecutive cohorts that is eminently suitable to assess whether contextual changes have deteriorated alignment between the instrument and its context.

[1] Still, as was the case during the initial construction phase when the measurement invariance was tested as part of a broad set of psychometric tests, meeting all necessary requirements (Ten Dam et al., 2011; Geijsel et al., 2012; based on personal communication, because these results were not available in print).

2 Theory

2.1 Meaningful comparison of groups

In this section, building on He and Van de Vijver (2013) and Isac et al. (2019), we outline three examples of how the operationalisation of measurement does not measure the same across groups. We also provide a 'counter-example' of a difference between groups that is not the result of measurement non-invariance but an example of a 'true' difference. To illustrate our examples, we used existing items from the CCQ (measuring citizenship attitudes, skills, and knowledge of students). Students needed to indicate to what extent the items applied to them on a four-point Likert scale ranging from 'does not apply at all to me' to 'applies completely to me' for attitude items or ranging from 'not good at all' to 'very good' for skill-items. Students had to pick the best answer in the knowledge items by choosing one out of three. The use of these items at this place is merely illustrative.

First, we outline how measurement non-invariance is caused by the fact that some underlying items are not considered indicative of the construct to be measured for some group members. To illustrate this for SEP, we look at the item: 'If we talk about the news in class, I want to add something to the conversation too.' Compared to students with a high SEP, students with a low SEP may have limited access to news sources, such as a subscription to a daily newspaper or a personal mobile device to check news websites. Therefore, students with a low SEP may answer this item with 'does not apply at all to me', whereas they are willing to contribute to the conversation. Therefore, their answer indicates their accessibility to news, not their democratic attitude. This may cause students with a low SEP to be wrongly labelled as 'less capable' of the construct 'democratic attitude' because this specific indicator of the construct is less applicable to their context.

Second, we outline how measurement non-invariance is caused by some group members understanding items differently due to linguistic differences. To illustrate this for migration background, we look at the item: 'How good are you at… holding on to your opinion, if you are really right?' Linguistic differences between students with and without a migration background may influence how students understand constructs or underlying items. Language that involves ambiguity in meaning (e.g., 'holding on to something', 'if you are really right') or metaphorical language is particularly susceptible to differences in interpretation, for example, by students who have another mother language (more likely being students with a migration background). However, vice versa, multilingual students (more likely students with a migration background) may also benefit from their knowledge and skill in language comprehension. This may be less common for monolingual students (more likely students without a migration background). Either way, such linguistic differences may distort the results and wrongly label students as more or less skilled in citizenship competences.

Likewise, group members may understand items differently due to cultural differences. To illustrate this for sex, we look at the item: 'It is normal to help in the household (for example, by preparing the dinner table, tidying up or cleaning)'. In some cultures, helping in the household is considered predominantly a task for girls and not for boys. Boys who grow up in such a culture may be more likely to answer this question with 'does not apply to me at all', whereas their attitude towards 'acting in a socially responsible manner' may, in reality, be different than this answer reveals. Or vice versa: girls who grow up in such a culture may be more likely to answer this question with 'applies completely to me', whereas their attitude towards 'acting in a socially responsible manner' may, in reality, be different than this answer reveals.

Third, we outline how measurement non-invariance is caused by some groups characterised by a specific response style. To illustrate this for sex, we look at the item: 'People who earn sufficient salary should together take care of people with less wealth'. Suppose that girls are more likely to answer items in a more socially desirable way and systematically answer this item with 'applies completely to me', whereas boys answer this item in a less socially desirable way. This may cause girls to overestimate their attitude toward the construct 'acting in a socially responsible manner'.

At last, we outline how a difference can be not the result of measurement non-invariance but a 'true' difference between groups. To illustrate this, we look at the item: 'In a sports game, the referee takes a wrong decision *against* your team. What should you do?' And the following answer options: (a) Go to the referee and debate the decision; (b) Get the coach of your sports team; (c) Keep on playing because the decision of the referee is directive during the game. The latter is appointed the preferred answer. Student background characteristics such as sex, SEP, or migration background may influence how group members respond to this item. However, these differences are not an example of measurement non-invariance when the differences do not adhere to differences in understanding, interpretation, or applicability to the context, but rather are an example of 'true' differences in what is valued about acting in a socially responsible manner. Regardless of these 'true' differences, the conceptualisation of acting in a socially responsible manner holds that if participants of a sports game agree on rules beforehand, they conform to these rules during the sports game – even if they do not agree during the game.

2.2 Assessment of measurement invariance

Multiple-group confirmatory factor analysis (MGCFA) is the most commonly used method to assess measurement invariance (Putnick & Bornstein, 2016). In MGCFA, hierarchical, subsequent models with increasing restrictions are specified and compared. These are the configural model, the metric model, and the scalar model.

The *configural model* assesses whether the instrument measures the same latent factors across groups and whether the indicators are the same across groups (Isac et al., 2019). To test for configural invariance, models with the same pattern (i.e., the same configuration) should be specified across groups (Vandenberg & Lance, 2000), meaning an equal number of latent variables, indicators, et cetera. If the configural model yields poor model fit, it indicates that in one of the groups, a different pattern fits the data. For example, for boys, one of the questions may not be an indicator of the latent construct, whereas, for girls, it is. If the configural model fits the data well, it provides ground for testing metric invariance.

The *metric model* assesses whether the factor loadings differ across groups (Horn & Mcardle, 1992). To test for metric invariance, the factor loadings are constrained to equality across groups (Reeskens & Hooghe, 2010). Thus, for example, two equal path models with the same pattern are specified with the first factor loading in the model for boys constrained to equality to the first factor loading in the model for girls, and so on for the remaining factor loadings. If the metric model yields poor model fit, it indicates that in one of the groups, a factor loading relates differently to the latent variable as compared to the other group (e.g., for boys, the third item is strongly related to the latent variable, but for girls, this is not the case). If the metric model fits the data well, it allows for subsequent analyses of testing scalar invariance. Reaching metric invariance justifies the comparison of latent variables and their associations across groups (Isac et al., 2019). In addition, achieving (partial) metric invariance is seen as a minimal prerequisite for meaningful cross-group comparisons (Little, 2013).

The *scalar model* assesses whether the intercepts (i.e., the constant or the scalar) of the indicators are the same across groups (Steenkamp & Baumgartner, 1998). To test for scalar invariance, the intercepts of the indicators are constrained to equality across groups. Thus, in addition to the identical pattern (configural model) and equal factor loadings (metric model) across groups, the intercepts are constrained to equality. This means that the intercept of the first item for, e.g., boys is constrained to equality to the intercept of the first item for girls, and so on for the intercepts of the remaining items. If the scalar model yields poor fit, at least one intercept differs across groups (Isac et al., 2019). If the scalar model fits the data well, it justifies comparing latent means across groups (Reeskens & Hooghe, 2010).

3 METHODOLOGY

In this study, we examined the measurement invariance in the assessment of citizenship competences (i.e., the competence of students to function and participate in society) across sex, SEP, and migration background. We did so by looking at competences of students in terms of attitude (i.e., thoughts, desires, and willingness), knowledge (i.e., knowing, understanding, insight), and skill (i.e., an estimate by students of what they think they are able to).

3.1 Data

We used data of consecutive cohorts from the CCQ (Ten Dam et al., 2011). Schools use the CCQ to measure students' citizenship knowledge, attitude and skill in terms of four so-called ‘social tasks’: acting democratically, acting socially responsible, dealing with conflicts, and dealing with differences (Ten Dam et al., 2011). The CCQ is suitable for grades 5 and 6 of primary education (approximate age is 10 to 12 years old) and grades 7, 8, and 9 from secondary education (approximate age is 12 to 16 years old). We retained this study’s scope to data gathered in primary education for pragmatic reasons. In this sample, we merged data from grades 5 and 6 to obtain larger group sizes and more robust results.

3.2 Sample and procedure

The sample consists of 6035 students from 87 primary schools participating in the Dutch’ Alliance Citizenship’ (Table 1). The Alliance Citizenship is a partnership of schools that organises annual measurements of citizenship competences of students. This study used data from 2015, 2016, 2017, 2018, and 2019. We used a sample that was both recent and large enough to detect possible changes in society and population composition. Each year, the CCQ is online available to participating schools during spring. The sample of schools differed each year: some schools participated more than once; some only once.[2]

Table 1

Descriptive information of the sample

	2015	2016	2017	2018	2019	Total
Primary schools (N)	26	18	16	11	19	87
Students (N)	1587	1349	1176	564	1359	6035
Grade						
Grade 5	809	675	605	296	678	3063
Grade 6	778	674	571	268	681	2972
Age						
10 years or younger	309	225	198	121	339	1192
11 years	676	638	570	272	643	2799
12 years	389	391	315	137	301	1533
13 years	55	33	30	9	19	146
14 years	1	3	0	0	0	4
15 years	0	0	0	0	0	0
16 years or older	1	0	1	1	1	4
Sex						
Boy	734	642	555	254	628	2813
Girl	697	648	559	285	675	2864
SEP						
Low SEP [a]	430	402	393	192	535	1952
High SEP [b]	608	526	425	178	364	2101

[2] The possibility that some students were measured twice (i.e., from grade 5 in one years’ measurement to grade 6 in the following years’ measurement), as well as the nested structure of the data, may have caused a degree of dependency that we did not account for. In practice, it means that our results may be somewhat too conservative.

Migration background						
No migration background [c]	1102	989	847	401	1086	4425
Migration background [d]	334	302	267	139	217	1259

[a] Highest educational level of mother and father is 'no school', 'only primary education' or 'only secondary education'.

[b] Highest educational level of mother and father is 'higher education'.

[c] Both parents are born in The Netherlands.

[d] At least one parent is born outside of The Netherlands.

3.3 Variables

This section provides information on the conceptual framework underlying the questionnaire that we examined in our analyses (Table 2). The values for Cronbach's alpha are calculated with the sample used in the present study. The values were largely consistent with the original alpha's provided by Ten Dam et al. (2011), based on 16,000 students from grade 6 and grade 9 who participated in 2005, 2006, and 2007.[3] The variables Attitude – Acting democratically and Skill – Acting democratically are both presented in two interpretable factors. The variable Skill – Acting in a socially responsible manner is presented in a combined factor with the scale Skill – Dealing with conflicts.

For the Knowledge items, students chose the answer they thought was the best response. An example is: *All children have a right to… (a) an allowance; (b) choose who they want to live with;* or *(c) education*. The correct answer here is c. All knowledge items were dichotomised into correct (1) or incorrect (0). The phrasing of the Attitude items is: *How well does this statement apply to you?* A sample statement is: *I like knowing something about different religions*. The response options are: *(1) does not apply at all to me, (2) does not apply much to me, (3) applies a fair amount to me,* or (*4) applies completely to me*. The phrasing of the Skill items is: *How good are you at…?* A sample statement is: … *finding a solution that everyone is satisfied with for a conflict?* The response options are: *(1) not good at all; (2) not very good; (3) pretty good;* or *(4) very good.*

For sex, boys (49.55%) were appointed with value 1 and girls with value 2. For SEP, we converted maternal and paternal educational levels into a new variable, indicating whether the mother or father's highest obtained educational level is either no school, primary school, or secondary school (value 1) or higher education (value 2). In our sample, 48,16% of all students had a low SEP. For migration background, we converted maternal and paternal country of birth into a new variable, indicating whether both parents of the students were born in the Netherlands (value 1) or whether at least one parent of a student has a migration background, regardless of the country of origin (value 2). In our sample, 77,85% of all students had both parents born in the Netherlands.

[3] In order to only share robust results, the results of the questionnaire are reported at the overarching scale-level (e.g., knowledge, attitude, or skill), and not on subscale-level. We performed the analyses of measurement invariance on both the scale- and subscale-level to gain more information on exact items that could be hindering reaching a higher level of measurement invariance. Hence, we provided the values for Cronbach's alpha also on the subscale-level (of which some values are below the advised threshold of 0.70 (Cortina, 1993)).

3.4 Analytic plan

We conducted MGCFA and compared groups based on sex (boys versus girls), SEP (low versus high), and migration background (having no migration background versus having a migration background). Analyses were conducted with R version 3.5.0 (R Core Team, 2021), using the lavaan package (Rosseel, 2012) and the semTools package (Jorgensen, 2021), with particular focus on the guidelines in the measEq.syntax. While doing so, analyses were performed according to a the following principles.

First, we treated the indicators that were directly observed (questionnaire items) as ordered categorical factors because they were measured using a four-point Likert scale (attitude and skill) or a three-option multiple-choice (knowledge). Therefore, they could not be treated as continuous indicators. Assessing measurement invariance using ordered categorical factors makes the procedure more complex because it also involves testing for threshold invariance. Testing for threshold invariance is a means to do justice to the ordered categorical nature of indicators. The threshold model assumes that a normally distributed latent item-response lies underneath each observed categorical indicator (Kite, Jorgensen, & Chen, 2018). The threshold can be seen as a 'tipping point' between the different category responses (e.g., between answering 'does not apply at all to me' and 'does not apply much to me'). The threshold model is estimated before the metric model, but only if polytomous indicators are involved (thus: only for attitude and skill).

In the case of dichotomous indicators (such as the knowledge items), it is not possible to distinguish between equality constraints on the thresholds (testing threshold invariance), factor loadings (testing metric invariance), and intercepts (testing scalar invariance) (Wu & Estabrook, 2016). Instead, the equality constraints need to be added simultaneously. Therefore, we tested only for configural and scalar invariance for dichotomous indicators.

Table 2

Conceptual framework of citizenship competences (derived from Ten Dam et al., 2011)

Components	**Knowledge (α = .79)** Knowing, understanding, insight	**Attitudes (α = .89)** Thoughts, desires, willingness	**Skills (α = .87)** Estimate of what one can do
Social tasks	A young person with such knowledge…	A young person with such attitudes…	A young person with such skills…
Acting democratically Acceptance of and contribution to a democratic society	… knows what democratic principles are and what acting in accordance with them involves (8 items, α = .67)	… *wants to hear everyone's voice, enter into a dialogue* (3 items; α = .65) *and make an active, critical contribution* (3 items; α = .65)	… *is able to assert own opinions* (3 items; α = .74) *and listen to the opinions of others* (3 items; α = .68)
Acting in a socially responsible manner Taking shared responsibility for the communities to which one belongs	… knows social rules (i.e. legal or unspoken rules for social interaction) (6 items, α = .54)	… wants to uphold social justice, is prepared to provide care and assistance, does not want to harm another or the environment as a result of his or her behavior (6 items, α = .68)	… can adopt a socially just position (5 items, α = .76)
Dealing with conflicts Handling of minor situations of conflict or conflicts of interest to which the child is a party	… knows methods to solve conflicts such as seeking win-win solutions, calling in help from others, admission of mistakes, prevention of escalation (7 items, α = .62)	… is willing to explore conflicts, prepared to consider the standpoint of another, jointly searches for an acceptable solution (6 items, α = .79)	… can listen to others, put oneself in someone else's place, seek win-win solutions (5 items, α = .76)
Dealing with differences Handling of social, cultural, religious, and outward differences	… is familiar with cultural differences, has knowledge of rules of behavior in different social situations, knows when one can speak of prejudice or discrimination (6 items, α = .63)	… has a desire to learn other people's opinions and lifestyles, has a positive attitude towards differences (6 items, α = .85)	… can adequately function in unfamiliar social situations, adjust to the desires or habits of others (4 items, α = .67)

3.5 Data preparation

All empty cases were dropped via list-wise deletion by removing all rows with less than nine items answered. This removed the rows in which only certain standard school characteristics (e.g., unique identifier for school, year and class) were automatically filled in (8 items) but none of the actual questionnaire items. The resulting data had 1.28% missing observations that did not show any pattern of missingness. The percentage of missing data is considered acceptable as basis for further analyses (Bennett, 2001). Second, four variables were excluded from our subsequent analyses because they consisted of three items, which was insufficient as input to our fit measures in the configural model. These excluded variables are: ‘Attitude – Acting democratically 1’, ‘Attitude – Acting democratically 2’, ‘Skill – Acting democratically 1’, and ‘Skill – Acting democratically 2’.

3.6 Model selection procedure

To assess how well the configural, threshold, metric, and scalar models fit our data, we consulted the Chi-square test of overall model fit (χ^2), and the relative difference in model fit ($\Delta\chi^2$), the comparative fit index (*CFI*), and the root mean square error of approximation (*RMSEA*) – where the CFI and RMSEA are functions from the Chi-square test statistic (Shi, Lee, & Maydeu-Olivares, 2019). We followed five decision rules in the model selection procedure using these model fit measures.

The first decision rule is based on the *CFI*. We consulted the *CFI* to indicate how well the model fit improved compared to the null model. The *CFI* considers the complexity of the model and ranges from 0 to 1. A value above 0.95 is preferred (Hu & Bentler, 1999).

The second decision rule is based on the *RMSEA*. The RMSEA indicates the “badness-of-fit” (Adelson, 2012) and considers the model complexity by estimating approximation error per model degree of freedom. Larger values of the *RMSEA* indicate a worse model fit. An *RMSEA* lower than 0.05 indicates ‘close fit’, and an *RMSEA* between 0.05 and 0.08 indicates ‘satisfactory fit’ (Browne & Cudeck, 1993). In this study, we perceived a value for RMSEA ≤ 0.08 as acceptable.

The third decision rule is based on the Chi-square test of overall model fit. This fit index tests the assumption of respectively configural, threshold (if applicable), metric (if applicable), and scalar invariance. A downside of the Chi-square test for overall model fit is that even minor deviations from the suggested models can reject the null hypotheses of model fit for large samples (Shi et al., 2019). Therefore, we proceeded to test additional invariance models as long as values for *CFI* and *RMSEA* of the configural model are acceptable – even if the Chi-square value is significant. At the same time, we performed permutation tests to ensure that the rejection of the overall model fit is most likely due to the same underlying reason across groups (Jorgensen, 2017). In permutation tests, all observations are randomly reassigned (a thousand times) to groups. These permuted group compositions are expected to fit the data equally poor or well as the hypothesised group composition (in which we ‘manually’ assigned, e.g., all boys to the one group and all girls to the other group). If this is the case, it is indicated by a non-significant p-value. However, a significant p-value suggests that there might be different underlying reasons across groups causing the significant Chi-square test of overall fit.

The fourth decision rule is based on the p-value of the Chi-square test of the difference in model fit ($\Delta\ \chi^2$). This fit index was applied to test the null hypothesis of equal model fit for two models. If the p-value of the Chi-square test of difference is significant, it indicates a meaningful difference between the two models in comparison, and we reject the more restricted model.

The fifth decision rule entails whether or not to test for partial invariance. A partial model can be seen as an ‘in-between model’ because some, but not all, of the equality constraints may need to be rereleased. To assess what constraints this involves, we inspected standardised mean residuals and modification indices. Standardised mean residuals can be consulted to identify which factor loadings, thresholds, or intercepts differ most across groups and may need to be rereleased. In addition, modification indices estimate how much the Chi-square test statistic of a model decreases if a constrained parameter is set to free again. We used a Bonferroni-adjusted alpha level to see whether modification indices were significant and could be considered released. When values for *CFI* and *RMSEA* are within the thresholds, but the p-value for the Chi-square test of difference is significant, we additionally tested for partial measurement invariance. To maintain analyses within the scope of this article, we only searched for partial invariance if it established a meaningful difference, i.e. reaching partial metric invariance (for polytomous indicators) or partial scalar invariance (for polytomous and dichotomous indicators). We did not test for partial invariance if the meaningful minimum of partial metric invariance (Little, 2013) was still more than one step away (e.g., when the configural model fitted well, but the threshold model did not).

4 RESULTS

We specified two types of first-order one-factor models: (1) models in which social tasks such as ‘acting democratically’ are specified as indicators of overarching constructs such as citizenship attitude, knowledge, or skill; and, zooming in, (2) models in which items from the questionnaire are specified as indicators of the aforementioned social tasks. We summarised the findings in Tables 3, 4 and 5. In this section, we elaborate on the results by first sharing the results of the models with social tasks as indicators of citizenship knowledge, attitude, and skill, and secondly by sharing the results of the models with questionnaire items as indicators of the social tasks.

Table 3

Models with social tasks as indicators of citizenship constructs

		Configural		Metric			Scalar			Partial Scalar		
		CFI	R	CFI	R	$\Delta \chi^2$	CFI	R	$\Delta \chi^2$	CFI	R	$\Delta \chi^2$
Attitude												
	Sex	+	+	+	+	+	-	-	-			
	SEP	+	+	+	+	+	+	+	-	+	+	+
	Migr.	+	+	+	+	-						
Knowledge												
	Sex	+	-									
	SEP	+	-									
	Migr.	+	-									
Skill												
	Sex	+	+	+	+	+	+	+	-	+	+	+
	SEP	+	+	+	+	+	+	+	-	+	+	+
	Migr.	+	+	+	+	+	+	+	+			

Note. Migr. = migration background; R = RMSEA. $\Delta \chi^2$ = Chi-square test of difference. A '+' for CFI indicates that the value for CFI is ≥ 0.95; a '-' indicates that the value is ≤ 0.95. A '+' for RMSEA indicates that the value for RMSEA is ≤ 0.08; a '-' indicates that the value is ≥ 0.08. A '+' for $\Delta \chi^2$ indicates that the p-value of the Chi-square test of difference is ≥ 0.05; a '-' for $\Delta \chi^2$ indicates that the p-value of the Chi-square test of difference is ≤ 0.05.

Table 4

Models with polytomous questionnaire items as indicators of social tasks

	Configural		Threshold			Metric			Partial Metric			Scalar			Partial Scalar		
	CFI	R	CFI	R	$\Delta \chi^2$	CFI	R	$\Delta \chi^2$	CFI	R	$\Delta \chi^2$	CFI	R	$\Delta \chi^2$	CFI	R	$\Delta \chi^2$
Attitude – Acting in a socially responsible manner																	
Sex	+	+	+	+	+	+	+	+				+	+	-			
SEP	+	+	+	+	+	+	+	-	+	+	+				+	+	+
Migr.	+	+	+	+	+	+	+	+				-	+	-			
Attitude – Dealing with conflicts																	
Sex	+	+	+	+	+	+	+	-									
SEP	+	+	+	+	+	+	+	+				+	+	+			
Migr.	+	+	+	+	+	+	+	+				+	+	-	+	+	+
Attitude – Dealing with differences																	
Sex	+	-															
SEP	+	-															
Migr.	+	-															
Skill – Acting in a socially responsible manner / Dealing with conflicts																	
Sex	+	-															
SEP	+	-															
Migr.	+	-															
Skill – Dealing with differences																	
Sex	+	+	+	+	+	+	+	+				+	+	+			
SEP	+	+	+	+	+	+	+	+				+	+	-			
Migr.	+	+	+	+	+	+	+	-	+	+	+				+	+	-

Note. Migr. = migration background; R = RMSEA. $\Delta \chi^2$ = Chi-square test of difference. A '+' for CFI indicates that the value for CFI is ≥ 0.95; a '-' indicates that the value is ≤ 0.95. A '+' for RMSEA indicates that the value for RMSEA is ≤ 0.08; a '-' indicates that the value is ≥ 0.08. A '+' for $\Delta \chi^2$ indicates that the p-value of the Chi-square test of difference is ≥ 0.05; a '-' for $\Delta \chi^2$ indicates that the p-value of the Chi-square test of difference is ≤ 0.05.

Table 5

Models with dichotomous questionnaire items as indicators of social tasks

	Configural		Scalar			Partial Scalar		
	CFI	R	CFI	R	$\Delta \chi^2$	CFI	R	$\Delta \chi^2$
Knowledge – Acting democratically								
Sex	+	+	+	+	+			
SEP	+	+	+	+	+			
Migr.	+	+	+	+	+			
Knowledge – Acting in a socially responsible manner								
Sex	+	+	+	+	+			
SEP	+	+	+	+	+			
Migr.	+	+	+	+	+			
Knowledge – Dealing with conflicts								
Sex	-	+						
SEP	-	+						
Migr.	+	+	+	+	-	+	+	+
Knowledge – Dealing with differences								
Sex	+	+	-	+	-	+	+	+
SEP	+	+	+	+	+			
Migr.	+	+	+	+	+			

Note. Migr. = migration background; R = RMSEA. $\Delta \chi^2$ = Chi-square test of difference. A '+' for CFI indicates that the value for CFI is ≥ 0.95; a '-' indicates that the value is ≤ 0.95. A '+' for RMSEA indicates that the value for RMSEA is ≤ 0.08; a '-' indicates that the value is ≥ 0.08. A '+' for $\Delta \chi^2$ indicates that the p-value of the Chi-square test of difference is ≥ 0.05; a '-' for $\Delta \chi^2$ indicates that the p-value of the Chi-square test of difference is ≤ 0.05.

4.1 Attitude

4.1.1 Social tasks as indicators

Comparing citizenship attitude across sex and SEP, metric invariance was reached. This justifies comparing associations between the construct citizenship attitude and its four underlying social tasks across sex and SEP. Comparing citizenship attitude across migration backgrounds, metric invariance could not be reached. Inspecting modification indices and standardised mean residuals pointed to no possibility of establishing a partial model. Therefore, only configural invariance was reached. This means that the path models specified for students with and without migration backgrounds are likely to follow the same pattern. For an overview, see Table 3.

4.1.2 Items as indicators

Comparing the social tasks within citizenship attitude, (partial) metric invariance was reached in most cases. This allows for comparing the associations between the concerning social tasks and the underlying items across groups. Threshold invariance was achieved for *Attitude – Dealing with conflicts* across sex, but metric invariance could not be established. This means that the thresholds are the same across boys and girls, but the factor loadings between questionnaire items and the social task differ across boys and girls. For comparisons of *Attitude – Dealing with differences* across sex, SEP and migration background, the configural model indicated poor model fit. This suggests that we cannot assume that the path models specified for boys and girls, students with low and high SEP, and students with and without migration background, follow the same pattern; the same number of indicators relating to the same number of latent constructs. For an overview, see Table 4.

4.2 Knowledge

4.2.1 Social tasks as indicators

Comparing citizenship knowledge across sex, SEP and migration background, the configural model indicated poor model fit. This suggests that the data follow a different pattern in one of the groups. For example, it may be the case that one of the indicators of citizenship knowledge, e.g., dealing with differences, is an indicator of citizenship knowledge for students with a low SEP but not for students with a high SEP. For an overview, see Table 3.

4.2.2 Items as indicators

Comparing the social tasks within citizenship knowledge, the minimum of (partial) scalar invariance was reached in most cases. This means that we can meaningfully compare the mean scores of the concerning social tasks across groups. For comparisons of the social task *Knowledge – Dealing with conflicts* across sex and SEP, the configural model indicated poor model fit. For an overview, see Table 5.

4.3 Skill

4.3.1 Social tasks as indicators

Comparing citizenship skill across sex and SEP, metric invariance was reached. This justifies the comparison of citizenship skill and its associations with the underlying social tasks across boys and girls and students with a low and high SEP. Comparing citizenship skill across migration backgrounds, scalar invariance was reached. This means that we found no evidence that the intercepts differ across students with and without migration backgrounds, and we can meaningfully compare the mean scores of citizenship skill. For an overview, see Table 3.

4.3.2 Items as indicators

Comparing the social tasks within citizenship skill, the required minimum of (partial) metric invariance was reached for comparisons of *Skill – Dealing with differences* across sex, SEP, and migration background. This means we may compare the associations between the social task and its underlying questionnaire items across groups. However, for comparisons of *Skill – Acting in a socially responsible manner/Dealing with conflicts* across sex, SEP, and migration background, the configural model indicated poor model fit. This means that the path models will likely follow a different pattern across the groups. For an overview, see Table 4.

4.4 Permutation tests

The permutation tests resulted predominantly in non-significant p-values, indicating that the significant Chi-square tests of overall model fit depended on the same underlying reasons across groups. However, in the comparison of *Skill – Dealing with differences* across sex and *Attitude – Acting in a socially desirable way* across migration background, the p-value was respectively 0.045 and 0.012. Here, the significant Chi-square tests of overall model fit may depend on different underlying reasons across groups. This means that we need to be precautious when interpreting the results of these two comparisons.

4.5 Exploratory analyses

We established a partial metric model for Skill – Dealing with differences across migration backgrounds. For Attitude – Acting in a socially responsible way across SEP, Knowledge – Dealing with conflicts across migration background, and Knowledge – Dealing with differences across sex, we established partial scalar models. This means that in these models, some indicators were less invariant than others and hindered reaching a higher level of measurement invariance. We performed exploratory analyses to zoom in on these indicators by looking at the content of the questionnaire items, the response frequencies, the factor loadings and the item-rest correlations. In most cases, we found that the difference in factor loadings across groups was the largest for the less invariant indicators.

For example, comparing Attitude – Acting in a socially desirable way across SEP, we established a partial scalar model with freed equality constraints on the second and fourth indicator intercepts. Exploratory analyses of these two indicators demonstrated that the factor loading was higher for students with a high SEP for the second indicator. For the fourth indicator, the opposite was true. This means that the second indicator ('If a classmate is being called names in the streets, I want to stick up for him or her') more strongly relates to the construct Attitude – Acting in a socially desirable way for students with a high SEP as compared to students with a low SEP. Contrary, the fourth indicator ('You should say sorry if you did something that hurt another person') more strongly relates to the construct for students with a low SEP as compared to students with a high SEP.

We also found a difference in factor loadings across groups. For example, comparing Efficacy – Dealing with differences across migration background, we established a partial metric model with freed equality constraints on the factor loading of the first indicator. Exploratory analyses demonstrated that for the first indicator ('How good are you at… adapting your behaviour to the rules and habits of others?'), the factor loading was larger for students without a migration background, whereas for the remaining indicators ('How good are you at… behaving normally in an unfamiliar environment?'; '… adapting your language to the person you are speaking with?'; '… considering the wishes of others when making a decision together?'), the factor loadings were larger for students with a migration background.

The permutation tests resulted predominantly in non-significant p-values, indicating that the significant Chi-square tests of overall model fit depended on the same underlying reasons across groups. However, in the comparison of Skill – Dealing with differences across sex and Attitude – Acting in a socially desirable way across migration background, the p-value was respectively 0.045 and 0.012. Here, the significant Chi-square tests of overall model fit may depend on different underlying reasons across groups. This means that we need to be precautious when interpreting the results of these two comparisons.

5 Discussion

Standardised questionnaires are widely used to measure the outcomes of citizenship education in terms of knowledge, attitude, and skill (Ireland et al., 2006; Schulz et al., 2018). The outcomes of these questionnaires are often compared across groups based on student characteristics, such as boys and girls. Insight into the outcomes of citizenship education and the extent to which these outcomes differ across groups of students is important to evaluate the effectiveness of the delivery methods and content of the curriculum, and ensure that all students benefit optimally. However, a prerequisite for valid cross-group comparisons based on standardised questionnaires entails addressing whether the instrument measures the same across groups and later points in time. This can be done by examining the measurement invariance (Meredith, 1993).

The establishment of measurement invariance is, however, not a given static because the interpretation of constructs can change over time. This may particularly be the case for value-sensitive concepts like citizenship and, accordingly, citizenship education, that is considered a dynamic construct that takes on meaning in context and moves along with changes in society. Hence, it is important to periodically examine the measurement invariance of measurement instruments capturing citizenship competences. In line with this rationale, this study aimed to investigate the measurement invariance of the CCQ (Geijsel et al., 2012; Ten Dam et al., 2011), which is a standardised questionnaire based on a generic conceptualisation of citizenship competences used over a long period. Schools use the instrument to gain insight into students' citizenship knowledge, attitudes, and skills across four social tasks: acting democratically, acting in a socially responsible manner, dealing with conflicts and dealing with differences. To check whether this instrument over time still measures the same across groups in recent samples, we conducted comparisons within three student characteristics that are known to relate to robust differences in citizenship competences: sex, SEP, and migration background (Dijkstra et al., 2015; Geijsel et al., 2012).

Our study showed that in two-thirds of the comparisons across groups, the meaningful minimum of at least (partial) metric invariance (Little, 2013) was reached. This allows for comparisons of the associations between latent constructs and their indicators across groups (i.e., 'the relation between indicator 1 and latent construct X is stronger for boys than for girls'). We even reached (partial) scalar invariance across groups in some comparisons. This allows for comparing latent means across groups (i.e., 'girls, on average, obtain a higher score on latent construct X than boys'). Our findings also mark a warning for researchers conducting multi-group comparative analyses based on citizenship constructs or social tasks where the minimum required level of measurement invariance could not be established. In this study, this applied to one-third of the comparisons, being one citizenship construct (i.e., knowledge) and two social tasks (i.e., Skill – Acting in a socially responsible manner/Dealing with conflicts, and Attitude – Dealing with differences), where fit indices of the configural model indicated overall poor model fit. This means that, for these three aspects, we have no grounds to assume that the instrument measures the same across sex, SEP and migration background, and we need to be cautious when making cross-group comparisons. Nevertheless, the results on these aspects can still be used to describe all students in a class or school as a whole. Moreover, it is advised to assess measurement invariance over these aspects again when using new samples. Whereas we advocate periodic assessment of measurement invariance of *all* citizenship constructs, our findings underscore this is indeed important for future studies examining the construct and social tasks mentioned above.

In some comparisons, specific indicators hindered establishing a higher level of invariance. Therefore, we searched for partial invariance where these particular indicators remained on the lower level of invariance while the other indicators were specified at a higher level of invariance. Exploratory analyses of the partial models pointed to no unidirectional reason why these particular indicators may have been answered differently across groups. That is, we could not identify measurement non-invariance due to either of the three explanations mentioned by Van de Vijver (2013) and Isac et al. (2019): (1) some group members not considering some underlying indicators to be indicative of the construct; (2) cultural or linguistic differences across group members; or (3) a specific response-style of some group members. To improve the questionnaire regardless, we suggest performing additional qualitative data analysis on these non-invariant indicators to see whether this reveals unidirectional leads of how to improve the indicators. This can be done in individual interviews or panel group interviews with students originating from all comparison groups. Alternatively, new indicators can be added to the scales and tested among students to see whether they do show measurement invariance across groups. Eventually, new invariant indicators can replace the non-invariant indicators.

While the number of studies that assessed the measurement invariance of citizenship competences within a country is scarce, we can, to some extent, compare our findings to the studies that performed a between-country assessment of measurement invariance, which is more common in large-scale international assessments of citizenship competences. An example is ICCS (Schulz et al., 2016), where most scales were invariant across countries. The technical report places caution in comparing only two scales (i.e., students' perceptions of the importance of citizenship behaviours; and students' attitudes toward civic institutions and their country of residence). Isac et al. (2019) also used data from ICCS and, similarly, found that most scales measuring students' attitudes towards immigrants were invariant across countries. Only the scale measuring students' attitudes towards equal rights for immigrants was more heterogeneous across countries. The indicator 'Immigrants should have the opportunity to continue their own customs and lifestyle' was the weakest in the scale. In addition, Munck et al. (2018) found that the measurement of students' attitudes toward immigrants was largely invariant over time (from 1999 to 2009) across countries. However, the indicator 'immigrants should have the opportunity to keep their own language' was invariant in only 36 out of 92 countries. Whereas the studies of Isac et al. (2019) and Munck et al. (2018) showed that between-country measurement invariance was more difficult to establish in measuring students' attitude in dealing with differences, our findings of non-invariance in Attitude – Dealing with differences indicate the same in within-country comparisons. However, these previous studies do not comply with the non-invariance that we found for the construct Knowledge and the social task Skill – Acting in a socially responsible manner/Dealing with conflicts.

Although we grounded our findings on extensive analyses, it is important to bear in mind the limitations of this study. Caution needs to be placed on the way we composed the groups. We complied with scholars who underscore that the robustness and precision of measurement invariance analyses increase as sample sizes increase (Koh & Zumbo, 2008; Meade, 2005; Meade & Lautenschlager, 2004). Therefore, we combined data from students of five subsequent years sharing the same student characteristics, and consequently, we were limited in comparing measurement invariance over time. However, the comparison over time is an important lead for future research. We were also somewhat limited in the extent to which we could differentiate within each group. For example, we combined data from students whose parents were born in countries other than the Netherlands into 'having a migration background'. Whereas it is not inconceivable that the effect on measurement invariance differs within these countries, this operationalisation suffices to provide insight into the main effect of migration background in assessing measurement invariance. The same may hold for SEP, where we combined different attained parental educational levels into low SEP.

Despite the limitations, this study contributes to the empirical knowledge base of the assessment of measurement invariance in instruments for citizenship competences. Whereas previous research on the assessment of measurement invariance in instruments for citizenship competences has focused on between-country comparisons (Isac et al., 2019; Schulz et al., 2016), the assessment of measurement invariance is equally important in within-country comparisons (Steinmetz et al., 2009). This study highlights this importance. However, at the same time, the findings of this study make clear that the results of the assessment of measurement invariance are not static. Citizenship and, as a result, citizenship education is dynamic and subject to changes in society and school. These changes may lead to changes in how an instrument aligns with its context and underscores the importance of periodic assessment of measurement invariance.

References

Adelson, J. L. (2012). Examining relationships and effects in gifted education research: An introduction to structural equation modeling. *Gifted Child Quarterly*, *56*(1), 47–55. https://doi.org/10.1177/0016986211424132

Bennett, D. A. (2001). How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health*, *25*(5), 464–469. https://doi.org/10.1111/j.1467-842X.2001.tb00294.x

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Thousand Oaks: Sage.

Byrne, B. M., & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology*, *34*(2), 155–175. https://doi.org/10.1177/0022022102250225

Cleaver, E., Ireland, E., Kerr, D., & Lopes, J. (2006). *Citizenship education longitudinal study second cross-sectional survey, 2004. Listening to young people: Citizenship education in England. Research report RR626*. Nottingham: ERIC Clearinghouse.

Daas, R., Ten Dam, G., & Dijkstra, A. B. (2016). Contemplating modes of assessing citizenship competences. *Studies in Educational Evaluation*, *51*, 88–95. https://doi.org/10.1016/j.stueduc.2016.10.003

Dijkstra, A. B., Geijsel, F., Ledoux, G., Van der Veen, I., & Ten Dam, G. (2015). Effects of school quality, school citizenship policy, and student body composition on the acquisition of citizenship competences in the final year of primary education. *School Effectiveness and School Improvement*, *26*(4), 524–553. https://doi.org/10.1080/09243453.2014.969282

Eidhof, B. B. F., Ten Dam, G., Dijkstra, A. B., & Van de Werfhorst, H. G. (2016). Consensus and contested citizenship education goals in Western Europe. *Education, Citizenship and Social Justice*, *11*(2), 114–129. https://doi.org/10.1177/1746197915626084

Etzioni, A. (1996). *The new Golden Rule. Community and morality in a democratic society.* New York: Basic Books.

Eurostat (2021). *Migrant integration statistics*. Brussels: European Commission.

Geijsel, F., Ledoux, G., Reumerman, R., & Ten Dam, G. (2012). Citizenship in young people's daily lives: Differences in citizenship competences of adolescents in the Netherlands. *Journal of Youth Studies*, *15*(6), 711–729. https://doi.org/10.1080/13676261.2012.671932

He, J., & Van de Vijver, F. J. R. (2013). Methodological issues in cross-cultural studies in educational psychology. In G. A. D. Liem & A. B. I. Bernardo (Eds.), *Advancing cross-cultural perspectives on educational psychology: A festschrift for Dennis McInerney* (pp. 39–56). Charlotte: Information Age Publishing.

Horn, J. L., & Mcardle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3), 117–144. https://doi.org/10.1080/03610739208253916

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Ireland, E., Kerr, D., Lopes, J., Nelson, J., & Cleaver, E. (2006). *Active citizenship and young people: Opportunities, experiences and challenges in and beyond school. Citizenship education longitudinal study: Fourth annual report (DfES Research Report 732)*. Nottingham: Department for Education and Skills.

Isac, M. M., Palmerio, L., & Van der Werf, M. P. C. (2019). Indicators of (in)tolerance toward immigrants among European youth: An assessment of measurement invariance in ICCS 2016. *Large-Scale Assessments in Education*, *7*(1), 1–21. https://doi.org/10.1186/s40536-019-0074-5

Jorgensen, T. D. (2017). Applying permutation tests and multivariate modification indices to configurally invariant models that need respecification. *Frontiers in Psychology*, *8*(1455). https://doi.org/10.3389/fpsyg.2017.01455

Jorgensen, T. D. (2021). *semTools: Useful tools for structural equation modeling*. Retrieved from https://cran.r-project.org/web/packages/semTools/index.html

Kerr, D., Lopes, J., Nelson, J., White, K., Cleaver, E., & Benton, T. (2007). *VISION versus PRAGMATISM: Citizenship in the secondary school curriculum in England : citizenship education longitudinal study : fifth annual report*. Annesley: Department for Education and Skills.

Kite, B. A., Jorgensen, T. D., & Chen, P. (2018). Random permutation testing applied to measurement invariance testing with ordered-categorical indicators. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 573–587. https://doi.org/10.1080/10705511.2017.1421467

Kline, R. B. (2015). *Principles and practices of structural equation modelling*. New York: The Guilford Press.

Koh, K., & Zumbo, B. (2008). Multi-group confirmatory factor analysis for testing measurement invariance in mixed item format data. *Journal of Modern Applied Statistical Methods*, *7*(2). https://doi.org/10.22237/jmasm/1225512660

Little, T. D. (2013). *Longitudinal structural equation modelling*. New York: The Guilford Press.

Mattei, P., & Broeks, M. (2018). From multiculturalism to civic integration: Citizenship education and integration policies in the Netherlands and England since the 2000s. *Ethnicities*, *18*(1), 23–42. https://doi.org/10.1177/1468796816676845

Meade, A. W. (2005). *Sample size and tests of measurement invariance*. Presented at the 20th annual Conference of the Society for Industrial and Organizational Psychology, Los Angeles.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*(4), 361–388. https://doi.org/10.1177/1094428104268027

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

Meuleman, B., Żółtak, T., Pokropek, A., Davidov, E., Muthén, B., Oberski, D. L., … Schmidt, P. (2022). Why measurement invariance is important in comparative research. A response to Welzel et al. (2021). *Sociological Methods & Research*, *0*(0), 1–19. https://doi.org/10.1177/00491241221091755

Munck, I., Barber, C., & Torney-Purta, J. (2018). Measurement invariance in comparing attitudes toward immigrants among youth ccross Europe in 1999 and 2009: The alignment method applied to IEA CIVED and ICCS. *Sociological Methods & Research*, *47*(4), 687–728. https://doi.org/10.1177/0049124117729691

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557–585. https://doi.org/10.1007/BF02296397

Organisation for Economic Co-operation and Development (2017). *Understanding the socioeconomic divide in Europe*. Paris: OECD Publications.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

R Core Team (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Reeskens, T., & Hooghe, M. (2010). Beyond the civic–ethnic dichotomy: Investigating the structure of citizenship concepts across thirty-three countries. *Nations and Nationalism*, *16*(4), 579–597. https://doi.org/10.1111/j.1469-8129.2010.00446.x

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36. https://doi.org/10.18637/jss.v048.i02

Schulz, W., Ainley, J., Fraillon, J., Losito, B., Agrusti, G., & Friedman, T. (2018). *Becoming citizens in a changing world: IEA International Civic and Citizenship Education Study 2016 International Report*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-73963-2

Schulz, W., Carstens, R., Losito, B., & Fraillon, J. (2016). *ICCS 2016 Technical report. IEA International Civic and Citizenship Education Study 2016*. International Association for the Evaluation of Educational Achievement.

Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the model size effect on SEM fit indices. *Educational and Psychological Measurement*, *79*(2), 310–334. https://doi.org/10.1177/0013164418783530

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78–90. https://doi.org/10.1086/209528

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S., & Schwartz, S. H. (2009). Testing measurement invariance using multigroup CFA: Differences between educational groups in human values measurement. *Quality & Quantity*, *43*(4), 599–616. https://doi.org/10.1007/s11135-007-9143-x

Ten Dam, G., Geijsel, F., Reumerman, R., & Ledoux, G. (2011). Measuring young people’s citizenship competences. *European Journal of Education*, *46*(3), 354–372. https://doi.org/10.1111/j.1465-3435.2011.01485.x

US Census Bureau (2020). *Estimates of the components of resident population change by race and Hispanic origin for the United States*. Suitland: US Census Bureau. Retrieved from US Census Bureau website: https://www.census.gov/data/tables/time-series/demo/popest/2010s-national-detail.html

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002

Vertovec, S. (2007). Super-diversity and its implications. *Ethnic and Racial Studies, 30*(6), 1024–1054. https://doi.org/10.1080/01419870701599465

Wattles, J. (1996). *The Golden Rule*. Oxford: Oxford University Press.

Wray-Lake, L., Metzger, A., & Syvertsen, A. K. (2017). Testing multidimensional models of youth civic engagement: Model comparisons, measurement invariance, and age differences. *Applied Developmental Science*, *21*(4), 266–284. https://doi.org/10.1080/10888691.2016.1205495

Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika*, *81*(4), 1014–1045. https://doi.org/10.1007/s11336-016-9506-0

Acknowledgements

We would like to express our gratitude towards dr. Terrence D. Jorgensen for his consult during the analyses of this research, and to dr. Femke Geijsel for her kind responses to our request for additional information on the psychometrics of the CCQ.

Role of the funding source

This research was funded by the Netherlands Organisation for Scientific Research (NWO), Initiative for Education Research (NRO, 40.5.18540.099) and the Academische Werkplaats Sociale Kwaliteit Onderwijs. The funding source had no involvement in the conduct of the research, nor in the preparation of this article.

Author biography / Biographies

Lianne Hoek is a PhD Candidate at the Research Institute of Child Development and Education at the University of Amsterdam, The Netherlands. Her research focuses on the measurement of social outcomes of education and output-driven citizenship education. E-mail: l.h.m.hoek@uva.nl.

Bonne Zijlstra is Assistant Professor at the Research Institute of Child Development and Education at the University of Amsterdam, the Netherlands. His research focuses on the statistical modelling of dependent data. E-mail: b.j.h.zijlstra@uva.nl.

Anke Munniksma is Assistant Professor at the Research Institute of Child Development and Education at the University of Amsterdam, The Netherlands. Her research focuses on the social development of youth in ethnically diverse societies. E-mail: a.munniksma@uva.nl.

Anne Bert Dijkstra is professor of Supervision and Effects of Socialisation in Education at the University of Amsterdam. His research focuses on school effectiveness and social outcomes of education. He also is Program Director of the Education and Social Cohesion Program at the Inspectorate of Education in The Netherlands. E-mail: a.b.dijkstra@uva.nl.

6 APPENDIX/SUPPLEMENTARY MATERIALS

Attachment 1: *Model fit of configural, threshold, metric, and scalar model across sex*

		χ^2	df	P	CFI	RMSEA	CI RMSEA	Anova	$\Delta \chi^2$	P
Attitude										
	Configural	296.642	10	0.000***	0.964	0.100	[0.091; 0.110]	-	-	
	Metric	304.165	14	0.000***	0.964	0.085	[0.077; 0.094]	Configural	7.5236	0.1107
	Scalar	567.460	18	0.000***	0.932	0.104	[0.096; 0.111]	Metric	263.30	<0.001***
Attitude – Acting in a socially desirable way										
	Configural	200.257	18	0.000***	0.964	0.060	[0.053; 0.067]	-	-	-
	Threshold	220.041	24	0.000***	0.961	0.054	[0.047; 0.060]	Configural	2.2175	0.8986
	Metric	210.176	29	0.000***	0.964	0.047	[0.041; 0.053]	Threshold	6.237	0.2838
	Scalar	268.797	34	0.000***	0.954	0.049	[0.044; 0.055]	Metric	45.632	<0.001***
Attitude – Dealing with conflicts										
	Configural	164.454	18	0.000***	0.993	0.054	[0.046; 0.061]	-	-	-
	Threshold	175.315	24	0.000***	0.993	0.047	[0.041; 0.054]	Configural	4.3649	0.6274
	Metric	175.180	29	0.000***	0.993	0.042	[0.036; 0.048]	Threshold	12.401	0.02968*
	Scalar	183.975	34	0.000***	0.993	0.039	[0.034; 0.045]	Metric	12.074	0.03379*
Attitude – Dealing with differences										
	Configural	1132.638	18	0.000***	0.962	0.148	[0.141; 0.155]	-	-	-
	Threshold	1211.898	24	0.000***	0.959	0.132	[0.126; 0.139]	Configural	9.8652	0.1304
	Metric	1105.878	29	0.000***	0.963	0.115	[0.109; 0.120]	Threshold	4.1677	0.5255
	Scalar	1119.360	34	0.000***	0.963	0.106	[0.101; 0.112]	Metric	37.56	<0.001***
Knowledge										
	Configural	140.400	4	0.000***	0.971	0.110	[0.094; 0.126]	-	-	-
	Metric	148.345	7	0.000***	0.970	0.084	[0.073; 0.096]	Configural	7.9455	0.04715*
	Scalar	302.460	10	0.000***	0.937	0.102	[0.092; 0.112]	Metric	154.11	<0.001***
Knowledge – Acting democratically										
	Configural	103.537	40	0.000***	0.981	0.024	[0.018; 0.029]	-	-	-
	Scalar	113.542	46	0.000***	0.980	0.023	[0.18; 0.28]	Configural	9.7076	0.1375

Knowledge – Acting in a socially desirable way										
	Configural	21.512	18	0.254	0.997	0.008	[0.000; 0.020]	-	-	-
	Scalar	26.429	22	0.234	0.997	0.008	[0.000; 0.019]	Configural	4.6404	0.32622
Knowledge – Dealing with conflicts										
	Configural	226.743	28	0.000***	0.943	0.050	[0.044; 0.056]	-	-	-
	Scalar	232.386	33	0.000***	0.943	0.046	[0.041; 0.052]	Configural	13.260	0.02106*
Knowledge – Dealing with differences										
	Configural	99.869	18	0.000***	0.955	0.040	[0.033; 0.048]	-	-	-
	Scalar	115.965	22	0.000***	0.948	0.039	[0.032; 0.046]	Configural	13.0913	0.01084*
	Partial Scalar	105.885	20	0.000***	0.953	0.039	[0.032; 0.046]	Configural	4.9802	0.0829
Skill										
	Configural	99.441	4	0.000***	0.987	0.092	[0.077; 0.108]	-	-	-
	Metric	101.927	7	0.000***	0.987	0.069	[0.058; 0.081]	Configural	2.486	0.4778
	Scalar	247.994	10	0.000***	0.967	0.092	[0.082; 0.102]	Metric	146.07	<0.001***
	Partial Scalar	102.50	9	0.000***	0.987	0.060	[0.050; 0.071]	Metric	0.56847	0.7526
Skill – Acting in a socially desirable way/Dealing with conflicts										
	Configural	413.516	10	0.000***	0.975	0.120	[0.110; 0.130]	-	-	-
	Threshold	434.607	15	0.000***	0.974	0.100	[0.092; 0.108]	Configural	6.2712	0.2807
	Metric	384.895	19	0.000***	0.978	0.083	[0.076; 0.090]	Threshold	1.9376	0.7472
	Scalar	474.821	23	0.000***	0.972	0.083	[0.077; 0.090]	Metric	69.79	<0.001***
Skill – Dealing with differences										
	Configural	33.351	4	0.000***	0.996	0.051	[0.036; 0.068]	-	-	-
	Threshold	38.813	8	0.000***	0.995	0.037	[0.026; 0.049]	Configural	3.3767	0.4969
	Metric	43.312	12	0.000***	0.995	0.032	[0.022; 0.043]	Threshold	5.0567	0.1677
	Scalar	44.883	14	0.000***	0.995	0.028	[0.019; 0.037]	Metric	2.591	0.4591

Attachment 2: *Model fit of configural, threshold, metric, and scalar model across socioeconomic position*

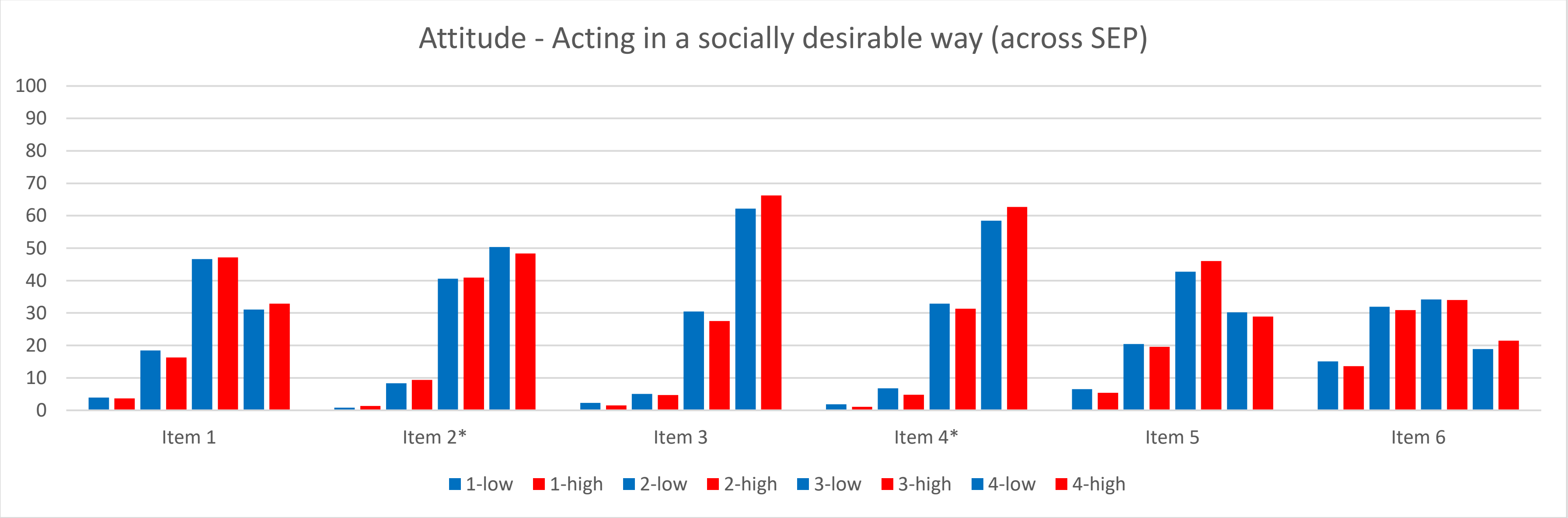
		χ^2	df	P	CFI	RMSEA	CI RMSEA	Anova	$\Delta\ \chi^2$	P
Attitude										
	Configural	195.933	10	0.000***	0.968	0.096	[0.084; 0.108]	-	-	-
	Metric	198.077	14	0.000***	0.968	0.081	[0.071; 0.091]	Configural	2.1444	0.7092
	Scalar	225.477	18	0.000***	0.964	0.075	[0.067; 0.084]	Metric	27.399	<0.001***
	Partial Scalar	198.083	15	0.000***	0.968	0.078	[0.068; 0.087]	Metric	0.006	0.9381
Attitude – Acting in a socially responsible manner										
	Configural	164.591	18	0.000***	0.962	0.063	[0.055; 0.073]	-	-	-
	Threshold	182.332	24	0.000***	0.959	0.057	[0.050; 0.065]	Configural	3.0837	0.7983
	Metric	194.089	29	0.000***	0.958	0.053	[0.046; 0.060]	Threshold	16.238	0.006197**
	Partial Metric	173.415	27	0.000***	0.963	0.052	[0.045; 0.059]	Threshold	1.7855	0.6181
	Partial Scalar	180.591	32	0.000***	0.962	0.048	[0.041; 0.055]	Partial Metric	8.5067	0.1304
Attitude – Dealing with conflicts										
	Configural	114.129	18	0.000***	0.993	0.051	[0.043; 0.061]	-	-	-
	Threshold	128.988	24	0.000***	0.993	0.047	[0.39; 0.055]	Configural	9.3279	0.156
	Metric	112.524	29	0.000***	0.994	0.038	[0.031; 0.045]	Threshold	1.0021	0.9624
	Scalar	111.648	34	0.000***	0.995	0.034	[0.027; 0.041]	Metric	3.5128	0.6215
Attitude – Dealing with differences										
	Configural	740.961	18	0.000***	0.964	0.141	[0.133; 0.150]	-	-	-
	Threshold	788.920	24	0.000***	0.962	0.126	[0.118; 0.133]	Configural	2.3189	0.8882
	Metric	731.332	29	0.000***	0.965	0.110	[0.103; 0.117]	Threshold	8.1128	0.1501
	Scalar	719.695	34	0.000***	0.966	0.100	[0.094; 0.106]	Metric	12.266	0.03132*
Knowledge										
	Configural	109.354	4	0.000***	0.970	0.114	[0.096; 0.133]	-	-	-
	Metric	110.875	7	0.000***	0.970	0.086	[0.072; 0.100]	Configural	1.5212	0.6774

	Scalar	146.012	10	0.000***	0.961	0.082	[0.070; 0.094]	Metric	35.137	<0.001***
Knowledge – Acting democratically										
	Configural	68.933	40	0.003**	0.987	0.019	[0.011; 0.026]	-	-	-
	Scalar	80.816	46	0.001**	0.984	0.019	[0.012; 0.026]	Configural	10.189	0.1169
Knowledge – Acting in a socially responsible manner										
	Configural	15.598	18	0.621	1.000	0.000	[0.000; 0.017]	-	-	-
	Scalar	19.792	22	0.596	1.000	0.000	[0.000; 0.016]	Configural	4.0691	0.3967
Knowledge – Dealing with conflicts										
	Configural	187.003	28	0.000***	0.947	0.053	[0.046; 0.060]	-	-	-
	Scalar	180.329	33	0.000***	0.951	0.047	[0.040; 0.054]	Configural	1.6152	0.8994
Knowledge – Dealing with differences										
	Configural	72.946	18	0.000***	0.958	0.039	[0.030; 0.048]	-	-	-
	Scalar	72.592	22	0.000***	0.961	0.034	[0.025; 0.043]	Configural	3.3881	0.4951
Skill										
	Configural	63.587	4	0.000***	0.988	0.086	[0.068; 0.105]	-	-	-
	Metric	65.780	7	0.000***	0.988	0.064	[0.051; 0.079]	Configural	2.1931	0.5333
	Scalar	80.967	10	0.000***	0.986	0.059	[0.048; 0.071]	Metric	15.187	0.001664**
	Partial Scalar	67.729	9	0.000***	0.989	0.057	[0.045; 0.070]	Metric	1.949	0.3774
Skill – Acting in a socially responsible manner/Dealing with conflicts										
	Configural	263.175	10	0.000***	0.979	0.112	[0.101; 0.124]	-	-	-
	Threshold	274.545	15	0.000***	0.978	0.093	[0.083; 0.102]	Configural	2.1858	0.8229
	Metric	255.037	19	0.000***	0.980	0.079	[0.070; 0.087]	Threshold	7.6401	0.1057
	Scalar	254.560	23	0.000***	0.981	0.071	[0.063; 0.079]	Metric	4.1723	0.3832
Skill – Dealing with differences										
	Configural	19.616	4	0.001**	0.997	0.044	[0.026; 0.064]	-	-	-
	Threshold	22.100	8	0.005**	0.997	0.030	[0.015; 0.045]	Configural	1.5899	0.8106
	Metric	25.492	11	0.008**	0.997	0.026	[0.012; 0.039]	Threshold	3.8552	0.2775
	Scalar	36.466	14	0.001**	0.996	0.028	[0.017; 0.040]	Metric	8.6192	0.03481*

Attachment 3: *Model fit of configural, threshold, metric, and scalar model across migration background*

		χ^2	df	P	CFI	RMSEA	CI RMSEA	Anova	$\Delta \chi^2$	P
Attitude										
	Configural	276.123	10	0.000***	0.968	0.097	[0.087; 0.107]	-	-	-
	Metric	288.138	14	0.000***	0.967	0.083	[0.075; 0.091]	Configural	12.015	0.01724*
	Scalar	554.333	18	0.000***	0.936	0.102	[0.095; 0.110]	Metric	266.19	<0.001***
Attitude – Acting in a socially responsible manner										
	Configural	210.179	18	0.000***	0.968	0.061	[0.054; 0.069]	-	-	-
	Threshold	230.528	24	0.000***	0.965	0.055	[0.049; 0.062]	Configural	4.506	0.6085
	Metric	224.131	29	0.000***	0.967	0.049	[0.043; 0.055]	Threshold	9.2647	0.09896
	Scalar	457.328	34	0.000***	0.928	0.066	[0.061; 0.072]	Metric	103.48	<0.001***
Attitude – Dealing with conflicts										
	Configural	172.091	18	0.000***	0.993	0.055	[0.048; 0.063]	-	-	-
	Threshold	181.417	24	0.000***	0.993	0.048	[0.042; 0.055]	Configural	5.6431	0.4643
	Metric	164.825	29	0.000***	0.994	0.041	[0.035; 0.047]	Threshold	5.4018	0.3688
	Scalar	185.670	34	0.000***	0.993	0.040	[0.034; 0.045]	Metric	14.326	0.01366*
	Partial Scalar	170.475	33	0.000***	0.994	0.038	[0.033; 0.044]	Metric	7.4826	0.1125
Attitude – Dealing with differences										
	Configural	1196.779	18	0.000***	0.958	0.152	[0.145; 0.160]	-	-	-
	Threshold	1260.470	24	0.000***	0.955	0.135	[0.129; 0.141]	Configural	17.618	0.007262**
	Metric	1118.213	29	0.000***	0.961	0.115	[0.110; 0.121]	Threshold	8.3448	0.1382
	Scalar	1104.191	34	0.000***	0.961	0.105	[0.100; 0.111]	Metric	31.088	<0.001***
Knowledge										
	Configural	138.223	4	0.000***	0.972	0.109	[0.094; 0.125]	-	-	-
	Metric	145.666	7	0.000***	0.971	0.083	[0.072; 0.096]	Configural	7.4431	0.05904
	Scalar	176.543	10	0.000***	0.965	0.077	[0.067; 0.087]	Metric	30.878	<0.001***

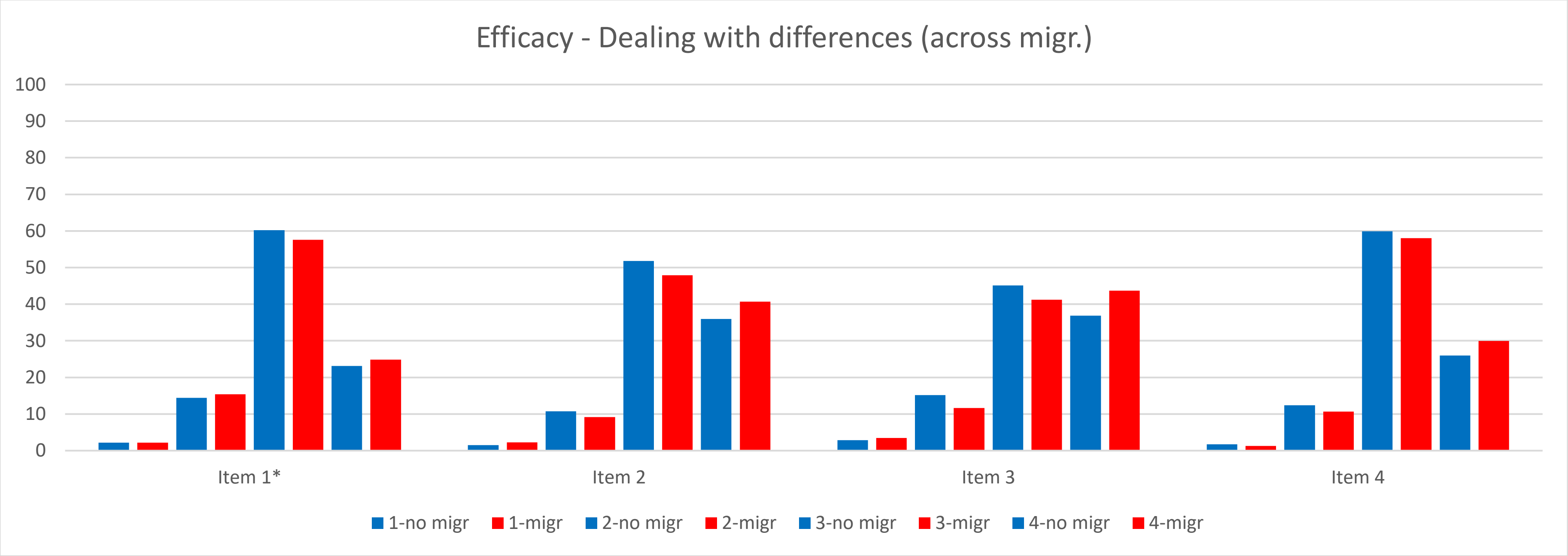
Attachment 4: *Exploratory analyses of Attitude – Acting in a socially desirable way (across SEP)*



Attitude – Acting in a socially desirable way (across SEP)		Low SEP			High SEP			Difference Δ		
		μ	λ	Rir	μ	λ	Rir	Δμ	Δλ	ΔRir
1	Wealthy people taking care of less wealthy people	3.05	0.486	0.36	3.09	0.514	0.36	0.04	0.028	0.00
2*	Standing up for a classmate when s/he is being called names	3.41	0.401	0.29	3.36	0.494	0.35	-0.05	0.093	0.06
3	Cleaning up after having a picknick in the parc	3.53	0.698	0.43	3.59	0.660	0.39	0.06	-0.038	-0.04
4*	Saying sorry if you hurt somebody	3.48	0.757	0.48	3.56	0.656	0.40	0.08	-0.101	-0.08
5	Helping in the household	2.97	0.454	0.33	2.99	0.448	0.31	0.02	-0.006	-0.02
6	Visiting a classmate who has been sick for a long time	2.57	0.487	0.36	2.63	0.471	0.34	0.06	-0.016	-0.02

Note. Items were originally formulated in Dutch. A description of the item in English is provided.

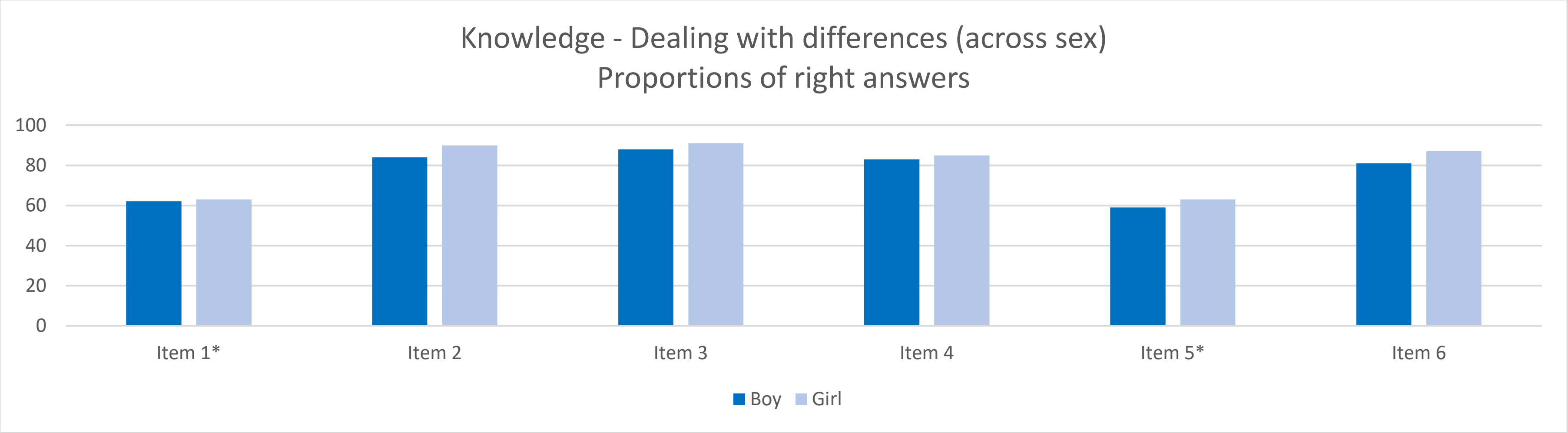
Attachment 5: *Exploratory analyses of Efficacy – Dealing with differences (across migration background)*



Efficacy – Dealing with differences (across migration background)		No migr.			Migr.			Difference Δ		
		μ	λ	Rir	μ	λ	Rir	Δμ	Δλ	ΔRir
1*	Adapting to other people's rules and habits	3.04	0.705	0.49	3.05	0.672	0.50	0.01	-0.033	0.01
2	Behaving normal in an unknown environment	3.22	0.675	0.48	3.27	0.708	0.53	0.05	0.033	0.05
3	Adapting your language to the one who you speak with	3.16	0.625	0.46	3.25	0.698	0.51	0.09	0.073	0.05
4	Taking into account wishes of others when making a decision	3.10	0.669	0.48	3.17	0.751	0.55	0.07	0.082	0.07

Note. Items were originally formulated in Dutch. A description of the item in English is provided.

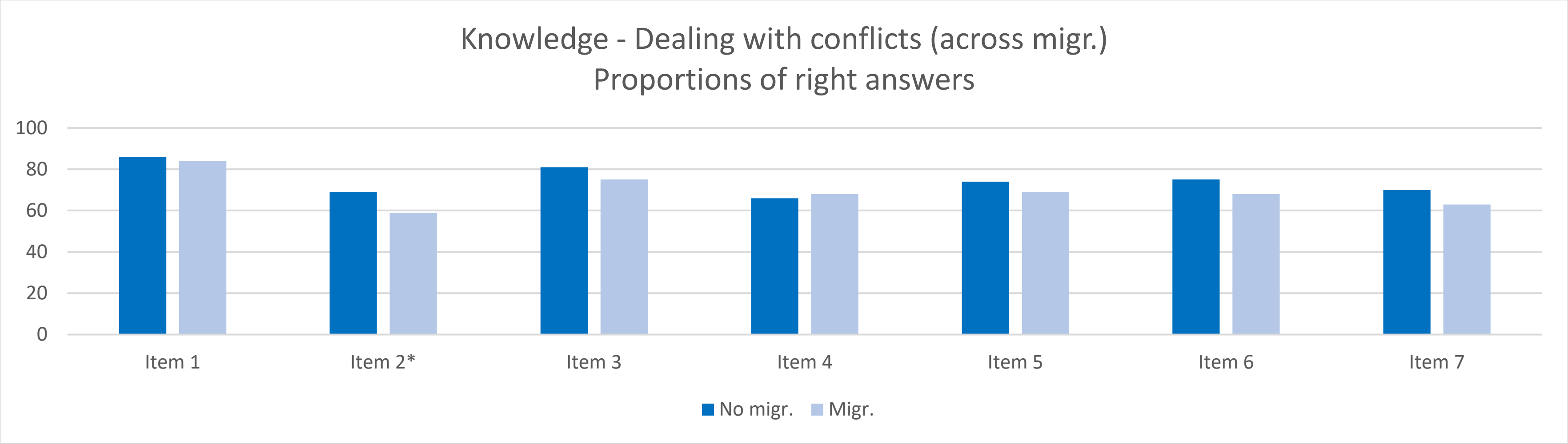
Attachment 6: *Exploratory analyses of Knowledge – Dealing with differences (across sex)*



Knowledge – Dealing with differences (across sex)		Boys			Girls			Difference Δ		
		μ	λ	Rir	μ	λ	Rir	Δμ	Δλ	ΔRir
1*	Recognising a prejudice in three examples	0.62	0.355	0.20	0.63	0.505	0.25	0.01	0.150	0.05
2	Understanding why the teacher places a classmate who behaves busy outside the classroom	0.84	0.655	0.30	0.90	0.528	0.22	0.06	-0.127	-0.08
3	Recognising discrimination in three examples	0.88	0.770	0.34	0.91	0.747	0.30	0.03	-0.023	-0.04
4	Picking the correct fact between three statements about Islam, Hinduism and Catholicism	0.83	0.427	0.21	0.85	0.341	0.15	0.02	-0.086	-0.06
5*	Choosing whether a statement is discrimination, a prejudice or a fact	0.59	0.487	0.28	0.63	0.498	0.25	0.04	0.011	-0.03
6	Deciding whether or not to take off your shoes when visiting someone where this is a habit	0.81	0.555	0.27	0.87	0.474	0.20	0.06	-0.081	-0.07

Note. Items were originally formulated in Dutch. A description of the item in English is provided.

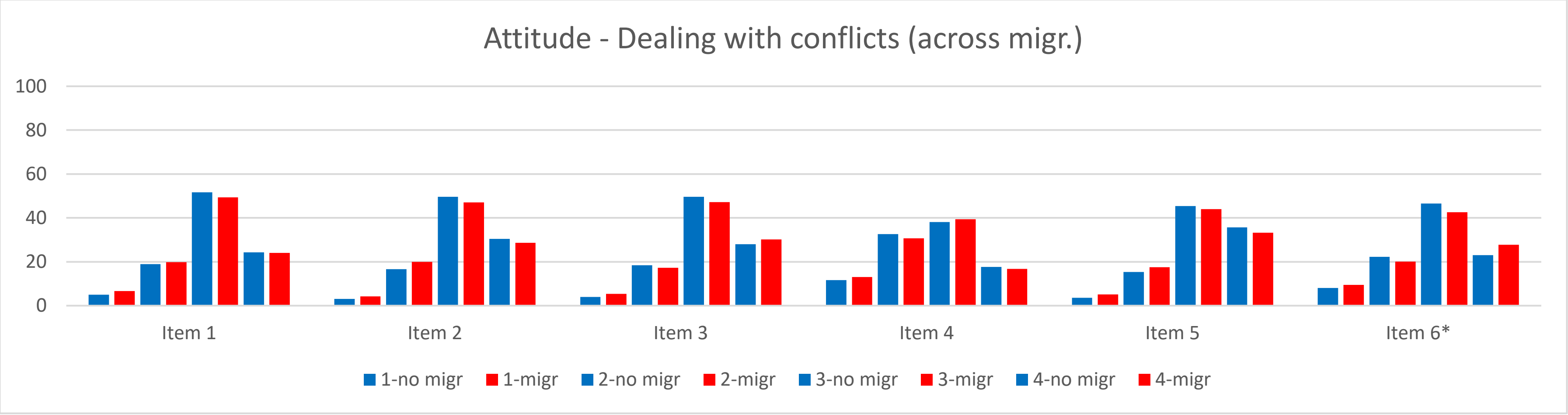
Attachment 7: *Exploratory analyses of Knowledge – Dealing with conflicts (across migration background)*



Knowledge – Dealing with conflicts (across migration background)		No migr.			Migr.			Difference Δ		
		μ	λ	Rir	μ	λ	Rir	Δμ	Δλ	ΔRir
1	Making a decision with two friends when all three want something different	0.86	0.651	0.35	0.84	0.685	0.37	-0.02	0.034	0.02
2*	Deciding what you can do best when your friend is in a fight that does not seem to stop	0.69	0.609	0.32	0.59	0.543	0.30	-0.10	-0.063	-0.02
3	Deciding what you can do best when you had a fight but afterwards, it appeared that you were wrong	0.81	0.677	0.37	0.75	0.713	0.41	-0.06	0.036	0.03
4	Deciding what you can do best when your friends bully your friendly neighbour.	0.66	0.418	0.26	0.68	0.464	0.29	0.02	0.046	0.03
5	Deciding what you can do best when you are in a fight with a friend that only seems to become worse	0.74	0.233	0.14	0.69	0.313	0.19	-0.05	0.080	0.05
6	Deciding what you can do best when your sister frequently skips the household chores that you need to take care of together	0.75	0.627	0.35	0.68	0.593	0.35	-0.07	-0.034	0.00
7	Deciding what you can do best when most classmates want to play a game but not everyone agrees	0.70	0.633	0.37	0.63	0.627	0.38	-0.07	-0.006	0.01

Note. Items were originally formulated in Dutch. A description of the item in English is provided.

Attachment 8: *Exploratory analyses of Attitude – Dealing with conflicts (across migration background)*



Attitude – Dealing with conflicts (across migration		No migr.			Migr.			Difference Δ		
background)		μ	λ	Rir	μ	λ	Rir	Δμ	Δλ	ΔRir
1	Taking into account the other when being into a fight	2.95	0.815	0.62	2.91	0.763	0.67	-0.04	-0.052	0.05
2	Still seeing the other in a normal way, even after disagreeing in a fight	3.08	0.735	0.59	3.00	0.727	0.60	-0.08	-0.008	0.01
3	Trying to take the other serious when being into a fight	3.02	0.729	0.60	3.02	0.738	0.62	0.00	0.009	0.02
4	Searching for points of agreement and disagreement when being into a fight	2.62	0.703	0.55	2.60	0.662	0.59	-0.02	-0.041	0.04
5	Getting your way when being into a fight	3.13	0.757	0.63	3.05	0.766	0.63	-0.08	0.009	0.00
6*	Willing to search a compromise when being into a fight	2.85	0.592	0.47	2.89	0.561	0.50	0.04	-0.031	0.03

Note. Items were originally formulated in Dutch. A description of the item in English is provided.

Attachment 9: *Results of permutation tests*

		Omnibus p value based on parametric chi-squared difference test			Omnibus p values based on nonparametric permutation method	
Group	Social tasks	Chisq diff	Df diff	Pr(>Chisq)	AFI. Difference	p.value
Sex	Attitude – acting in a socially desirable way	200.257	18.000	0.000	137.802	0.053
	Attitude – dealing with conflicts	164.454	18.000	0.000	80.876	0.27
	Attitude – dealing with differences	1132.638	18.000	0.000	588.838	0.998
	Efficacy – acting in a socially desirable way/dealing with conflicts	413.516	10.000	0.000	216.566	1
	Efficacy – dealing with differences	33.351	4.000	0.000	18.636	0.045
	Knowledge – acting democratically	103.537	40.000	0.000	78.725	0.561
	Knowledge – acting in a socially desirable way	21.512	18.000	0.254	16.574	0.513
	Knowledge – dealing with conflicts	226.743	28.000	0.000	163.157	0.251
	Knowledge – dealing with differences	99.869	18.000	0.000	76.279	0.816
SEP	Attitude – acting in a socially desirable way	164.591	18.000	0.000	112.273	0.948
	Attitude – dealing with conflicts	114.129	18.000	0.000	57.182	0.149
	Attitude – dealing with differences	740.961	18.000	0.000	386.515	0.761
	Efficacy – acting in a socially desirable way/dealing with conflicts	263.175	10.000	0.000	136	0.529
	Efficacy – dealing with differences	19.616	4.000	0.001	10.999	0.143
	Knowledge – acting democratically	68.933	40.000	0.003	51.633	0.994
	Knowledge – acting in a socially desirable way	15.598	18.000	0.621	11.526	0.693
	Knowledge – dealing with conflicts	218.543	28.000	0.000	154.525	0.96
	Knowledge – dealing with differences	103.432	18.000	0.000	79.226	0.374
Migr.	Attitude – acting in a socially desirable way	210.179	18.000	0.000	141.395	0.012
	Attitude – dealing with conflicts	172.091	18.000	0.000	83.047	0.184
	Attitude – dealing with differences	1196.779	18.000	0.000	599.084	0.2920
	Efficacy – acting in a socially desirable way/dealing with conflicts	418.726	10.000	0.000	220.929	0.743
	Efficacy – dealing with differences	25.207	4.000	0.000	13.297	0.911

Knowledge – acting democratically	104.332	40.000	0.000	78.414	0.598
Knowledge – acting in a socially desirable way	19.809	18.000	0.344	15.149	0.632
Knowledge – dealing with conflicts	218.543	28.000	0.000	154.525	0.948
Knowledge – dealing with differences	103.432	18.000	0.000	79.226	0.38