Regional Differential Item Function in the International Civics and Citizenship Education Study

Julian Fraillon, The Australian Council for Educational Research, fraillon@acer.edu.au
Wolfram Schulz, The Australian Council for Educational Research, schulz@acer.edu.au
John Ainley, The Australian Council for Educational Research, ainley@acer.edu.au

.

Abstract

This paper investigates patterns of differential item functioning (DIF) by country and geographical regions for the student civic knowledge test in the IEA Civic and Citizenship Education Study (ICCS). Data for 79 test items completed by approximately 140,000 students in more than 5,300 schools across 36countries were analyzed. The investigations were based on one-parameter IRT (Rasch) analyses of the international civic knowledge test items, by country and overall, to establish measures of DIF. These measures were then were used as data in a cluster analysis to investigate patterns of DIF across countries. Characteristics of groups of items showing DIF across geographical clusters of countries were then explored. The clustering of countries by item DIF appeared to relate primarily to language of testing but with an overlay of geo-cultural difference. There were some idiosyncrasies in the clustering, but two broad geographical country clusters – Europe and Latin America – were evident in the data. There were insufficient items showing DIF within the European country cluster to warrant further investigation. In the Latin American country cluster, DIF appeared to be most clearly associated with the target content and cognitive processes measured by items. A disproportionately high number of items that targeted ICCS content domain 1 (*civic society and systems)* and cognitive domain 1 (*knowing)* showed a pattern of DIF that caused them to appear to be relatively easier than expected within the cluster of Latin American countries. There were too few items to determine whether this pattern related to specific content topics. The results of these analyses support the need for further investigation of the relationship between language structures in items and item level DIF across countries. More specifically, it would be valuable to explore possible causes of the systematic DIF observed across countries in the Latin American cluster.

Keywords: *ICCS, Civic Education, comparative analysis, differential item function*

Introduction

The IEA International Civic and Citizenship Education Study (ICCS) studies the ways in which young people in lower secondary schools are prepared to undertake their roles as citizens in a wide range of countries. ICCS is the third IEA study designed to measure contexts and outcomes of civic and citizenship education and is linked to the 1999 IEA Civic Education Study (CIVED) (Amadeo, Torney-Purta, Lehmann, Husfeldt & Nikolova, 2002; Schulz & Sibberns, 2004; Torney-Purta, Lehmann, Oswald & Schulz, 2001). A central aspect of the study is the assessment of student civic knowledge which encompasses 'understanding as well what might be more conventionally thought as knowing facts' (Schulz, Ainley, Fraillon, Kerr & Losito, in press).

Differential item function (DIF) is evident in a test item when students of equal ability from different sub-groups respond differently (i.e. receive different scores) on that item (Hambleton, Swaminathan and Rogers, 1991). Evidence of systematic DIF in test items can therefore introduce a potential bias against or in favor of a given country or sub-group. Item DIF is routinely evaluated for key population sub-groups in national assessments as part of the process of item adjudication. In international surveys, item DIF is evaluated at the country level and is referred to as item-by-country interaction. Typically item-by-country interaction is investigated by comparing the international item parameters (difficulty) of each individual item with the national parameters of each item (Schulz and Sibberns, 2004, Mullis and Martin, 1998, OECD, 2009). Evidence of item-by-country interaction across a sufficient number of countries can be used as evidence for the removal of that item from the item set (Schulz and Sibberns, 2004)

This paper extends the evaluation of item-by-country interaction from empirical consideration of evidence of item DIF at the individual country level to consider patterns of DIF in clusters of countries.

The substantive evaluation of patterns of regional DIF will be based on item profiles derived from the ICCS Assessment Framework (Schulz, Fraillon, Ainley, Losito & Kerr, 2008). The contents of the ICCS international student test are defined by an assessment framework (Schulz et al, 2008). The framework describes possible test content in terms of four content domains: (civic society and systems; civic principles; civic participation; and civic identities and two cognitive domains (knowing; and reasoning and analyzing) (Schulz, Fraillon, Ainley, Losito & Kerr, 2008). Each ICCS test item can be characterized in terms of the content and cognitive process it represents. Those contents and processes will be the beginning point for analysis of the properties of items demonstrating regional (or sub-regional) patterns of DIF.

Ultimately the paper seeks to relate observable patterns of regional item DIF to the content and cognitive processes each item requires. It demonstrates the application of systematic consideration of item properties to provide insights into item characteristics that may be associated with regional DIF in the context of the international assessment of civics and citizenship.

Methodology

Participants

The ICCS main survey student data were collected from approximately 140,000 students in more than 5,300 schools across 38 countries[1] between October 2008 and May 2009. The target grade corresponded to the eighth year of schooling provided that the minimum age of students was 13.5 and overall the average age of participating students in ICCS was 14.4 years. The student sample was selected using a two stage probability proportional to size sampling methodology that is standard in international studies. In each country students were selected from intact classrooms within sampled schools.

The ICCS participating countries were from the broad geographical regions of Europe (26 countries), East Asia (5 countries) and Latin America (6 countries). New Zealand was the only participating country outside these three broad regions. The locations of countries participating in ICCS are shown in Figure 1[2].

[Take in Figure 1 about here]

The test instrument

The ICCS Civic Knowledge test administered to students comprised 79 items[3] that were typically presented as units, with some stimulus material followed by one or more items relating to a common context. Seventy-three items were multiple-choice and six items required students to respond by writing between one a four sentences, these six items were later scored by internationally trained scorers in each country. Seventeen of the multiple-choice items were secure items that had been used in the CIVED. The ICCS test was developed in English and later adapted for locally relevant conditions and translated to the

---

1 In ICCS the term 'country' is used to refer to each participating education system including the Hong Kong SAR and Flemish Belgium.

2 Taken from Schulz, Ainley, Fraillon, Kerr & Losito, in press

3 There were 80 items included in the test but only 79 of those items form the basis for the analysis.

target language (or languages) of instruction in each participating country. In total, 51 different versions of the test were administered in 31 different languages across the 38 participating countries.

The ICCS Assessment framework (Schulz, Fraillon, Ainley, Losito & Kerr, 2008) described four content and two cognitive domains for assessment. Table 1 shows the percentage of items by ICCS content and cognitive domains in the test.

[Take in Table 1 about here]

Data analysis: Overview

The data analysis involved a sequence of steps: first to scale the items; then to measure DIF by country; then to establish clusters of countries showing DIF on the same items; and finally to evaluate the properties of items that show DIF across the country clusters.

Data Analysis Step 1: Scaling

One parameter (Rasch) IRT was used for scaling the test items. A number of item fit and dimensionality analyses were conducted on the test item data as part of the scaling process (for further details of these see Schulz, Ainley & Fraillon, (forthcoming)). One multiple-choice item was removed from the scaling on the basis of these analyses so that 79 items contributed to the final scale. Two countries (the Netherlands and Hong Kong) did not satisfy the student sampling requirements for ICCS and the final scaling and subsequent analyses were conducted using data from 36 countries.

Data Analysis Step 2: Differential Item Function

The DIF analyses were completed by conducting a separate scaling of the 79 test items for each country. The item difficulties on the separate country scales and the international scale were standardised to a common mean so that the standardised item difficulties (in logits) for each individual country could be compared to the standardised item difficulties on the international scale. Differences in the standardised item difficulty of an item between the international scale and the country scale are indicative of item DIF. There is no fixed rule regarding the magnitude of this difference that should be called DIF (Zenisky, Hambleton and Robin, 2003). It is possible to conduct a chi-square significance tests based on the standardised differences, but the very large sample size in a study such as ICCS leads to relatively small absolute differences (that have little substantive meaning on the scale) being labelled as statistically significant. For the purpose of this review, and in the ICCS item analysis, a difference between the standardised country scale and ICCS scale item difficulty of 0.3 logits was deemed to be indicative of DIF for a given item in that country.

Data Analysis Step 3: Patterns of DIF

Patterns of DIF across countries were first explored using a hierarchical cluster analysis with complete linkage on the standardised differences between the country and ICCS scale difficulties for all items.

Data Analysis Step 4: Qualitative review of items showing DIF

The final aspect of the DIF analysis was to select the items that showed patterns of DIF across the clusters of selected European and Latin American countries. Each item was first classified for DIF according to one of three categories: 1) If the standardised scaled item difficulty for an item in a country was more than 0.3 logits greater than that for the ICCS scale the item was classified as being *harder in that country than overall*; 2) If the standardised scaled item difficulty in a country for an item was more than 0.3 logits less than that for the ICCS scale the item was classified as being *easier in that country than overall*; and 3) If the difference between the standardised item difficulty in a country and the ICCS item difficulty was less than 0.3 logits then the item was classified as *showing no DIF*.

This classification was then used to establish whether a *pattern* of DIF could be observed for each item. An item was judged to have a pattern of DIF if the difference between the number of countries for which the item showed DIF in one direction (harder or easier than the overall) and the number of countries for which the item showed DIF in the other direction was larger than half the total number of countries in the cluster. This criterion was established to ensure that only items for which a large proportion of DIF was evident in a single direction (and not 'cancelled out' by DIF in the opposite direction) would be selected. Ultimately this means that each item was classified as having no pattern of DIF or that the item was harder or easier for a given region. Items that showed a pattern of DIF in each direction by region were then analysed qualitatively to determine whether there were attributes in the items that may be linked to the pattern of DIF.

The attributes considered in this final analysis were: Item classification according to the ICCS assessment framework (content and cognitive process); trend or new item status; item difficulty; and response format. Classification of item *readability* was trialled both through the application of Flesh-Kinkaid readability analysis (see for example, Flesh, 1948; Kincaid et. al., 1975) to items and through a raw word count. Neither approach proved fruitful. The commonly used Flesh-Kinkaid analysis relies on the average number of words per sentence and number of syllables per word but was confounded by the use of domain specific terminology (such as 'parliament' and 'government') and the idiosyncratic sentence structures necessary for many test items. This consequently produced large variability in the measured reading level of items that had very similar superficial qualities and apparent reading loads.

Readability measures for any given item may also vary as items are translated to target languages and such variations may not be consistent across items and languages. As a consequence, readability may be better used to consider item characteristics within a given language rather than as relevant to a given item across language versions. It is quite possible that linguistic analysis of items in the target languages may inform the analysis of DIF, however they were beyond the scope of this paper. A similar preliminary exploration of patterns of the substantive content relating to item DIF was abandoned because there were insufficient items with sufficiently similar content to support analyses of this type. The ICCS content and cognitive domains were the narrowest level of substantive organiser that could be used to support item analysis of this type.

Findings and Discussion

Clustering of DIF by country

Figure 2 shows the tree diagram (dendogram) from the cluster analysis of the item DIF by country.

[Take in Figure 2 about here]

Figure 2 shows that the primary clustering factor in many cases relates to the language of testing within countries. Clear examples of this are the proximity in the clustering of Luxembourg, Switzerland, Austria and Lichtenstein in which German was a common language of testing for most students; England, New Zealand, Ireland and Malta in which English was a common language of testing most students; and Denmark, Norway, and Sweden all of which tested using languages of North Germanic origin. The influence of language origins on clustering extends broadly across the top of Figure 2, where there appears to be a cluster of European countries with Germanic languages and a second broad cluster of European countries mainly with non-Germanic languages. The clustering is not absolute and as has been observed in other international studies (see for example, OECD, 2002, pp 90-93), differential performance on items by country is attributable to a mix of culture, language and related geography that extends beyond language alone. Evidence of this interaction can be seen in three examples in Figure 2. Spain appears similar to Italy rather than to a cluster of five Latin American countries (Chile, Guatemala, Paraguay, Columbia and Mexico) that tested in Spanish. Estonia and Finland, despite linguistic and geographic proximity clustered more closely with other countries. England, Ireland and Malta, despite sharing English as a language of testing for most students, clustered with the largely non-Germanic language cluster within Europe.

Figure 2 does show evidence of two large regional clusters of countries that correspond to the ICCS regions of Europe (with only Russia not included) and Latin America (with only the Dominican Republic not included). There is no evidence of clustering by DIF for countries from East Asia shown in Figure 2. New Zealand appeared in a cluster with England (most likely on the basis of the language and cultural similarities between the countries) but was excluded from the further DIF analysis within regions.

The further analysis of DIF by region was therefore restricted to the 24 countries in the *European country* cluster and five countries in the *Latin American* country cluster.

Patterns of DIF across country clusters

Table 2 shows the number of items with patterns of DIF across the European and Latin American country clusters.

[Take in Table 2 about here]

From Table 2 it can be seen that considerably more DIF was evident across the Latin American country cluster than the European country cluster. Overall there were only three items showing a pattern of DIF across the European country cluster. The single item that was harder in the European country cluster than overall was a constructed response item that was marked locally. It should be remembered that item DIF is a relative measure of differences between subgroups. In the case of this analysis, the European cluster of countries comprises 24 of the 36 countries used to derive the cognitive knowledge scale. It is reasonable to expect that there will be relatively less DIF evident for a cluster that contributes a relatively large amount of information to the final scale. This, on its own, should not be used to dismiss the finding that there is relatively less systematic DIF across the European country cluster than across the Latin American country cluster as the relative lack of DIF may also be legitimately attributable to the characteristics of the items themselves.

In the Latin American country cluster there were 29 items showing a pattern of DIF with 12 items relatively harder on average and 17 relatively easier on average in this cluster of countries. Table 3 shows a summary of the characteristics of these 29 items.

[Take in Table 3 about here]

Across the Latin American country cluster it appears that item format does not relate to item DIF in either direction (easier or harder). All items showing DIF in either direction are multiple-choice. In the ICCS cognitive scale, the proportions of items for “level 1”, “level 2”

and “level 3” are approximately 28%, 48% and 22% respectively (Schulz, Ainley & Fraillon, forthcoming) and these proportions are similar to those for the items showing DIF either way in Table 3. This suggests that DIF for these items is not related overall to item difficulty.

The characteristic of the items that are harder in the Latin American country cluster do not show any clear pattern of difference from the overall ICCS item set. The proportions of items by content and cognitive domain, by trend status and by level are roughly the same for both these items and the overall ICCS item pool (see Table 1 for comparison).

However, there are differences in the item characteristics for the items that appear easier in the Latin American country cluster than the ICCS pool. The clearest areas of difference are in the proportions of items by content and cognitive domains. In the complete ICCS item pool, approximately 40% of *all* items, and 30% of all items that were *not* systematically easier in the Latin American country cluster, address content domain 1; *Civic Society and Systems*. From a different perspective 70% of the items that were systematically easier in the Latin American cluster address content domain 1. In other words items concerned with *Civic Society and Systems* were answered correctly relatively (meaning in relation to overall performance) more frequently in Latin America than were other items.

A second point of difference appears when comparing the overall ICCS proportion of items addressing the two ICCS cognitive domains (*knowing; and reasoning and analyzing).* In the full ICCS item set approximately 25% and 75% of items addressed cognitive domains 1 and 2 respectively. In the set of items that were not systematically easier in the Latin American country cluster these proportions were roughly 20% (*knowing*) and 80% (*reasoning and analyzing*). In contrast, approximately equal proportions of items that were systematically easier in the Latin American country cluster were from each of cognitive domains 1 and 2. In other words more of the relatively easy items for Latin America were about *knowing* and fewer were concerned with *reasoning and analyzing*.

ICCS was designed such that items addressing either content domain could be developed to address any of the four cognitive domains. That is to say, different cognitive processes could be applied to the range of ICCS content. Among the items that were relatively easier for the Latin American country cluster, there was no clear pattern of difference from the overall ICCS item pool. Although most of the items showing DIF were from cognitive domain 1 and content domain 1, there were items from the combination of cognitive domain 1 and content domain 2 (*civic principles)* and items addressing content domain 1 through cognitive domain 2. The data do not suggest that a higher proportion of items measuring both content domain 1 and cognitive domain 1 will necessarily be easier for countries in the Latin American country

cluster. In fact two items with such a profile were relatively *harder* for these Latin American countries.

There were 17 trend items in the ICCS items pool, the great majority of which addressed ICCS content domain 1 and cognitive domain 1. The relatively higher proportion of CIVED trend items (approximately 41%) that were easier in the Latin American country cluster compared to the overall proportion of items corresponds to the content and cognitive domains they measure rather than to the fact of the items being from CIVED.

Conclusion and Implications

This paper provides a first review of DIF by country in the ICCS cognitive test item set. It shows that broad patterns of DIF can be detected for clusters of countries geographically located within Europe and Latin America and that, for the Latin American cluster of countries this DIF may be associated with the content and cognitive processes addressed by the items. There was no clear pattern of item characteristics that could explain why a set of items with DIF were relatively more difficult for the Latin American country cluster.

There was some evidence to suggest that ICCS test items that related either (but not necessarily both) to content domain 1, c*ivic society and systems,* or cognitive domain 1 *knowing* are more likely than items addressing other domains to be relatively easier in the Latin American country cluster.

This preliminary exploration raised many more questions than it answered. The primary clustering of countries by language could be examined further through a more detailed language analysis of the items showing DIF at this level. This may uncover, for example, particular language or vocabulary structures that account for DIF by language sub-group. The second broad level of clustering seen in the European cluster of countries could be investigated to see whether the interaction between linguistic and cultural differences can provide an explanation for which items show DIF at this level.

It may also be possible to more deeply analyze the substance of the items that showed DIF in the Latin American country cluster. In particular it would be interesting to see whether there are common content themes that caused items to be easier (presumably because they have a stronger curriculum focus), and to look for differences in civic and citizenship education delivery between these countries and the other ICCS countries that may have led to the evidence of DIF shown in some items. Should future cycles of ICCS be conducted this may support more detailed substantive analysis of item contents from a larger total item pool.

References

Amadeo, J., Torney-Purta, J., Lehmann, R., Husfeldt, V., and Nikolova, R. (2002). *Civic knowledge and engagement: An IEA study of upper secondary students in sixteen countries*. Amsterdam: International Association for the Evaluation of Educational Achievement (IEA).

Flesch R (1948). *A new readability yardstick*. Journal of Applied Psychology. 32:221–33.

Kincaid JP, Fishburne RP, Rogers RL, Chissom BS. (1975). Derivation of new readability formulas (Automated Readability Index, Fog count and Flesch reading ease formula) for navy enlisted personnel. Research Branch Report 8-75, Millington, TN, Naval Technical Training, U. S. Naval Air Station, Memphis, TN, 1975.

Hambleton, R, Swaminathan, H & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage

Mullis, I. & Martin, M. (1998) *Item Analysis and Review*. In Martin & Kelly (Ed.) *TIMSS Technical Report (III).* Chestnut Hill: Boston College

OECD (2002). Reading for Change. Performance and Engagement Across Countries. Results from PISA 2000. Paris: OECD

Schulz, W., Ainley, J. & Fraillon, J. (Eds.) (forthcoming). *ICCS 2009 Technical Report*. Amsterdam: IEA.

Schulz, W., Fraillon, J., Ainley, J., Losito, B., & Kerr, D. (2008) *International Civic and Citizenship Education Study. Assessment framework.* Amsterdam: International Association for the Evaluation of Educational Achievement (IEA).

Schulz, W. & Sibberns, H. (2004). Scaling Procedures for Cognitive Items. In W. Schulz & H. Sibberns (Eds.) *IEA Civic Education Study. Technical Report (*pp. 69-91). Amsterdam: IEA.

Schulz, W, Ainley, J, Fraillon, J, Kerr, D, Losito, B (in press). Initial Findings from the IEA International Civic and Citizenship Education Study. IEA. Amsterdam

Schulz, W., Ainley, J. & Fraillon, J. (Eds.) (forthcoming). *ICCS 2009 Technical Report*. Amsterdam: IEA.

Torney-Purta, J., Lehmann, R., Oswald, H., & Schulz, W. (2001). *Citizenship and education in twenty-eight countries*. Amsterdam: International Association for the Evaluation of Educational Achievement (IEA).

Torney-Purta, J., Schwille, J., & Amadeo, J. A. (1999). *Civic education across countries: Twenty-four case studies from the IEA Civic Education Project*. Amsterdam: International Association for the Evaluation of Educational Achievement (IEA).

Zenisky, A. L., Hambleton, R. K., Robin, F. (2003). DIF Detection and Interpretation in Large-Scale Science Assessments: Informing Writing Practices. Educational Assessment. 9(1&2) 61-78

Fig.1: Countries participating in ICCS 2009

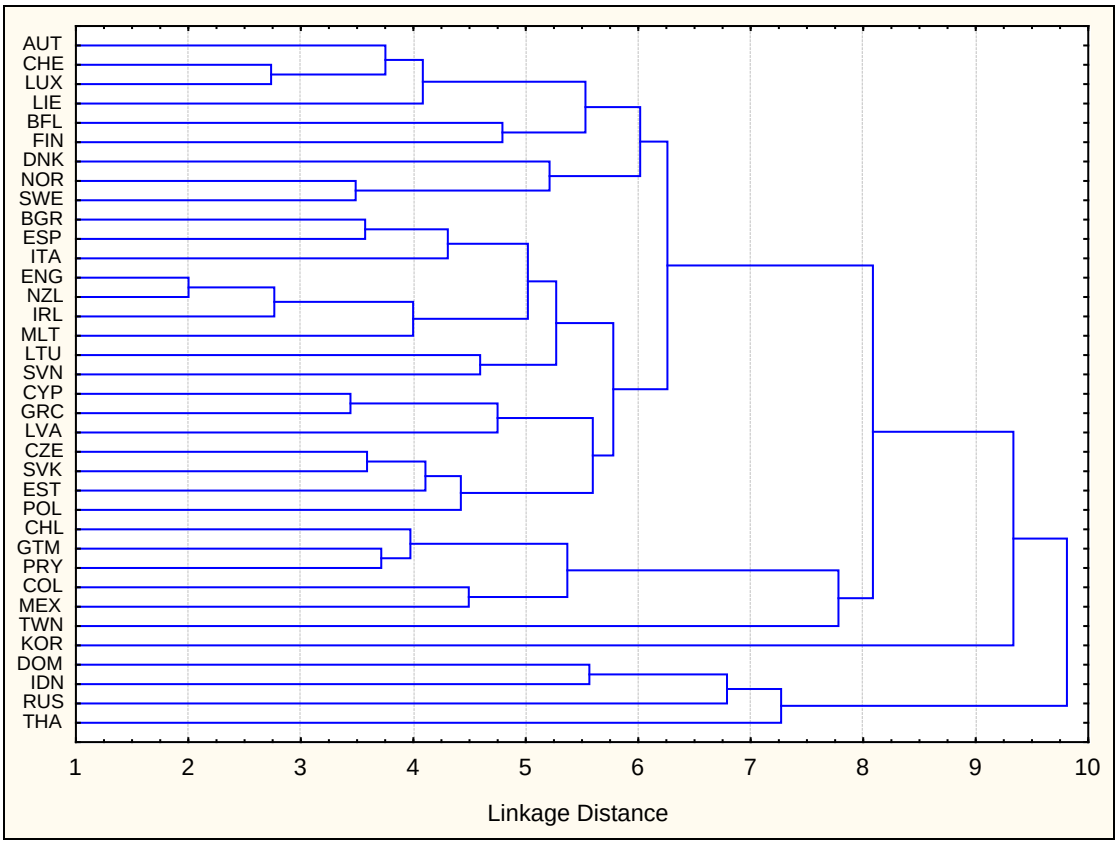

Fig. 2: Tree diagram of cluster analysis of ICCS item DIF by country

Table 1: Proportions of ICCS test item contents by ICCS content and cognitive domain

Content Domain		Cognitive Domain	
Civic society and systems	40%	Knowing	25%
Civic principles	30%	Reasoning and analyzing	75%
Civic participation	20%		
Civic identities	10%		

Table 2: Items showing patterns of DIF across country clusters

	European Cluster	Latin American Cluster
Harder across cluster	1 item	12 items
Easier across cluster	2 items	17 items
No pattern of DIF cluster	76 items	50 items

Table 3: Characteristics of items showing DIF for the Latin American country cluster

Item Characteristic		**Harder in Latin American country cluster**	**Easier in Latin American country cluster**
Question Format	*MCQ*	12	17
	CR	0	0
ICCS Content Domain	*Domain 1*	5	12
	Domain 2	3	4
	Domain 3	4	1
	Domain 4	1	0
ICCS Cognitive Domain	*Domain 1*	3	8
	Domain 2	9	9
Trend Status	*New*	10	10
	Trend	2	7
ICCS Cognitive Scale Level	*Below Level 1*	0	1
	Level 1	3	5
	Level 2	7	7
	Level 3	2	4