Do Culture and Reading Literacy Associate with Inconsistent Responding on Mixed-Worded Scales?

Fatemeh Montazerikafrani

Supervisors: Dr. Isa Steinmann, Prof. Dr. Johan Braeken

Master of Assessment, Measurement and Evaluation

120 credits

University of Oslo

Centre for Educational Measurement (CEMO)

Fall 2021

Popular Abstract

Questionnaire scales often include a combination of both positively and negatively worded items such as “I make friends easily at school” and “I feel lonely at school”. In this case, a respondent with a high sense of belonging to school is expected to agree with the positively worded item and disagree with the negatively worded item. However, studies show that some of the respondents fail to switch the side of their responses according to item wording, a behavior we call inconsistent responding. The proportion of inconsistent respondents varies between countries in international assessments. This study shows that on a scale from PISA 2018 across 75 countries, between 4% to 30% of respondents show this behavior. By looking at two variables for culture and one variable for reading literacy, I expected to find significant associations between all of the variables and inconsistency. But I found that when having them all in the same model, only reading is significantly different from zero. I conclude that researchers should be cautious when using these types of scales in international assessments, especially when there are countries with low reading achievement levels.

Acknowledgments

I would first like to thank my thesis supervisors Dr. Isa Steinmann of the Centre for Educational Measurement at the University of Oslo, who was always there with fresh insight whenever I needed it, and Prof. Dr. Johan Braeken of the Centre for Educational Measurement at the University of Oslo, who influenced practicality and clarity throughout the project.

I would also like to acknowledge my deepest gratitude to my parents, sisters and brother in law, who provided me with all the support I needed and more throughout my study period.

Abstract

Mixed-worded scales which are sometimes used in questionnaires have shown to have unintended consequences caused by inconsistent responding behavior. The proportion of respondents with this behavior varies between countries, and the present study aims to investigate whether cultural differences and reading literacy are associated with the shares of inconsistent respondents in cross-national surveys. I expected to find larger shares of inconsistency in individualist and culturally loose countries and countries with lower mean reading scores. I used a constrained factor mixture analysis model to identify inconsistent respondents on a mixed worded scale from PISA 2018 in 75 countries. I found proportions of between 4% to 30% of inconsistent respondents. Findings of country-level regression analyses suggest that it is mostly the mean reading literacy levels that are strongly associated with the inconsistency rate, not the cultural differences. The study discusses implications for the use of mixed-worded scales in international assessments when reading literacy varies largely among respondents.

Keywords: individualist/collectivist, tight/loose, reading proficiency, cross-cultural differences, inconsistent respondents, mixed-worded scales

Do Culture and Reading Literacy Associate with Inconsistent Responding on Mixed-Worded Scales?

The use of mixed-worded items has become increasingly popular among scale developers. A mixed-worded scale is a Likert scale in a questionnaire (i.e., when the construct to be measured is a typical non-cognitive performance formatted as statements (Chyung, Barkin & Shamsy, 2018), where there are no correct or incorrect responses) that consists of both positively (e.g., I make friends easily at school) and negatively (e.g., I feel lonely at school) worded items. The response categories are the same for both item types (e.g., from strongly agree to strongly disagree) and the construct being measured by both sets of items is the same.

Reverse-worded items are mixed with other items in the scale to avoid boredom and inattentiveness that leads to not thoroughly reading all items and giving the same response to all items if all of them are worded in the same direction. These scales are also thought to control for biases such as acquiescence, which is the tendency to agree and give positive responses. These behaviors create a systematic variance in the data that is irrelevant to the construct, so using a mixed-worded scale is suggested to reduce this (Chyung et al., 2018; DiStefano & Motl, 2006).

However, the use of these scales has been criticized for having unintended consequences (van Sonderen, Sanderman & Coyne, 2013). Correlations between the oppositely worded items are observed to be closer to zero than correlations between items with the same wording (Marsh, 1986); whereas theoretically there is no reason to expect that they would not be of equal levels. As shown in several studies, there has been traces of multi-dimensionality in the use of mixed-worded scales, where the negatively worded items seem to capture a systematic variance beyond the general factor (Gnambs & Schroeders, 2020). This makes the one-dimensional CFA

models to have poor fit to the data and show unexpected item intercorrelation patterns, which improves significantly after accounting for the uniqueness of the negatively worded items (Cheng & Hamid, 1997; Jiang, Fang, Stith, Liu & Huebner, 2018; Marsh, 1996; Schmitt & Stuits, 1985; Wong, Rindfleisch & Burroughs, 2003; Zhang, Noor & Savalei, 2016). It suggests that the idea of balancing the responses by canceling out unintended systematic responses with an equal number of positive and negative items is ineffective (Gnambs & Schroeders, 2020; Pilotte & Gable, 1990; Zhang et al., 2016). Steinmann, Strietholt & Braeken (2021) argue that this phenomenon is caused by an interaction of both the respondents' characteristics and the instruments' properties. This study is centered on person characteristics, i.e., noticing the change in item wording and managing to adjust the responses in mixed-worded scales that demands a certain level of cognitive ability, or the willingness to put in the effort to read every item carefully enough to notice the difference that requires a certain level of commitment from the respondents.

Response behavior reflects systematic tendencies of responding to questionnaire items in ways that are unrelated to what they are intended to measure (Paulhus, 1991). It seems to bring about a new kind of "response bias", something that was originally intended to be avoided by using mixed-worded items (Chyung et al., 2018). Response behavior is one of several factors that threaten the validity of the instrument because they do not reflect the respondents' actual opinions, feelings, or perspectives about the underlying construct (Kemmelmeier, 2016; van Sonderen et al., 2013). Different reasons lead to carelessness or cognitive difficulties that can influence response behavior, and it can reveal itself in many forms. The response behavior under this study is "inconsistent response behavior". When there are items worded in positive and

negative directions that target the same construct and have the same response scale, one expects that a respondent who agrees with the positive items would disagree with the negative items, and vice versa. This is what is called “consistent responses”. Inconsistent responses are then when a respondent agrees with both sets of items, or disagrees with both; also referred to as “misresponse” (Swain, Weathers & Niedrich, 2008). Schmitt & Stuits (1985) emphasize that this type of carelessness is not about giving responses randomly, but a systematic way of reading through some items and assuming the rest of the items are the same. They showed in a simulation study that when as low as 10% of the respondents express this behavior, a clear negative factor is generated and the scale is no longer unidimensional, hence the aforementioned threats to the factor structure and validity of the scale appears.

This problem seems more pronounced in some countries than in others. When it comes to promoting global research on people living in different societies, researchers need to ensure their measures are valid and appropriate across all countries and cultures, and failing to do so results in confounded cross-cultural applicability of mixed-worded scales (Cheng and Hamid, 1997; Jiang et al., 2018; Wong et al., 2003). These studies investigated the issue that some domestically developed mixed-worded scales in the western culture don’t work well when used in a different cultural setting like East Asia. The reason behind this is not fully clear yet, but they demonstrated that the respondents might see and interpret the reversed-worded items differently and actually treat them as measures of a separate construct. This might be at least partly related to the differences in culture.

Possible cultural and cognitive factors associated with between-country differences in this phenomenon

This study is about exploring some of the possible influencing factors for why these scales appear to work well in some countries while having questionable validity and reliability in cross-cultural comparisons (Cheng & Hamid, 1997; Jiang et al., 2018; Wong et al., 2003). I want to see if at least a part of the variation in the shares of inconsistent respondents between countries can be associated with cultural or reading ability differences. Possible influencers for this phenomenon have been studied before, such as language differences with regard to polarity (i.e., whether a statement is an affirmation or a negation) or truth value (i.e., whether a statement is true or false for the respondent), and item verification difficulty (i.e., reverse worded items can be more confusing or difficult to read and comprehend) (Swain et al., 2008). Kemmelmeier (2016) studied the cultural differences in three different types of survey responding biases: acquiescent (i.e., tendency to give positive responses), extreme (i.e., only choosing the two ends of the Likert scale and not in between) and socially desirable (i.e., responding according to what they think the society would like and accept). But there has not been a study so far that investigates the association between cultural factors and inconsistent response behavior on mixed-worded scales.

As a representative multi-national sample, I carried out this research on participants of an international, large-scale and low-stakes assessment which are 15 year-old high school students from more than 70 different countries. In order to inscribe the potential differences between various cultural groups as dependent variables in a comprehensive and simple manner, cultures can be differentiated following different taxonomy. In dimensional cultural theories, the

influence of sociocultural context on a member of a certain cultural group is quantified by calculating an average score of particular cultural dimensions and then the culture's characteristics can be compared (Čeněk, 2020). Two that appear relevant to understanding differences in response behavior are the individualism versus collectivism theory, and tightness versus looseness theory.

Hofstede's cultural dimensions theory

People have different patterned ways of thinking, feeling and acting, who face the same challenges, problems or situations in life. Understanding the structure of these differences is important for bringing about a mutual understanding worldwide. A cultural dimension is an aspect of a culture that can be measured relative to other cultures (Hofstede, Hofstede & Minkov, 2010).

In 1954, the American sociologist Alex Inkeles and the American psychologist Daniel Levinson suggested that individuals' relation to authority, their self-concept in relation to other individuals and society, the concept of masculinity/femininity, and ways of dealing with conflicts can be four categories of "problems" that would have different "solutions" in different cultures. Later on, the Dutch social psychologist and professor of organizational anthropology at Maastricht University, Gerard Hendrik (Geert) Hofstede studied a large body of data from a large multinational corporation; International Business Machine (IBM) which he also was an employee of. The data was collected between 1967 and 1973 in two survey rounds, with more than 116,000 questionnaires, each with about 150 questions, from 72 countries in 20 languages (Hofstede, 1983, 2001). The study was about country differences in answering questions about

employee values and work goal importance. A country-level factor analysis was used on data from comparable employee samples across countries (Hofstede, 2001). The participants were employees who were similar in all aspects except nationality; which made them good representatives for a cultural study. The statistical analysis supported the predicted areas by Inkeles and Levinson which were then named "cultural dimensions". *Power distance*, which reflects the answers to questions about how the fact that people are unequal is handled in a society, *individualism versus collectivism*, which refers to the role and the power of the group versus the role and the power of individuals in different societies, *masculinity versus femininity*, which is about which behavior is considered masculine or feminine in different societies, and *uncertainty avoidance*, which is about differences in ways of handling uncertain and ambiguous situations. All four concepts already existed in social sciences (Hofstede et al., 2010).

Hofstede and other people replicated the IBM study on several occasions over the years and made it possible to compare more countries. A review of nineteen small replications by the Danish researcher Mikael Søndergaard found that together they did statistically confirm all four dimensions, and the strongest confirmation was for the individualism versus collectivism dimension, which is the dimension I use for this study. The individualism index on the IBM study is based on fourteen survey questions about work goals, in which people were asked to express the importance of personal time, freedom, challenge, training, physical conditions and use of skills for them in an ideal job, on a scale from 1 (utmost importance) to 5 (very little to no importance) (Hofstede et al., 2010).

Of course, this is not enough to represent the culture of the whole society. But the correlations of the IBM individualism country scores with the non-IBM data about other

characteristics of societies confirm and validate the claim that this dimension from the IBM data does indeed measure individualism (Hofstede et al., 2010).

The individualist countries which are the minority of the countries in the world, have a loose tie between individuals; which means everyone is at first concerned about themselves and possibly their first-degree families. People rarely see their extended family or get involved with their friends' and neighbors' problems. Children learn quickly that their most important goal is to gain independence and leave their parents' home, to which they would rarely go back. A successful and healthy person in these societies is in no way dependent on a group of any sort (Hofstede et al., 2010). On the contrary, in collectivist countries which are the majority of the countries in the world, people learn from birth onward to live unquestionably loyal to groups and be practically and psychologically dependent on the groups, and receive the same treatment in return. Starting from family to extended family, school and the rest of the social groups they continue to get involved with. In other words; the interest of the group prevails over the interest of the individual (Hofstede et al., 2010).

Stepwise regression was used to see what quantitative information about the countries is associated with the differences in individualism scores. For example, individualist countries were high on wealth. Wealth; a possible compounding factor in culture, was measured by GNI (gross national income; the value produced by a country's economy in a given year) at the time of the IBM surveys. It explained 71 percent of the differences in the individualism scores for the original 50 IBM countries. However, it was not determined as a causal association (Hofstede et al., 2010). In my analysis, I used Gross Domestic Product (GDP) as a control variable for wealth, which is a documentation for a given country's economic health.

We can expect to find tendencies in mean differences between countries in their individualism/collectivism, but the construct is too broad to be referred to as a dichotomy (Hofstede et al., 2010; Schwartz, 1990). It is also important to consider within-nation differences in ethnic and cultural traits and avoid overgeneralization; as studied by Green, Deschamps & Páez (2005). They did a typological analysis by creating four combinations: self-reliance (which is an individualist behavior) and competitiveness (which was found to be a neutral behavior), self-reliance and non-competitiveness, group-oriented interdependence (which is a collectivist behavior) and competitiveness, and group-oriented interdependence and non-competitiveness, and ran the analysis between 20 countries to see whether they fit into presumed typologies.

Individualism and collectivism in a school setting

It is evident that this cultural dimension also exists among pupils in school; the target population of this study. In a school classroom, as often reported by teachers who moved from an individualist country to a collectivist setting, students from collectivistic countries are reluctant to speak up, unless they represent something more than themselves; e.g., if they are speaking on behalf of their group for a classroom discussion. The purpose of education in collectivist countries is stressed on becoming an acceptable group member and attaining higher social associations by the means of a degree, whereas in individualist countries it is more about learning and continuing to learn as much as one can to become a knowledgeable independent individual, attain self respect and sense of accomplishment as well as improve personal economic status (Hofstede et al., 2010).

Hypothesis I

With that in mind, I hypothesize that in a low-stakes assessment (an assessment that has neither positive outcomes/rewards nor negative consequences/punishments for the individual participants), there is variation in students' attitude towards the assessment based on their cultural background. Students from collectivist cultures care more about being representatives of their group (school/country) and therefore are more attentive and careful on these assessments. On the other hand, students from individualist countries where the individual benefit is the overarching mindset, are not as determined to perform well on a low-stake assessment; explicitly because the stake is low for them as individuals. A participant with an individualist cultural background would then be more careless on the assessment, read the items less attentively, and therefore would be more likely to miss the change in item wording direction. So it is more probable to have inconsistent responses to the mixed-worded scales from these respondents.

Gelfand's looseness versus tightness theory

The second cultural taxonomy is demonstrated by Michele J. Gelfand, an American cultural psychologist and psychology professor at the University of Maryland, who looks at culture from the attitude towards following social norms (i.e., rules for acceptable behavior in society). She claims that social norms are rather ignored in cultural comparison theories (Kofinas, 2019). Understanding them not only helps us make sense of different cultures, but also predict the behavior, and avoid unwanted consequences, e.g., drawing inappropriate and invalid comparisons between countries with different social norms.

The tightness or looseness of the cultural division determines the strength of social norms and the strictness of their enforcement in society. Tight cultures have very strong social norms and little tolerance for disobedience and deviant behavior. Loose cultures on the other hand, have weaker social norms with higher flexibility and tolerance for deviant behavior. As it sounds, the strengths of one can be the liabilities of the other, so there is no saying that one trade is in general "better" than the other for every country; and it is important to recognize that tight and loose cultures function in accordance with their ecological and historical context (Gelfand et al., 2011; Gelfand, Harrington & Jackson, 2017). Gelfand et al. (2011) suggest a relationship between tightness and ecological or human-made threats (e.g., natural disaster, invasion and chaos). The more a nation is exposed to threats, the more it feels the need to tighten rules and punishments to maintain order and to survive. Gelfand et al. (2011) found more cultural tightness among countries under ecological or historical threats, with higher population density and with shortage of natural resources. Tight countries also suffered from more natural disasters and have been more exposed to threats from their neighboring countries. Tight nations in an institutional context can be recognized by having more autocratic governments, restricted media, more severe punishment for criminals, and more power of religion. It can also manifest itself via everyday situations, or psychological adaptation which increases self-monitoring and self-regulation (Gelfand et al., 2011).

Both extreme freedom and extreme constraint are harmful, as shown in a study by Harrington & Gelfand (2014) which looks at the relationship between the level of tightness and constructs like happiness, life expectancy and economic status. In conclusion, a balance between freedom and constraints works to the country's benefit; and this is flexible for every country

according to the situation and their needs; that how much flexibility or tighter rules they can introduce into their activities to profit from.

Pertti J. Pelto (1968), an associate professor of anthropology at the University of Minnesota, in earlier anthropological research on tightness/looseness dimension emphasizes the importance of having a mutual criterion while referring to a population as tight or loose. He even suggests that it is highly likely that a nation wouldn't fit into one category, and the phenomenon should be treated as a continuum. This is also demonstrated by Harrington & Gelfand (2014); who found a wide variation in tightness/looseness across the 50 states of America.

Gelfand et al. (2011) measured the overall strength of social norms and tolerance of deviance with a six-item Likert scale, e.g., "There are many social norms that people are supposed to abide by in this country". They measured the degree of constraint in social situations by having the participants rate the appropriateness of 12 behaviors in 15 different situations, e.g., arguing in a classroom. They also measured psychological adaptation by well-validated measures. The data was initially collected from 6823 respondents across 33 countries. The participants were adults from different professions and backgrounds, as well as university students. Even though this dimension was not directly connected to a school setting in any of the primary studies, I would still assume that it exists among pupils as well.

Hypothesis II

Based on the distinction between tight and loose cultures with regard to situational strength, a tight setting is more strict and limited with regard to the variety of acceptable behavior in an everyday situation (like school). Individuals are more aware and cautious of their

behavior because they know they are being monitored and evaluated, even in an unofficial manner, e.g., self-regulation (Gelfand et al., 2011). Therefore, in a tight country, we could expect the respondents to pay high attention in filling out a questionnaire if they are asked to do so, whereas in a loose country it could be challenging to get an adequate amount of dedication.

Reading ability

Other than culture, another respondent characteristic that varies between countries and can potentially have an association with the proportion of inconsistent respondents is their cognitive ability; explicitly reading achievement. Responding correctly to mixed-worded items first requires detection of the difference in item wordings, and second adjustment of the responses accordingly by the respondents (Steinmann et al., 2021). For a full comprehension of a statement, young adolescents match the assembled phonology with the lexical representation of the word, i.e., "decode" the statement (Steacy, Elleman, Lovett & Compton, 2016). Gnambs & Schroeders (2020) showed that since responding to negatively worded items demands more complex cognitive processes than responding to positively worded items, the unidimensionality of a scale increases when there are higher levels of reading competency and reasoning among the respondents. It is therefore relevant to investigate the association between inconsistent responding to mixed-worded scales and the average reading literacy in different countries.

Hypothesis III

Deficits in decoding skills mean poor word identification (Steacy et al., 2016) which would lead the poor readers to fail to identify and distinguish between positively and negatively worded items. Negatively worded items create confusion and extra difficulty in interpreting the

item content for everyone, but especially for those with lower reading abilities who struggle with general processing deficiency and/or linguistic impairment issues more than typical readers (Hu, Vender, Fiorin & Delfitto, 2018; Zhang et al., 2016), so more shares of inconsistency is expected among respondents with lower reading achievement scores.

Research question

I aim to investigate the between-country variation in the proportions of inconsistent respondents to mixed-worded scales by looking at three country characteristics that might be associated with the phenomenon. My three research questions are: are there more shares of inconsistency on mixed worded scales among more individualist countries, more culturally loose countries, and countries with lower average reading literacy?

Methods

Data

The data for the study was from PISA (Program for International Student Assessment) 2018. PISA is an International Large-Scale Assessment that is conducted every 3 years across several countries. The participants are 15-year-old students enrolled in grade 7 or higher, and the aim is to measure their ability to use their reading, mathematics and science knowledge and skills to meet real life challenges (OECD, 2019a). There is also a set of context questionnaires along with the ability tests, which some of them include mixed-worded scales. Originally, 79 countries participated in the 2018 cycle, from which I included 75 in my study. Some countries were excluded because they didn't have any data on the mixed-worded scale that I used from the questionnaire. The sample size differed for each country, ranging between 3,000 to 35,000

participants. I chose PISA because compared to other International Large Scale Assessments like TIMSS (Trends in International Mathematics and Science Study) and PIRLS (Progress in International Reading Literacy Study) which have participants of 9-10 years old, 15-year-old adolescents are better representatives of the countries' culture than younger students who are still developing their social personality.

Measures

Identifying Inconsistent Respondents

The Scale. In order to identify inconsistent respondents for this study, I chose the students' sense of belonging to school (BELONG, ST034) mixed-worded scale from PISA 2018 student questionnaire. This scale consisted of a balanced and adequate number of positively (i.e., items Pos1, Pos2, & Pos3) and negatively (i.e., items Neg1, Neg2 and Neg3) worded items with the same four-point Likert response scale (see Table 1). As shown below, the negatively worded items don't have a negative particle (e.g., not), but rather are polar opposites having a negative connotation to them. This implies that noticing them requires attention, and is not merely about language and comprehending negation. I excluded countries for which there was no data on this scale; as inconsistent responding to the mixed-worded scale was my focus (Israel, Lebanon and North Macedonia).

Table 1

Item Wording of the BELONG Scale in PISA 2018 Student Questionnaire in Original Order

Items	To what extent do you agree with the following statements?
Neg1	I feel like an outsider (or left out of things) at school.
Pos1	I make friends easily at school.
Pos2	I feel like I belong at school.
Neg2	I feel awkward and out of place in my school.
Pos3	Other students seem to like me.
Neg3	I feel lonely at school.

Note. The response categories were 1 = Strongly agree, 2 = Agree, 3 = Disagree, 4 = Strongly disagree.

FMA. I fit a constrained factor mixture analysis (FMA) model (Steinmann et al., 2021) on the BELONG scale items separately for each of the 75 PISA participating countries that were included in the study, in order to classify students into inconsistent and consistent respondents. The item scores were treated as continuous and the original scores were used for the analysis. The FMA analyzes were run in Mplus demo version (Muthén & Muthén, 1998-2017) and further processed in R with the packages MplusAutomation (Hallquist & Wiley, 2018). Cases with missing values on all the BELONG items were excluded from the FMA analyzes. In using large-scale assessment data, I accounted for students being nested into schools, and used weights in the analyses (SENWGT). I used MLR estimator (maximum likelihood estimation with robust standard errors), generated 5,000 initial stage random starting values, and 500 final stage optimizations, to prevent the models from making mistaken, short-sighted estimations.

The constrained FMA models identified two latent classes of inconsistent and consistent respondents for each of the 75 countries in the dataset, and estimated the proportion of the respondents that belonged to each class. I also retrieved sample statistics such as means and factor loadings for each class from the models. The models were evaluated by checking the unstandardized factor loadings and the entropies. I allowed the first factor loading for the first positive item to be estimated freely. The factor variance was set to 1 and the factor mean was set to 0. But in the first class (the inconsistent class), the variance of the factor was estimated freely. The entropy values showed how precisely the model performed in classifying the respondents into two classes.

It is noteworthy that the BELONG scale was not a part of the simplified version of the questionnaire, the so-called *une heure* test which was administered to low-performing students and students with special education needs (OECD, 2019a). Therefore, students with presumably lower reading abilities were not included in the inconsistent/consistent classification.

Predictor Variables

Individualism. I used the country comparison tool (Hofstede Insights, 2021) which was developed based on Hofstede's 6-Dimension model to sort and compare the individualism/collectivism dimension of the countries. The tool scores several countries on individualism from close to zero (the most collectivist country) to 100 (the most individualist one) (Hofstede et al., 2010), including 72 out of 75 countries in this study. The dimensions were defined and the scores were generated over time, initially between 1967 and 1973 by studying a large body of data which was collected to measure how values in the workplace are influenced

by culture. The questionnaire used to collect this data included items such as: “How important is it to you to have a job which leaves you sufficient time for your personal or family life?” and “How important is it to you to work with people who cooperate well with one another?” (originally, these items did not target one specific cultural dimension).

Tightness. I referred to two of Gelfand’s publications for Tightness Index, to sort and compare the strength of norms in a nation and the tolerance for people who violate them. The first study (Gelfand et al., 2011) assigned tightness scores to 33 countries. The scale underlying the scores had 6 items, for example: “There are many social norms that people are supposed to abide by in this country”, “There are very clear expectations for how people should act in most situations”, “In this country, if someone acts in an inappropriate way, others will strongly disapprove”, and “People in this country almost always comply with social norms”. This measure was then expanded and validated across 57 nations (Eriksson et al., 2021).

The scores were laid on different scales in the two studies, and I used the scale of the 2021 one which included more countries. The scores were in a range from -1.0 (the most culturally loose) to 1.5 (the most culturally tight) countries. Belgium, Chinese Taipei, France, Hong Kong, New Zealand, and Norway were a part of the original study published in 2011, but not the expanded one. I wanted to include them anyway, so I transformed their scores from the first study to the scale of the expanded study published in 2021, assuming stability across the years (i.e., if such a country had a matching tightness score with another country in the original study, their transformed score was matched and set equal to this country’s score in the expanded study; if no exact match was found, the transformed score was an interpolation average between

the two closest score neighbours in the original study). I ended up having this variable for 49 countries.

Reading Proficiency. The average reading score for each country was taken from PISA 2018 international report (OECD, 2019a). It is important to mention that the mean scores were for all of the PISA 2018 participants, including those who took the simplified version of the assessment; *une heure*; whereas they did not answer the BELONG scale. In 2018, an adaptive testing method was used for reading proficiency to improve measurement precision. This means that students completed sets of reading items based on their proficiency (they were presented with easier items in the next step if they showed low reading ability in one step) (OECD, 2019b). The reading assessments were structured as units which were presented to students and they had to answer items according to the unit text. The scores in the report were in a range from 340 to 555. The score for Spain was unavailable due to anomalies indicating that students responded unnaturally quickly to the items. The average reading score for Vietnam was not included in the original report because they used the paper-based assessment and their results were not approved at the time the report was published. I chose to keep the PISA official report as my main resource and did not assign reading scores to Spain and Vietnam, and therefore had this variable for 73 countries.

Gross Domestic Product (GDP). Gross domestic product (GDP) functions as a comprehensive scorecard of a given country's economic health. Small, rich countries and more developed industrial countries tend to have the highest per capita GDP (GDP divided by population). As mentioned before, there seems to be a relationship between some countries' wealth and their level of individualism. To check for this, I used per capita GDP scores from

PISA 2018 international report (OECD, 2019a). The GDP values represent per capita GDP in 2018 at current prices, expressed in US Dollars. The conversion from local currencies to equivalent USD accounts for differences in purchasing power across countries and economies. The GDP values in the PISA 2018 report ranged from 7,000 to 131,000 USD, and were available for 72 of my countries (not available for China, Spain and Vietnam).

Statistical Analysis

To answer the research question, I first looked at country-level descriptive statistics (i.e., correlations) between the four measures and inconsistency. In a further step, I used stepwise linear regression analysis using the R package lavaan (Rosseel, 2012) to regress the inconsistency rate in the countries on their individualism, tightness, reading and GDP scores, to determine the significance and magnitude of each predictor variable's relationship with the shares of inconsistency. I specified tolerance statistics to check for multicollinearity between the 4 predictors.

Results

FMA Model Results and Sample Statistics

Unstandardized factor loadings of the positively worded items were positive and equal in both classes. For negatively worded items, unstandardized factor loadings were set to be positive in the inconsistent class, and negative in the consistent class, and were equal in size. The entropy values of the FMA model ranged between 0.555 and 0.976 across countries. Four countries had an entropy value lower than 0.6: Thailand, Brunei Darussalam, Morocco and Panama.

I found that in general, the negatively worded items had higher means than the positively worded items. Also, the negatively worded items tended to be negatively skewed, while the positively worded items tended to be positively skewed. This showed that on average, the respondents answered negatively to positively worded items and positively to negatively worded items, showing low sense of belonging to school. Negatively worded items also carried higher factor loadings in both classes, but all factor loadings (for positively and negatively worded items) were on average higher in the inconsistent class, and had higher reliability omegas (see Table 2).

Table 2

Model Estimated Results of the Constrained Factor Mixture Analysis (FMA) for the BELONG Scale in PISA 2018

	Inconsistent class	Consistent class
Factor mean	0.655 (0.95)	0 (0)
Factor variance	1 (2.143)	1 (1)
λ Pos1	0.772 (0.49)	0.653 (0.49)
λ Pos2	0.735 (0.47)	0.616 (0.47)
λ Pos3	0.73 (0.418)	0.606 (0.418)
λ Neg1	0.775 (0.517)	-0.66 (-0.517)

λ Neg2	0.776 (0.515)	-0.663 (-0.515)
λ Neg3	0.817 (0.536)	-0.713 (-0.536)
M Pos1	2.396	2.025
M Pos2	2.47	2.118
M Pos3	2.405	2.063
M Neg1	2.303	3.146
M Neg2	2.434	3.112
M Neg3	2.33	3.277
Ω Pos1	0.596	0.426
Ω Pos2	0.54	0.38
Ω Pos3	0.533	0.367
Ω Neg1	0.6	0.435
Ω Neg2	0.602	0.44
Ω Neg3	0.667	0.508

Note. The reported statistics in the table are the averages of the statistics across 75 countries. λ represents the factor loadings, *M* represents the means and *Ω* represents reliability. The values in the parantheses are the unstandardized statistics.

As expected, items with the same wording correlated positively with each other (Table 3), with negatively worded items correlating much stronger with each other than positively worded items (average item intercorrelation = 0.252 for positively and 0.838 for negatively worded items). The average item intercorrelation between positively and negatively worded items was -0.411.

Table 3

Correlations Between the Positively and Negatively Worded Items in the FMA Analysis

	Pos1	Pos2	Pos3	Neg1	Neg2	Neg3
Pos1	1					
Pos2	0.232	1				
Pos3	0.3	0.226	1			
Neg1	-0.434	-0.321	-0.372	1		
Neg2	-0.499	-0.241	-0.566	0.85	1	
Neg3	-0.52	-0.194	-0.554	0.83	0.836	1

Note. The reported correlations in the table are the averages of the correlations across 75 countries.

The means were on average 3.178 for the negatively worded items and 2.068 for the positively worded items in the consistent class. A score above 2 on negatively worded items and almost equal to 2 on positively worded items expresses a rather low sense of belonging to school.

In the inconsistent class, the average mean was 2.355 for the negatively worded items and 2.423 for the positively worded items. Evidently, there is not a big difference between the average means of the negatively and positively worded items in the inconsistent class, and the average scores are closer to the midpoint of the four-point Likert scale rather than the ends, which means a respondent from this class did not switch sides in the response scale and either agreed or disagreed with both sets of items.

Descriptive Statistics for the Predictor Variables

Individualism. The mean of the individualism scores for the 72 countries (as measured by Hofstede on a scale from 0 to 100) was 43, and the standard deviation was 23. UK, Australia and USA were the most individualist countries, with scores of 89, 90 and 91, respectively. Indonesia, Colombia and Panama were the least individualist (most collectivist) countries, with scores of 14, 13 and 11, respectively. Figure 1 shows that the distribution of the scores was bimodal; one mode around the value of 23 for more collectivist, and one mode around the value of 70 for more individualist countries. The distribution also shows that there are more collectivist countries than there are individualist.

Tightness. The mean of the tightness scores for the 49 countries (as measured by Gelfand on a scale from -1 to 1.5) was -0.06, and the standard deviation was 0.35. Indonesia, Saudi Arabia and Qatar were the most culturally tight countries, with scores of 0.5, 0.62 and 0.85, respectively. Netherlands, Colombia and Hungary were the most culturally loose countries, with tightness scores of -0.54, -0.58 and -0.6, respectively. Figure 1 shows that the distribution of the scores was bimodal, one mode around the value of -0.3 for culturally looser, and one mode

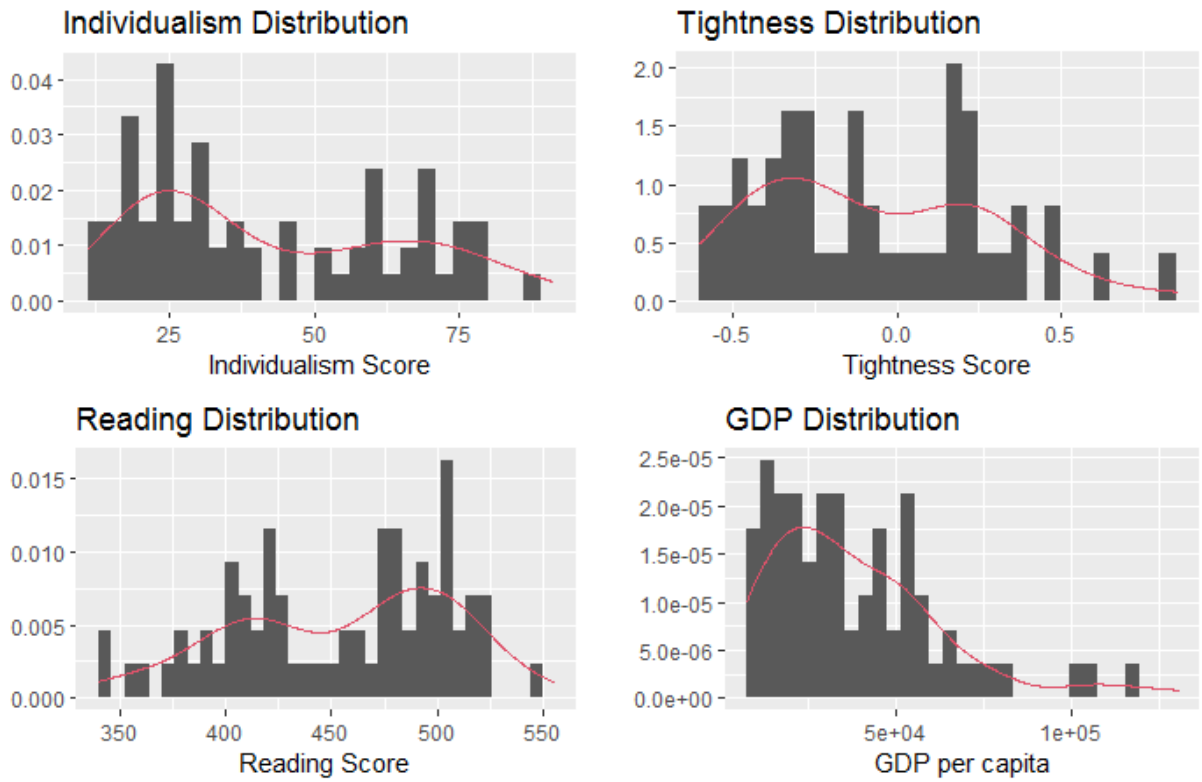
around the value of 0.3 for culturally tighter countries. The distribution also shows that there are more culturally loose countries than there are culturally tight.

Reading. The mean of the reading scores for the 73 countries (as reported by PISA on a scale from 340 to 555) was 455, and the standard deviation was 52.6. Macao, Singapore and China had the highest mean reading scores, with scores of 525, 549 and 555, respectively. Kosovo, Dominican Republic and Philippines had the lowest mean reading scores, with scores of 353, 342 and 340, respectively. Figure 1 shows that the distribution of the scores was bimodal, one mode around the value of 500 for countries with higher reading scores, and one mode around the value of 415 for countries with lower reading scores. The distribution also shows that more countries scored above average on reading than they did below average.

GDP. The mean per capita GDP for the 72 countries (as reported by PISA on a scale from 7,000 to 131,000 USD) was 38903, and the standard deviation was 25935. Luxembourg, Macao and Qatar had the highest GDP per capita, with values of 106704, 116807 and 130475 USD, respectively. Philippines, Morocco and Moldova had the lowest GDP per capita, with values of 8935, 8932 and 7304 USD, respectively. Figure 1 shows that the scores were clustered around 25,000 USD (slightly below average). The distribution also shows that more countries were below average on per capita GDP than they were above average.

Figure 1

Distributions of the Predictor Variables



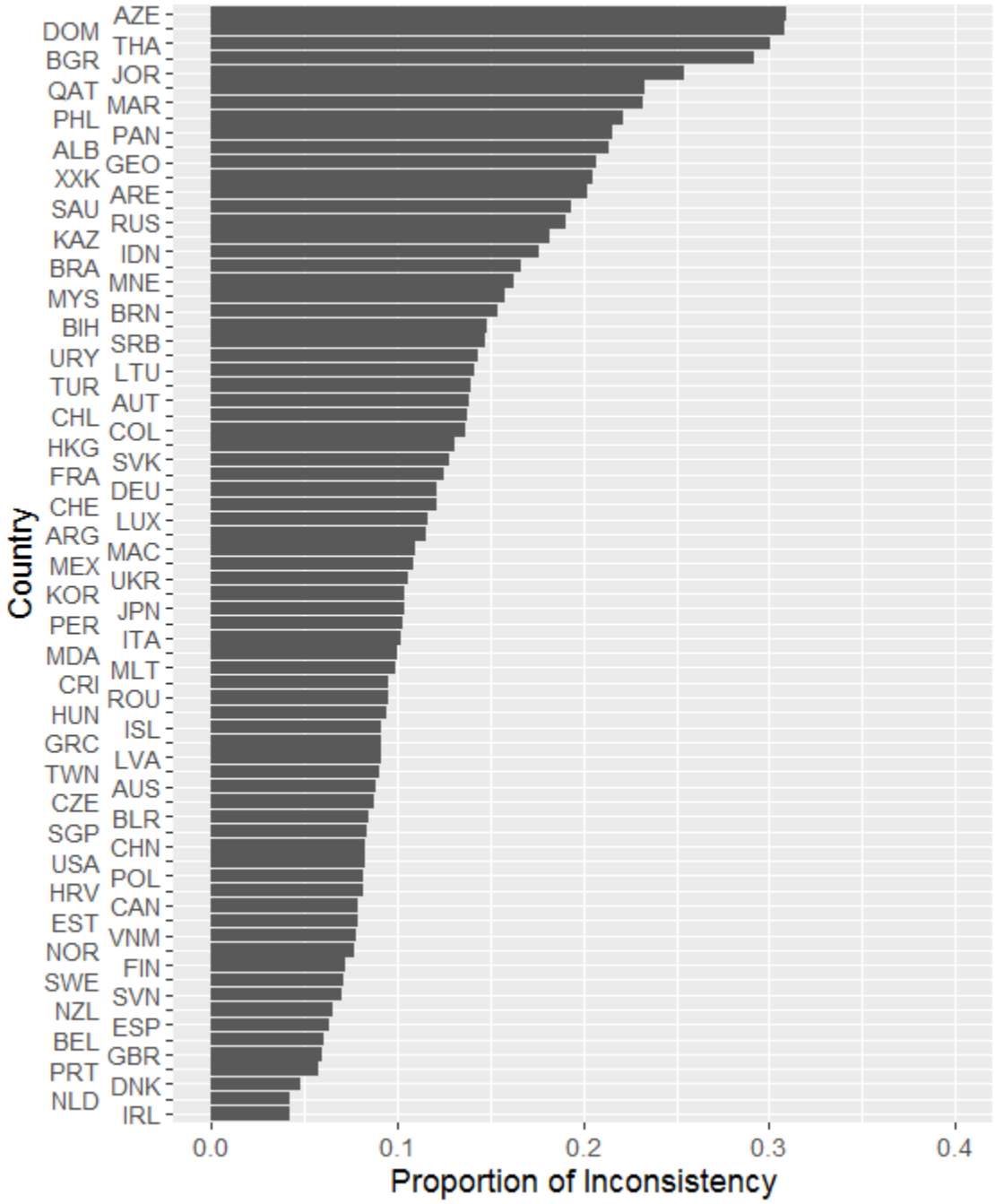
Note. The horizontal axes represent the variables' scores, and the vertical axes represent the magnitude of each score across 75 countries. The red smooth lines visualize the density of the score distributions.

Inconsistent Responding on the Mixed-worded Sense of Belonging to School Scale

The final proportion of inconsistent respondents per country based on the estimated models were all smaller than the proportion of consistent respondents, as expected. The inconsistency proportion in the countries ranged between 4% (Ireland, Netherlands and Denmark) and 30% (Azerbaijan, Dominican Republic and Thailand) with an average of 13% across countries (Figure 2).

Figure 2

Prevalence of Inconsistent Respondents to the BELONG Scale across Countries



Note. The figure displays the shares of inconsistent respondents to the BELONG self-concept scale on the horizontal axis.

Intercorrelations

Table 4 shows the correlation coefficients between all of the predictor variables and outcome. Individualism had a weak/moderate negative association with inconsistency. This contradicts the expectations. However, the weakness of this association can be seen in examples like Costa rica, Vietnam and Singapore who had low scores on individualism (15, 20 and 20; respectively) but low percentage of inconsistency (9, 7 and 8 percent, respectively).

Tightness had a weak positive association with inconsistency. This also contradicts the expectations. However, countries like Vietnam, Singapore and Sweden which are moderately high on tightness (0.39, 0.36 and 0.34; respectively) but low on inconsistency (7, 8 and 7 percent; respectively) show the weakness of this association.

Reading had a strong negative association with inconsistency. This conforms to expectations. For example, Azerbaijan, Dominican Republic and Thailand had the highest percentage of inconsistency (about 30%) and their mean reading scores were among the lowest (342-393) while Ireland, Netherlands and Denmark had the lowest percentage of inconsistency (about 4%) and their reading scores were among the highest (485-518).

GDP had a weak negative association with inconsistency, and individualism, reading and GDP all correlate positively with each other. From this I can suggest that my assumption that the richer countries are also more individualist is true, and since the relationship between reading scores and inconsistency is strongly negative, it makes sense that the relationship between individualism and GDP with inconsistency would also be negative.

Table 4

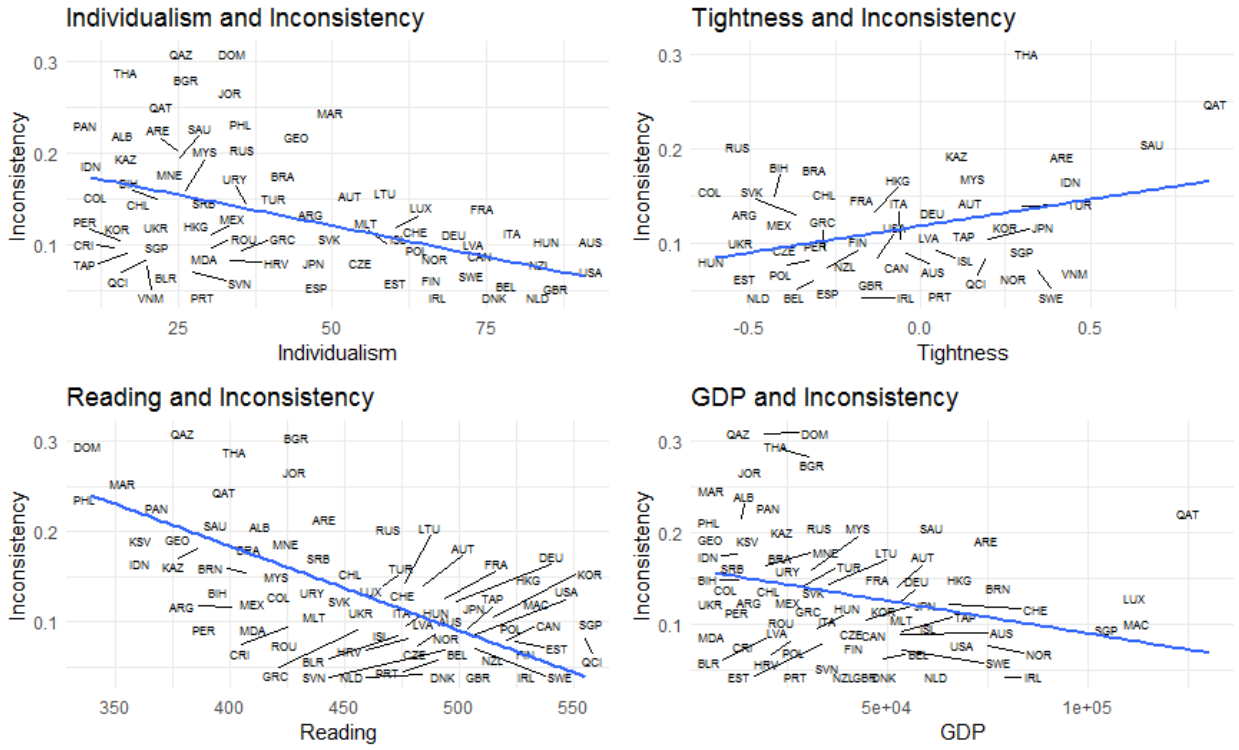
Correlation Coefficients Between the Predictor and the Outcome Variables in the FMA Analysis

	Individualism	Tightness	Reading	GDP	Inconsistency
Individualism	1				
Tightness	-0.201	1			
Reading	0.512	-0.07	1		
GDP	0.326	0.49	0.525	1	
Inconsistency	-0.477	0.308	-0.754	-0.293	1

Note. The reported correlations in the table are the averages of the correlations across countries for which the variables were available.

Figure 3

The Relationships Between Shares of Inconsistency and the Predictor Variables



Note. The horizontal axes represent the variables' scores, and the vertical axes represent the proportions of inconsistency across 75 countries. The blue smooth line displays the linear relationship between the inconsistency rate and each predictor variable.

Linear Regression Analysis Results

In order to simplify the analyzes, I re-scaled all of the predictor variables into a range of 0 to 1 (same as the outcome variable: inconsistency) before fitting the models to the data. I ran a stepwise linear regression analysis to regress the inconsistency rate on individualism, tightness, reading and GDP. I introduced the predictors one by one into the regression model. I used

case-wise (full-information) maximum likelihood estimator (FIML) with robust standard errors to account for missing data on countries that did not have some of the predictor variables in the dataset. The results showed that the regression coefficients were significant in the first three steps for individualism, individualism and tightness, and individualism and tightness and GDP (Table 5). But once reading was introduced into the multiple regression model, the regression coefficients were no longer significant for any of the other predictor variables. Reading also significantly increased the R-squared value of the models. The other 3 predictors only explained 22-37 percent of the variance in inconsistency in the first three steps, but together with reading, they explained 64 percent. This once again shows the strength of association between reading ability and shares of inconsistency in the mixed-worded scale, that it overpowered the other predictors (that were included) in the multiple predictor analysis.

The tolerance values for each of the predictor variables were all above 0.4, which showed that there is no concern for multicollinearity (Allison, 1999). The variables were independent of each other enough to be safely used in the analysis (Table 5).

Table 5

Stepwise Regression of Inconsistency on the Predictors

Inconsistency ~	Step 1 β *(SE)*	Step 2 β *(SE)*	Step 3 β *(SE)*	Step 4 β *(SE)*	Tolerance
Individualism	-0.471 ***	-0.370 ***	-0.231 *	-0.057	0.638
	(0.019)	(0.019)	(0.023)	(0.016)	
Tightness		0.320 *	0.466 **	0.275	0.535
		(0.037)	(0.041)	(0.04)	
GDP			-0.434 *	-0.053	0.404
			(0.052)	(0.037)	
Reading				-0.677 ***	0.545
				(0.023)	
R^2	0.222	0.315	0.374	0.64	

Note. Significance codes: p = 0 ***, p < 0.001 **, p < 0.01 *, p < 0.05 ., p < 0.1. β represents regression coefficient.

Discussion

The aim of this study was to identify any association between the proportions of inconsistency with two cultural dimensions and with reading literacy among 75 countries. Associations which, to the best of my knowledge, have not been investigated in previous studies

about the inconsistent responding phenomenon. Findings from the stepwise regression analysis showed that when only including the individualism variable, there was a significant negative association between individualism and inconsistency, while it was hypothesized to be positive. When including the individualism and tightness variables, the association between individualism and inconsistency remained significantly negative, and there was a significant positive association between tightness and inconsistency, while it was hypothesized to be negative. When including the individualism, tightness and GDP variables, the association between individualism and inconsistency and between tightness and inconsistency remained the same, and GDP had a significant negative association with inconsistency. When including the individualism, tightness, GDP and reading variables, individualism, tightness and GDP became insignificant, and reading had a strong negative association with inconsistency, which was in line with the hypothesis.

The results indicated that instead of culture, the students' reading ability was the most relevant predictor for between-country differences in inconsistency rate on mixed-worded scales, and I found that on average, richer countries, i.e., countries with higher GDP, and more individualist countries had higher reading scores. The negative association between individualism and inconsistency is then explained by these findings.

Limitations

It is relevant to emphasize that Hofstede's and Gelfand's measures of the cultural dimensions might not be perfectly accurate and the constructs might not have been defined well enough. For example, measures of individualism and collectivism were about values in the workplace and the samples were adults; whereas the sample for this study was 15 year old high school students and the individualism construct might have had a different implication for them.

Findings from the typological study of Green et al. (2005) also show that not all presumably individualist countries fit into self-reliance categories, and not all presumably collectivist countries are in the group-oriented interdependence types. Attitudes relating to individualism/collectivism can be activated as a function of social contexts and relations; for example, the same person might act as a collectivist in a family situation while being completely individualist in a business meeting (Green et al., 2005).

The tightness data was also collected among adults, and the construct is simply too broad to be easily specified. For example, in a country with a high tightness score, there might be strict punishments for specific behaviors against the law, but people can easily get away with other behaviors. So an item like "In this country, if someone acts in an inappropriate way, others will strongly disapprove" is not specific enough to distinguish between inappropriate behaviors in different domains.

Other factors, like respondents wanting their country to appear as good and acceptable internationally can also influence the way they respond to measures of culture. We can also see how the respondents' age has an impact on their responses by looking at the central tendency in the responses in this study. Most 15 year old students expressed a low score for the sense of belonging to school construct, whereas usually in the self-concept scales (e.g., reading self concept scale in PIRLS among 9-10 year old students (Mullis, Martin, Foy & Drucker, 2012)), students tend to give higher responses expressing more positivity on the construct.

There are several other cultural dimension theories and taxonomies that were not considered for this study. There are also other predictors such as language differences, translation errors and test motivation that are potentially relevant to look at for the between-country differences in inconsistency rates.

My study does not show that lower mean reading scores cause higher proportions of inconsistent respondents. There might be other mechanisms in place, e.g., students who are careless in their responses can be flagged as inconsistent respondents and score low on achievement tests. There might even be an issue of spurious correlation; where an unknown variable has an independent impact on inconsistency and on one or more of the predictors, and it causes a correlation between the two variables to appear when there exists no correlation or a different correlation.

The participants for this study were high school students, and I only looked at one mixed-worded scale and one construct at one point in time, so the conclusions can't be generalized for all mixed-worded scales across all populations at all times. As suggested by Steinmann et al., (2021), inconsistent response attitudes that are related to persons' characteristics and not the instrument (i.e., culture and cognitive ability) should be stable across mixed-worded scales and over time, but this was not investigated in this study.

Implications of the study

The most dominant finding of this study was how much reading literacy is important when it comes to using mixed-worded scales in cross-country comparisons. After adding reading score to the multiple regression model, the r-square value increased from 37 to 64 percent, showing reading score on its own added 27 percent value to the percentage of variance in inconsistency that was explained by the model. It suggests that the mixed-worded scales are especially problematic in countries with low reading scores. To ensure cross-country comparability, it is better to avoid mixed-worded scales when reading literacy differs to a great extent among groups of participants. This is a confirmation for the conclusions of several other

studies about using mixed-worded scales (Cheng & Hamid, 1997; Gnambs & Schroeders, 2020; Jiang et al., 2018; Marsh, 1986, 1996; Pilotte & Gable, 1990; Schmitt & Stuits, 1985; Steinmann et al., 2021; van Sonderen et al., 2013; Wong et al., 2003; Zhang et al., 2016).

A part of my findings suggest that it is not necessarily problematic to use mixed-worded scales in every country and setting (i.e., some countries had lower than 10% of inconsistency while in theory we have evidence that a minimum of 10% compromises the scale unidimensionality (Schmitt & Stuits, 1985)). However, since my study focuses on international large-scale assessments with big variation in reading literacy among participants, inconsistency rates of up to 30% strongly suggests that it's not the best practice to use mixed-worded scales in this setting.

References

Allison, P. D. (1999). *Multiple regression: A primer*. Pine Forge Press.

Čeněk, J. (2020). Cultural dimension of individualism and collectivism and its perceptual and cognitive correlates in cross-cultural research. *Journal of Education Culture and Society*, *6*(2), 210–225. https://doi.org/10.15503/jecs20152.210.225

Cheng, S.-T., & Hamid, P. N. (1997). Dispositional optimism in Chinese people: What does the life orientation test measure? *International Journal of Psychology*, *32*(1), 15–22. https://doi.org/10.1080/002075997400935

Chyung, S. Y., Barkin, J. R., & Shamsy, J. A. (2018). Evidence‐based survey design: The use of negatively worded items in surveys. *Performance Improvement (International Society for Performance Improvement)*, *57*(3), 16–25. https://doi.org/10.1002/pfi.21749

DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(3), 440–464. https://doi.org/10.1207/s15328007sem1303_6

Eriksson, K., Strimling, P., Gelfand, M., Wu, J., Abernathy, J., Akotia, C. S., Aldashev, A., Andersson, P. A., Andrighetto, G., Anum, A., Arikan, G., Aycan, Z., Bagherian, F., Barrera, D., Basnight-Brown, D., Batkeyev, B., Belaus, A., Berezina, E., Björnstjerna, M., … Van Lange, P. A. M. (2021). Perceptions of the appropriate response to norm violation in 57 societies. *Nature Communications*, *12*(1), 1481. https://doi.org/10.1038/s41467-021-21602-9

Gelfand, M. J., Harrington, J. R., & Jackson, J. C. (2017). The strength of social norms across human groups. *Perspectives on Psychological Science*, *12*(5), 800–809. https://doi.org/10.1177/1745691617708631

Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., Duan, L., Almaliach, A., Ang, S., Arnadottir, J., Aycan, Z., Boehnke, K., Boski, P., Cabecinhas, R., Chan, D., Chhokar, J., D'Amato, A., Ferrer, M., Fischlmayr, I. C., … Yamaguchi, S. (2011). Differences between tight and loose cultures: A 33-nation study. *Science*, *332*(6033), 1100–1104.

Gnambs, T., & Schroeders, U. (2020). Cognitive abilities explain wording effects in the Rosenberg self-esteem scale. *Assessment*, *27*(2), 404–418. https://doi.org/10.1177/1073191117746503

Green, E. G. T., Deschamps, J.-C., & Páez, D. (2005). Variation of individualism and collectivism within and between 20 countries: A typological analysis. *Journal of Cross-Cultural Psychology*, *36*(3), 321–339. https://doi.org/10.1177/0022022104273654

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621–638. https://doi.org/10.1080/10705511.2017.1402334

Harrington, J. R., & Gelfand, M. J. (2014). Tightness-looseness across the 50 United states. *Proceedings of the National Academy of Sciences*, *111*(22), 7990–7995. https://doi.org/10.1073/pnas.1317937111

Hofstede, G. (1983). National cultures in four dimensions: A research-based theory of cultural differences among nations. *International Studies of Management & Organization*, *13*(1–2), 46–74. https://doi.org/10.1080/00208825.1983.11656358

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage Publications.

Hofstede, G., Hofstede, G. J., & Minkov, M. (2010). *Cultures and organizations: Software of the mind; intercultural cooperation and its importance for survival* (3rd ed.). McGraw-Hill.

Hofstede Insights. (2021). *Country Comparison*. https://www.hofstede-insights.com/country-comparison/

Hu, S., Vender, M., Fiorin, G., & Delfitto, D. (2018). Difficulties in comprehending affirmative and negative sentences: Evidence from Chinese children with reading difficulties. *Journal of Learning Disabilities*, *51*(2), 181–193. https://doi.org/10.1177/0022219417714775

Jiang, X., Fang, L., Stith, B. R., Liu, R., & Huebner, E. S. (2018). A psychometric evaluation of the Chinese version of the students' life satisfaction scale. *Applied Research in Quality of Life*, *13*(4), 1081–1095. https://doi.org/10.1007/s11482-017-9576-x

Kemmelmeier, M. (2016). Cultural differences in survey responding: Issues and insights in the study of response biases. *International Journal of Psychology*, *51*(6), 439–444. https://doi.org/10.1002/ijop.12386

Kofinas, D. (2019, October 7). Michele Gelfand | Rule makers and rule breakers: How tight and loose cultures wire our world (No. 103). In *Hidden forces.* Hidden Forces Productions. https://hiddenforces.io/podcasts/michele-gelfand-cultural-psychologist/

Marsh, H. W. (1986). Negative item bias in ratings scales for preadolescent children: A cognitive-developmental phenomenon. *Developmental Psychology*, *22*(1), 37. https://doi.org/10.1037/0012-1649.22.1.37

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology*, *70*(4), 810. https://doi.org/10.1037/0022-3514.70.4.810

Mullis, I. V. S., Martin, M. O., Foy, P., & Drucker, K. T. (Eds.). (2012). *PIRLS 2011 international results in reading*. IEA, TIMSS & PIRLS, International Study Center, Lynch School of Education, Boston College.

Muthén, L. K., & Muthén, B. O. (1998). *Mplus user's guide* (8th ed.). https://www.statmodel.com/html_ug.shtml

OECD. (2019a). *PISA 2018 results (volume I): What students know and can do*. PISA, OECD Publishing.

OECD. (2019b). *PISA 2018 technical report*. PISA, OECD Publishing. https://www.oecd.org/pisa/data/pisa2018technicalreport/

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). Academic Press.

Pelto, P. J. (1968). The differences between "tight" and "loose" societies. *Society*, *5*(5), 37–40. https://doi.org/10.1007/BF03180447

Pilotte, W. J., & Gable, R. K. (1990). The impact of positive and negative item stems on the validity of a computer anxiety scale. *Educational and Psychological Measurement*, *50*(3), 603–610. https://doi.org/10.1177/0013164490503016

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Schmitt, N., & Stuits, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, *9*(4), 367–373. https://doi.org/10.1177/014662168500900405

Schwartz, S. H. (1990). Individualism-collectivism: Critique and proposed refinements. *Journal of Cross-Cultural Psychology*, *21*(2), 139–157. https://doi.org/10.1177/0022022190212001

Steacy, L. M., Elleman, A. M., Lovett, M. W., & Compton, D. L. (2016). Exploring differential effects across two decoding treatments on item-level transfer in children with significant word reading difficulties: A new approach for testing intervention elements. *Scientific Studies of Reading*, *20*(4), 283–295. https://doi.org/10.1080/10888438.2016.1178267

Steinmann, I., Strietholt, R., & Braeken, J. (2021). A constrained factor mixture analysis model for consistent and inconsistent respondents to mixed-worded scales. *Psychological Methods*. https://doi.org/10.1037/met0000392

Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research*, *45*(1), 116–131.

van Sonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's learn from cows in the rain. *PloS One*, *8*(7), e68967. https://doi.org/10.1371/journal.pone.0068967

Wong, N., Rindfleisch, A., & Burroughs, J. E. (2003). Do reverse-worded items confound measures in cross-cultural consumer research? The case of the material values scale. *Journal of Consumer Research*, *30*(1), 72–91. https://doi.org/10.1086/374697

Zhang, X., Noor, R., & Savalei, V. (2016). Examining the effect of reverse worded items on the factor structure of the need for cognition scale. *PloS One*, *11*(6), e0157795. https://doi.org/10.1371/journal.pone.0157795

Appendix I (NSD application form)

NOTIFICATION FORM (ENGLISH TRANSLATION) – NSD
• Personal data
• Types of data
• Project Information Responsibility
• Sample and Criteria
• Third Persons
• Documentation
• Other approvals
• Processing
• Information Security
• Duration of project
• Additional Information
• Send in
Which personal data will be processed?
Name
No
National ID number or other personal identification number
No
Date of birth
No
Address or telephone number
No
Email address, IP address or other online identifier
No
Photographs or video recordings of persons
No
Audio recordings of persons
No
GPS data or other geolocation data
No
Demographic data that can identify a natural person
No
Genetic data
No
Biometric data
2
No
Other data that can identify a natural person
If you think that you will be processing personal data but cannot find a suitable alternative above, indicate this here.
No
Will special categories of personal data or personal data relating to criminal convictions and offenses be processed?

Racial or ethnic origin
No
Political opinions
No
Religious beliefs
No
Philosophical beliefs
No
Trade Union Membership
No
Health data
No
Sex life or sexual orientation
No
Criminal convictions and offenses
No
Project Information
Edit project Register new project Choose existing project under ‘Register new project’:
Title
“Do culture and reading literacy associate with inconsistent responding on mixed-worded scales?”
Project description
Proportions of inconsistent respondents on mixed-worded scales vary across countries. Could this variation be associated with the respondents’ cultural background and/or reading literacy? Do respondents from individualist, culturally loose and low reading achievement countries have larger percentages of inconsistency in their data?
3
Subject area
• Social sciences
Will the collected personal data be used for other purposes, in addition to the purpose of this project?
No
Explain why it is necessary to process personal data.
Personal data is not processed.
Project description Choose file...
External funding
No
Type of project
• Student project, Master’s thesis
Responsibility for data processing
Data controller
UiO
Project leader (research assistant/ supervisor or research fellow/PhD candidate)
Name: Dr. Isa Steinmann

Position: Postdoctoral Fellow at Centre for Educational Measurement, University of Oslo – Master's thesis supervisor
Email address: isa.steinmann@cemo.uio.no
Telephone number: +47 22844485
Will the responsibility for processing personal data be shared with other institutions (joint data controllers)?
No
Whose personal data will be processed?
Sample 1
15-year-old students from the 79 countries which participated in PISA 2018 assessment cycle
Recruitment or selection of the sample
Age
4
Will you include adults (18 år +) who do not have the capacity to consent?
No
Types of personal data - sample 1
-
Methods /data sources - sample 1
Select and/or describe the method(s) for collecting personal data and/or the source(s) of data
Tests for pedagogical research / psychological tests
Big data
Reseptformidleren
Forsvarets helseregister
Helsearkivregisteret
Helseundersøkelsen i Nord Trøndelag (HUNT)
Tromsø-undersøkelsen
SAMINOR
Den norske mor og barn undersøkelsen (MoBa)
Nasjonalt register for langtids mekanisk ventilasjon
Nasjonalt kvalitetsregister for barnekreft
Norsk Kvalitetsregister Øre-Nese-Hals –Tonsilleregisteret
Norsk vaskulittregister & biobank (NorVas)
Norsk Parkinsonregister & biobank
Norsk karkirurgisk register (NORKAR)
Norsk hjertinfarkregister
Gastronet
Norsk register for analinkontinens
Nasjonalt barnehofteregister
Norsk kvalitetsregister for artrittsykdommer (NorArtritt)
Norsk nakke- og ryggregister
Nasjonalt korsbåndregister
Nasjonalt register for leddproteser
NorKog
Norsk MS-register og biobank
Nasjonalt register for KOLS

Nasjonalt kvalitetsregister for lymfom og lymfoide leukemier
Nasjonalt kvalitetsregister for lungekreft
Nasjonalt kvalitetsregister for føflekkreft
Nasjonalt kvalitetsregister for brystkreft
Nasjonalt kvalitetsregister for prostatakreft
Nasjonalt kvalitetsregister for tykk- og endetarmskreft
Nasjonalt register for ablasjonsbehandling og elektrofysiologi i Norge (ABLA NOR)
Norsk register for invasiv kardiologi (NORIC)
Norsk hjertesviktregister
Norsk pacemaker- og ICD- register
Nasjonalt kvalitetsregister for gynekologisk kreft
Norsk register for gastrokirurgi (NoRGast)
Nasjonalt kvalitetsregister for behandling av spiseforstyrrelser (NorSpis)
5
Information - sample 1
Will you inform the sample about processing their personal data?
No
How?
Written information (on paper or electronically) Oral information
Information should be given in writing or electronically. Only in special cases is it applicable to give oral information, if a participant asks for this. See what you must give information about.
Upload information letter
Upload copy of oral information
No
Explain why the sample will not be informed about the processing of their personal data.
Because International Large-scale Assessment data (in this case PISA data) is completely anonymized and is available to the public, and no personal data is processed for this project.
Third persons
No
Documentation
Total number of data subjects in the project
(Data subjects: persons whose personal data you will be processing)
• 100.000+
How can data subjects get access to their personal data or how they can have their personal data corrected or deleted?
The data is anonymized, so they can’t retrieve their data and I don’t hold responsibility in that regard.
Other approvals
Will you obtain any of the following approvals or permits for the project?
No
• Ethical approval from The Regional Committees for Medical and Health Research Ethics (REC)
• Confidentiality permit (exemption from the duty of confidentiality) from the Regional Committees for Medical and Health Research Ethics (REC)

• Approval from own management for internal quality-assurance and evaluation of health services (intern kvalitetssikring) (The Health Personnel Act § 26)
• Confidentiality permit (exemption from the duty of confidentiality) from the Norwegian Directorate of Health, for quality-assurance and evaluation of health services
6
(kvalitetssikring) (The Health Personnel Act § 29b) Biobank – approval for?
• Confidentiality permit (exemption from the duty of confidentiality) from Statistics Norway (SSB) Statistics Norway has the authority to grant a confidentiality permit for the data that they manage, e.g. data about population, education, employment and social security.
• Approval from The Norwegian Medicines Agency (Statens legemiddelverk, SLV) E.g. for a clinical drugs trial
• Confidentiality permit (exemption from the duty of confidentiality) from a department or directorate
• Other approval E.g. from a Data Protection Officer
Indicate which approval
Upload document (oppdragsdokument)
Choose file... Upload approvals Chose file...
Processing
Where will the personal data be processed?
• Computer belonging to the institution responsible for the project
• Private device
Data collection, storing or archiving on private devices such as your own computer, mobile phone, memory stick etc. is not recommended and must be clarified with the institution responsible for the project.
Upload guidelines/approval for processing personal data on private devices Upload
Who will be processing/have access to the collected personal data?
• Project leader
• Student (student project)
Which others will have access to the collected personal data?
Will the collected personal data be made available to a third party or international organisation outside the EEA?
The data is anonymized and available to the public through OECD official website
Give the name of the institution/organisation
Give the country of the institution/organisation
On what basis will the collected personal data be transferred?
Upload necessary safeguards Choose file...
Next
7
Information Security
Will directly identifiable personal data be stored separately from the rest of the collected data (in a scrambling key)?
No
Explain why directly identifiable personal data will be stored together with the rest of the collected data.
The data is anonymized, stored, and available to the public

Which technical and practical measures will be used to secure the personal data?
The data is collected, anonymized, stored, and available to the public
Duration of project
Project period
August 2020 – November 2021
Will personal data be stored beyond the end of the project period?
• The data is anonymized, stored, and available to the public
For what purpose(s) will the collected personal data be stored?
• Other
Where will the collected personal data be stored?
• Other
Additional information
Will the data subjects be identifiable (directly or indirectly) in the thesis/publications for the project?
No
Explain why
Additional information
Other attachments
Choose file...
Send for preliminary assessment

Appendix II

The supplemental material can be found at a dedicated Google Drive folder: https://drive.google.com/drive/folders/1XAfQGy-5aawjbP0va4FBr6B92esTlUJ_?usp=sharing. It contains the R-code to prepare country specific datasets in an Mplus conform format (cf. prep.R), read the Mplus output results into R (cf. read.R), run regression analyses (cf. regress.R) and create figures (cf. fig.R). The Mplus syntax for the Factor Mixture Analysis of the BELONG scale for Albania is also available in the folder as an example (cf. Albania_FMA.inp). The same syntax was run for each of the 75 countries separately to retrieve consistent and inconsistent respondents in Mplus. The dataset with all the country-level variables that are used to investigate the research questions is also provided in the folder (cf. Country_data.xlsx).

The original PISA assessment data is available from the OECD: PISA 2018 Database at https://www.oecd.org/pisa/data/2018database/.