CHAPTER 2

Instrument Development, Sampling, Testing and Quality Control

HIGHLIGHTS RELATING TO METHODOLOGY

- A review of documents submitted by countries during the first phase of the civic education study, together with extensive item writing, pre-pilot and pilot testing, and input from country representatives resulted in the development of an instrument requiring two class-hours to administer. This instrument meets IEA's standards for psychometric quality.
- During 1999 nearly 90,000 students enrolled in the modal grade for 14-year-olds from 28 countries took the test of civic knowledge and skills, and the survey assessing concepts, attitudes and participatory actions.

International assessment in civic education has been much less frequent than testing in other content areas in comparative education. More detailed information about instrument development is therefore contained in this report than would be required in frequently tested areas.

In the first section of this chapter, we review the two-year process of identifying a common core of topics to form a content framework relating to citizenship and democracy valid across the 28 countries that participated in the civic education study. We also detail the three-year process of developing a fair and valid test (items designed with keys for correct answers) and survey (items assessing attitudes or beliefs for which there are no correct answers) to meet IEA standards.

In the next section, we describe the study's sampling. We chose the modal grade for 14-year-olds as the target population for two reasons. First, it is the standard IEA population, and it was the target population sampled in the 1971 study of civic education (Torney, Oppenheim & Farnen, 1975). Secondly, and more importantly, some National Research Coordinators noted during the development of the 1999 plans that testing an older group meant facing substantial student drop-out.

We devote the remainder of the chapter to a description of the international translation verification, testing, quality control and scaling. We present some characteristics of the achieved sample in a table, and summarize the modes of analysis and presentation. (For more detail, see the technical report of the study, Lehmann *et al.*, forthcoming.)

FRAMEWORK DEVELOPMENT DURING PHASE 1

The Phase 1 national case studies were the basis for Phase 2 of the study, in particular providing the material from which the testing framework was developed. This framework is similar to the intended curriculum on which tests in other IEA studies have been based.

The *data collected during Phase 1* included summaries of what panels of experts in participating countries believed that 14-year-olds should know about 18 topics relating to democratic institutions. These topics included elections,

individual rights, national identity, political participation, the role of the police and the military, organizations that characterize civil society, relation of economics to politics, and respect for ethnic and political diversity (Torney-Purta, Schwille & Amadeo, 1999).

Early in the study it was clear that there was a common core of topics and concepts that experts in these countries believed 14-year-olds should understand. Following examination of Phase 1 material and a vote on these topics by the National Research Coordinators, the International Steering Committee chose three domains of clustered topics as 'core international domains'. These were:

Domain I: Democracy
What does democracy mean, and what are its associated institutions and practices? The three sub-domains were:
A) Democracy and its defining characteristics
B) Institutions and practices in democracy
C) Citizenship—rights and duties.

Domain II: National Identity, Regional and International Relationships
How can the sense of national identity or national loyalty among young people be described, and how does it relate to their orientation to other countries and to regional and international organizations? The two sub-domains were:
A) National identity
B) International/regional relations.

Domain III: Social Cohesion and Diversity
What do issues of social cohesion and diversity mean to young people, and how do they view discrimination?

We also identified three other issues as important—the media, economics and local problems (including the environment)—but these were explored less systematically during Phase 2.

As a next step in developing a content framework, personnel at the Phase 1 Coordinating Center read the case study documents. They developed statements about what young people might be expected to know and believe about the three domains, and they elaborated on and illustrated these with quotations from the national case studies. This material formed the *Content Guidelines for the International Test and Survey,* which served as a concise statement of content elements in the three domains that were important across countries. The guidelines also provided a focus for those writing the test items. It was clear from the case study material that the greatest emphasis in the test should be on Domain I: Democracy, Democratic Institutions and Citizenship.

In addition to giving input on content domains to be covered, the National Research Coordinators were involved in defining the types of items to include in the instrument:

- Type 1 items: assessing *knowledge of content.*
- Type 2 items: assessing *skills in interpretation* of material with civic or political content (including short text passages and cartoons).

Types 1 and 2 items formed the *test.* These items had keyed correct answers.

Because civic education is an area where students' content knowledge and skills are important but not the sole focus, the National Research Coordinators suggested three other item types:

- Type 3 items: assessing how students understand *concepts* such as democracy and citizenship.
- Type 4 items: assessing students' *attitudes* (for example, feelings of trust in the government).
- Type 5 items: assessing students' current and expected participatory *actions* relating to politics.

Types 3, 4 and 5 items formed the survey. These items did not have correct answers.

Intersecting these five item types with the three study domains produced the following matrix, which served as the basis for the test and survey design.

Item Type:	1	2	3	4	5
Domain I Democracy/ Citizenship					
Domain II National Identity/ International Relations					
Domain III Social Cohesion/ Diversity					

A little less than half of the testing time was devoted to a test including cognitive items that could be 'keyed' with correct and incorrect answers. A little less than half of the remaining testing time was devoted to a survey including non-keyed items that assessed concepts, attitudes and actions. The rest of the instrument asked about students' perceptions of classroom climate and their confidence in participation at school, and obtained background information (including home literacy resources and the associations or organizations to which students belonged). A short period at the end of the second testing session was reserved for countries to administer nationally developed items.

THE PROCESS OF TEST AND SURVEY DEVELOPMENT DURING PHASE 2

Because there were no large existing sets of items that were likely to yield the number of items needed to fill in the matrix, extensive item writing was required. We began by reviewing materials in the Content Guidelines, other summaries of Phase 1 documents, and messages exchanged during an on-line conference on civic issues conducted with secondary school students in seven countries. We next invited all National Research Coordinators to submit items.

Our third task was to review the 1971 IEA Civic Education instrument, released items from United States and Canadian assessments, and the published research literature. Members of the International Steering Committee then wrote items, which were subsequently entered into an item database keyed to the content guidelines. Our fifth step involved asking groups of test specialists and content experts to review items in the database and their relation to the content framework.

The result of this activity was the development of 140 knowledge and skills items (Types 1 and 2), each with one correct answer and four distracters, each of which was entered into the database for the 14-year-old population. All the items were suitable for administration in the participating countries.

The items focused on principles, pivotal ideas and general examples, and not on the details of the political arrangements in any one country. For example, Type 1/Domain I items covered the principles of democracy and its associated institutions across the countries participating in the study. The test did not include items about specific mechanisms of the electoral process or government structure in any particular country. The Type 1/ Domains II and III items likewise dealt with internationally relevant or generalized matters shared across countries. This emphasis differs from that in many national tests where items about each country's political structure predominate. The IEA Civic Education Study Phase 2 items are congruent with information gathered during Phase 1 about what students are expected to know, and with recent expert statements such as that issued under the auspices of the Council of Europe about the role of history knowledge in civic education (Slater, 1995, 146–48).

Some of the Type 2 items (skills) asked students to distinguish between statements of fact and opinion. Others were based on a leaflet of the type issued during an election campaign, on the interpretation of a short article from a mock newspaper, or on a political cartoon. The general ideas for cartoons came from those published in newspapers. They were redrawn to communicate a single message that a 14-year-old across countries could be expected to understand.

Pre-Piloting of Item Types 1 and 2 (Knowledge and Skills)

Convenience samples of 14-year-olds in 20 countries were tested with 80 items of Types 1 and 2. The National Research Coordinators discussed the content of the pre-pilot items and the test statistics at a meeting held in March 1998. They agreed to retain 62 items, and prepared six items to fill gaps.

Piloting of Item Types 1 and 2 (Knowledge and Skills) and the Resulting Final Test

Between April and October 1998, 25 countries conducted pilot studies on Forms A and B of the test (Types 1 and 2 items described above) and survey (Types 3 through 5 items described below). In each country, judgement samples of about 200 students were tested (two class periods per student). The pilot countries included Australia, Belgium (French), Bulgaria, Chile, Chinese Taipei, Colombia, Cyprus, the Czech Republic, Estonia, Finland, Germany,

Greece, Hong Kong Special Administrative Region (SAR), Hungary, Italy, Latvia, Lithuania, Norway, Poland, Portugal, Romania, the Russian Federation, Slovenia, Switzerland and the United States. In addition to these countries, Denmark, England, the Slovak Republic and Sweden participated in the final testing of 14-year-olds. (Chinese Taipei was unable to obtain funding to continue past the pilot testing.)

The National Research Coordinators were provided with item statistics for their countries, and they discussed each item within its content category at a November 1998 meeting. The small number of items that was unacceptable to one-fifth of the countries was dropped, in accordance with the rule used by IEA to promote test fairness across countries. Through a process of negotiation, the research coordinators chose, by consensus, 38 items of Types 1 and 2 (knowledge and skills) from the 68 that had been piloted. The discrimination indices were greater than .30 for most items; coverage of the content framework and the research coordinators' preferences were the decisive factors.

The ratios of 'number of items written' to 'number piloted' to 'number accepted' were similar to IEA tests in other subject areas. Confirmatory factor analysis and IRT modeling, presented in Chapter 3, indicate a high-quality test across countries. These modern scaling methods (Frederiksen, Mislevy & Bejar, 1993) were our primary guide as we developed the test. Classical scaling methods also indicate a test of high quality. The alpha reliabilities for the final 38-item civic education test exceed .85 in each of the countries (see Chapter 3 and associated appendices for details).

With respect to content coverage, within Domain I there are items covering all three sub-domains (definitions of democracy 6, democratic institutions 12, citizenship in democracy 12); within Domain II there are items covering the two sub-domains (national identification 2, international relations 3); within Domain III there are three items. Appendix Table A.1 contains short descriptions of the 38 items and of the content categories in which they were classified, along with the percentage of students answering them correctly in the final test and the respective item parameters (discussed further in Chapter 3).

Piloting of Item Types 3, 4 and 5 (Concepts, Attitudes and Actions) and the Resulting Final Surveys

The National Research Coordinators reviewed lists of suggested topics for Types 3 to 5 items and some prototype items at the March 1998 meeting. Most item sets for piloting were suggested by the research literature. Some revisions were necessary to adapt items originally designed for administration to adults in an interview, and 'don't know' options were added.

In mid-1998 the research coordinators piloted the survey items along with two forms of the knowledge and skills test. Items for the survey were chosen through a process of negotiation similar to that described in the previous section. The final survey included 52 items of Type 3 (concepts), 62 items of Type 4 (attitudes) and 22 items of Type 5 (actions). Items assessing student background, school experience, organizational membership and peer group involvement were also included. Policy in some of the participating countries

prohibited questions about families' social or political values, and no such items were included. The final test and survey were designed so that they could be administered in two class periods. The texts of all of the Types 3, 4 and 5 items and of about half of the Types 1 and 2 items will be released for use by other researchers.

Chapters 4 through 7 of this publication describe the rationale for items and scales included in the student survey, along with relevant research literature.

The development of short survey instruments for teachers and for school heads (principals) began at the March 1998 meeting and covered the same content domains as the student instrument, along with questions about the school context and instruction. These instruments were piloted in the same countries and at the same time as the student instruments. The questions included and the results for the teacher survey are discussed in Chapter 9. The school questionnaire has been left for future international analysis and for national analysis.

SAMPLING, TESTING AND SCALING DURING PHASE 2

Sampling from an Internationally Defined Population

The internationally desired population was defined as follows:

> The population includes all students enrolled on a full time basis in that grade in which most students aged 14:00 to 14:11 [years; months] are found at the time of testing. Time of testing is the first week of the 8th month of the school year.

In most cases testing took place between March and June 1999 in the Northern Hemisphere and between August and October 1999 in the Southern Hemisphere. In England and Sweden, testing was conducted in the second or third month of the school year because of the countries' late entry into the study. In the United States the testing was done in the second month of the school year because of uncertainty as to the age distribution of students in the eighth month of the year (resulting from the varying school entry dates set by districts).

In the majority of countries, Grade 8 was selected. In nine countries, Grade 9 was chosen. In Switzerland, differences between regions led to the selection of Grades 8 or 9, depending on the structure of the educational system. In Portugal, Grade 8 was selected even though the proportion of 14-year-olds in this country tends to be slightly higher in the adjacent Grade 9. The average age of respondents in the selected Grade 8 was 14:5, which was similar to the average age in most other countries in this study. If Grade 9 had been used in Portugal, the average age would have been 15:4.

In two countries (Hong Kong/SAR and the Russian Federation), the average age was above 15:00 and therefore did not meet the study's age/grade specifications. In two countries (Belgium/French and Chile), the average age was between 14:00 and 14:11, but the proportion of 13-year-old students in the tested grade ended up being slightly higher than the proportion of 14-year-old students.

In Germany, three federal states ('Bundesländer') refused to participate in this study, and one federal state did not permit testing in high schools ('Gymnasium'). Therefore, the sample was not representative for the population of *all* 14-year-old students in this country but only for those in the participating federal states.

A two-stage stratified cluster design for sampling was employed in consultation with IEA sampling experts. At the first stage, schools were sampled using a probability proportional to size (PPS).[1] At the second stage the sample consisted of one intact classroom per school from the target grade. The chosen class was not to be tracked by ability and was, where possible, to be in a civic-related subject (for example, history, social studies).

Table 2.1 shows the participation rates of the 28 countries. National Research Centers made every attempt to meet the sampling requirements, but in some countries there was resistance from teachers and schools. Ten countries failed to reach a 75 percent overall participation rate before replacement as specified

Table 2.1 Participation Rates and Sample Sizes

Country	School Participation Before Replacement (Weighted Percentage)	School Participation After Replacement (Weighted Percentage)	Total Number of Schools That Participated	Student Participation Rate	Total Number of Students Assessed	Overall Participation Rate Before Replacement	Overall Participation Rate After Replacement
Australia	75	94	142	92	3331	69	86
Belgium (French)	57	75	112	93	2076	53	70
Bulgaria	86	86	148	93	2884	80	80
Chile	98	100	180	97	5688	94	97
Colombia	66	94	144	96	4926	64	90
Cyprus*	100	100	61	96	3106	96	96
Czech Republic	91	99	148	95	3607	86	94
Denmark	71	71	178	93	3208	66	66
England	54	85	128	93	3043	50	79
Estonia	84	85	145	90	3434	76	77
Finland	93	98	146	93	2782	86	91
Germany	63	94	169	89	3700	56	84
Greece	88	93	142	97	3460	85	90
Hong Kong (SAR)	90	100	150	99	4997	89	99
Hungary	99	99	146	95	3167	94	94
Italy	93	100	172	96	3808	89	96
Latvia	89	91	130	91	2572	81	82
Lithuania	93	97	169	90	3494	84	87
Norway	75	77	154	93	3321	70	71
Poland	83	90	179	94	3376	78	84
Portugal	98	99	149	95	3261	93	95
Romania	97	97	146	99	2993	96	96
Russian Federation	96	98	185	97	2129	94	95
Slovak Republic	79	97	145	94	3463	74	91
Slovenia	93	99	149	96	3068	89	95
Sweden	93	94	138	94	3073	88	88
Switzerland	71	87	157	97	3104	69	84
United States	65	83	124	93	2811	61	77

* In Cyprus two classes per school were sampled.

Source: IEA Civic Education Study, Standard Population of 14-year-olds tested in 1999.

in the sampling guidelines. In Belgium (French), Denmark and Norway, the overall participation rate, even after replacement of schools, was lower than 75 percent.[2] Student participation rates were at least 89 percent in all participating countries, however.

Sample sizes of schools per country varied between 112 and 185. In Cyprus, all 61 schools in the country were tested, and from each school two classes were sampled. Student sample sizes ranged between 2,076 and 5,688. In some countries, disproportional samples were drawn (for example, to include larger sub-samples of specific school types). Sampling weights were applied.

Table 2.2 summarizes three basic characteristics of the sample by country: mean and standard deviation of the age of students tested, the percentage of females, and the percentage of students who answered that they had not been born in the country. The age distribution is discussed in Chapter 3. The most serious gender disparity was in Colombia, where 58 percent of the student

Table 2.2 Sample Characteristics

Country	Age			Percentage of Females	Percentage of Students Not Born in Country
	Mean	Standard Deviation	Percentage of 14-year-olds		
Australia	14.6	0.5	67	55	10
Belgium (French)	14.1	0.7	34	49	10
Bulgaria	14.9	0.6	59	52	4
Chile	14.3	0.8	40	49	2
Colombia	14.6	1.2	35	58	3
Cyprus	14.8	0.4	75	51	9
Czech Republic	14.4	0.4	70	51	2
Denmark	14.8	0.4	66	49	7
England	14.7	0.3	79	50	6
Estonia	14.7	0.6	67	52	6
Finland	14.8	0.3	67	52	3
Germany[1]	n.a	n.a	n.a	51	19
Greece	14.7	0.5	83	52	6
Hong Kong (SAR)	15.3	0.8	38	49	20
Hungary	14.4	0.5	70	50	3
Italy	15.0	0.7	58	52	2
Latvia	14.5	0.6	62	52	5
Lithuania	14.8	0.6	67	51	3
Norway	14.8	0.3	71	51	6
Poland	15.0	0.4	54	52	1
Portugal	14.5	1.0	35	52	5
Romania	14.8	0.5	65	48	1
Russian Federation	15.1	0.5	48	53	14
Slovak Republic	14.3	0.4	69	53	2
Slovenia	14.8	0.4	74	50	4
Sweden	14.3	0.4	79	52	8
Switzerland	15.0	0.7	55	51	17
United States	14.7	0.6	74	51	11
International Sample	14.7	0.7	62	51	7

1 Information on age is not available for Germany. International sample figures based on 27 countries.

Source: *IEA Civic Education Study,* Standard Population of 14-year-olds tested in 1999.

respondents were female. The proportion of students not born in the country ranged from 1 percent (in Poland and Romania) to 19 and 20 percent in Germany and Hong Kong (SAR), respectively. This matter is discussed further in Chapter 5.

Instrument Translation

The pilot and final instruments were prepared in English and distributed by the International Coordinating Center (ICC). The National Research Centers then translated them into 22 languages. The ICC developed guidelines and detailed translation notes indicating alternative wordings adapted to the country's specific context.

Translated instruments for the final testing had to be submitted to the ICC for verification. Native speakers with a very good command of English and no working relationship with the National Research Centers verified the translations of school, teacher and student instruments according to guidelines issued by the ICC. The results of this verification were returned to the National Research Coordinators. In most countries, suggestions for improvement were taken into account in the final translations. In only three countries were translations not submitted in time for this process to take place. However, in these cases, the verifications did give some after-the-fact control over deviations (which were, in fact, few in number).

This process together with the translation verification of the pilot instruments in 1998 provided a high degree of quality control in this area. For 24 of the 28 countries the instruments were verified twice. For the four countries that entered the study after the pilot, there was only one verification. Instruments from English-speaking countries did not require translations but were reviewed for modifications necessary to adapt them to each country's political and cultural context.

Data Collection and Quality Control for Testing

Each participating country was responsible for data collection. Manuals for field operations, the school coordinators and the test administrators, together with tracking forms were adapted by the IEA Data Processing Center from those developed for TIMSS. The distribution of this material to the National Research Centers was carried out with the cooperation of the ICC. Where necessary, the manuals were translated into the country's language. Data collection at the schools followed strict guidelines for test administration and timing to safeguard comparability across countries. Full confidentiality of responses was guaranteed. Data entry was conducted by the National Research Centers.

The National Research Coordinators were asked to make follow-up calls on the day after testing to a 25 percent random sample of the tested schools. They were instructed to ask about deviations from testing procedures (using guidelines provided by the ICC). In a few countries, organizational problems made this task impossible, but every effort was made to examine these data with special care. In some countries, the national centers set up additional control monitoring.

After completion of the testing, the National Research Centers responded to a questionnaire on quality control that could be used as an additional check. Most centers completed this questionnaire within one month of testing.

Data Processing and Weighting

After collection, the data sets were submitted in a standard format to the IEA Data Processing Center (DPC) in Hamburg, which created the international database for the study. The DPC compared the database to the school, classroom, teacher and student tracking forms completed during data collection. They also checked and double-checked the data for inconsistencies. All deviations were documented and sent to the National Research Centers for clarification. The data-cleaning process consisted of several steps designed to guarantee high quality. The DPC also computed the weights to be applied to the sample according to the previously approved sampling design in each country (in line with IEA guidelines).

Confirmatory Factor Analysis and IRT Scaling

Structural equation modeling (SEM), including confirmatory factor analysis (CFA), was used to confirm theoretically expected dimensions or to re-specify the dimensional structure of the instruments. These procedures take into account the measurement error associated with indicators, providing more reliable estimates for latent variables and scales than classical psychometric methods.[3]

For both multiple-choice and categorical items, item response theory (IRT) scaling methods were used. For the cognitive test, a one-parameter Rasch model was fitted to the data; for attitudinal items the partial credit model was applied. The mean Rasch score for the scores derived from the test was set at 100, with a standard deviation of 20. The mean Rasch score for the scales derived from the survey (measuring concepts, attitudes and actions) was set at 10, with a standard deviation of 2.

There were several reasons for using IRT scaling. In this study all students were administered exactly the same 38-item test, so IRT scaling was not absolutely necessary (as it would have been if there had been a larger set of items from which several test forms had been constructed). In the case of the test items, however, the Rasch method provided a common scale for all countries, allowing the exclusion of items that did not fit the model in a few countries and without jeopardizing the comparability of the international scale. This method is prescribed in IEA studies.

Attitudinal items that had missing values, resulting from students who answered ‘don’t know’ or who left items out, was a potential problem. Here, IRT scaling provided an elegant way of computing estimates for latent dimensions, even those with missing information.

SUMMARY OF ANALYSIS

The analysis of the international data consisted of several steps:

- Computation of item statistics for all test and survey items.
- Exploratory factor analysis and computation of classical scale reliabilities for the theoretically expected scales for each country.
- Confirmatory factor analysis with structural equation modeling on an international random sample of 200 students from each country, followed by the checking of models for each country.
- Selection of scales based on theoretical and empirical grounds.
- Estimation of Rasch models for the selected scales on an international random sample of 200 students per country.
- Item adjudication to examine scales by country and to make further refinements.

The final scaling used a calibration sample of 500 students selected randomly from weighted country data. Item parameters were estimated for the calibration sample and used as anchors for subsequent scaling of country data sets.

Complex sampling (such as the multi-stage sampling used here) makes simple random sampling formulas for estimating standard errors inadequate. In order to estimate correct sampling errors for each statistic in this report, we applied the 'jack-knife' procedure. The overall estimate of a sample statistic plus or minus two standard errors gives a 95 percent probability of inferring the correct mean in the population based on the student sample.

GUIDE TO THE PRESENTATION OF DATA FOUND IN CHAPTERS 3–7

Many of the scales and items in the test and the survey were derived from previous research and, in some cases, had been the subject of extensive debate and empirical study by political or educational researchers, sociologists and psychologists. An attempt has been made in the panels in Chapters 3–7 to briefly review the methods and a selection of the findings of previous cross-national research, especially that conducted in countries participating in this study.

The 38 multiple choice items in the *test of knowledge and skills have correct answers.* The IRT scaling process for these items is covered in Chapter 3. The Rasch scores presented in this chapter were normed to have an international mean of 100 and a standard deviation of 20. The scores based on these items are presented in the same format as in many other IEA studies, with the countries in rank order by score.

The items in the *survey of concepts, attitudes and actions do not have correct answers.* These results are presented in Chapters 4–7. The large majority of items were statements to which the student was to respond on a four-point scale with an additional 'don't know' option. The labeling of scale points differed. For many scales they were 'strongly agree', 'agree', 'disagree', 'strongly disagree'. Other response formats asked about how important something was or how frequently something happened.

The examination of confirmatory factor analyses by the International Steering Committee led to the choice of 11 sets of survey items for scaling. An IRT scaling procedure was applied at the International Coordinating Center to each set of items, and the resulting Rasch scales were normed to have an international mean of 10 and a standard deviation of 2. These scores allow statistically sound comparisons between countries' means and the international mean, as well as between one country's mean and that of another. Appendix B includes an item-by-score map for each concept, attitude or action scale in Chapters 4–7, allowing the reader to ascertain the response on the four-point scale that corresponds to Rasch scores from 4 to 16. Although country means were in the range from 8 to 12, there were student respondents in each country with scores well below 8, and others that were well above 12. The Appendix B material also includes the percentage distribution of responses to each scaled item based on the entire group of countries.

The figures in Chapters 4–7 present country means on these scaled scores. All countries appear in alphabetical order, with a confidence bound for each mean of two standard errors. An upward or downward arrow also appears to indicate whether a country's mean is significantly higher or lower than the international mean. Chapter 10 summarizes these differences across all scales and countries.

A figure comparing country means is included for each of the Rasch scores in Chapters 4–7. An additional figure is included to illustrate gender differences only for those scales where half or more countries show a significant gender difference ($p < .05$ with a Dunn-Bonferroni correction for multiple comparisons). If fewer than half of the countries show a significant difference, the text lists the gender differences that are statistically significant at the .05 level, but no separate figure is included.

SUMMARY OF METHODS

The following were used to develop the two class-hours of the 1999 IEA Civic Education Study test and survey:

- an iterative process of review of Phase 1 documents submitted by countries;
- references to the research and theoretical literature;
- extensive item writing;
- review by experts internationally and within participating countries;
- pre-pilot and pilot testing;
- item choice by participating countries.

The test and survey was administered to nationally representative samples totaling nearly 90,000 14-year-old students in 28 countries. Confirmatory factor analysis and Rasch scaling were used to develop scales. Much of the data is presented in this volume in figures that allow an analysis of countries' positions significantly above, not significantly different from, or significantly below the international mean.

A similar process was undertaken for the development of the Teacher Questionnaire and a very short School Questionnaire (covered in Chapter 9).

Quality control procedures were undertaken, including the review of samples by sampling experts, two translation verifications by independent experts, and other measures prescribed by the IEA technical standards. Panel 2.1 presents a listing of these quality control procedures.

PANEL 2.1 Quality Control Processes

Control Processes Anchored in the Study's Conceptualization

The study's design, instruments and reports have been:

- Connected to 15 policy questions formulated to guide Phases 1 and 2.
- Referenced to 18 case study framing questions in Phase 1.
- Framed by the 'octagon model' and 'situated cognition perspective' in Phases 1 and 2.
- Anchored in Phase 1 country case study data (leading to definition of the three domains for Phase 2).
- Scaffolded by content guidelines (including quotations from Phase 1 documents).
- Organized under three domains forming the test and survey framework.
- Referenced to the research literature on political attitudes in youth and adults and to theories of democracy.
- Built using five item types matched to country expectations identified during Phase 1.

Control Processes Relating to IEA Standards and Participating Country Input

The study's design, instruments and reports have been:

- Guided and judged by IEA technical standards and procedures (for example, regarding sampling and testing).
- Influenced by participating countries' input (through National Expert Panels and National Research Coordinators forming a de-centered network for test adaptation).
- Shaped by the analysis of pre-pilot psychometric data for test items from 20 countries interpreted by National Research Coordinators.
- Shaped by the analysis of pilot psychometric data for test and survey from 25 countries and interpreted by National Research Coordinators.
- Shaped in meetings between National Research Coordinators and Data Processing Center Personnel (regarding sampling, weighting of samples and data submission).
- Reviewed periodically by the IEA Technical Executive Group.
- Informed by independent verification of test translations and of concept equivalence (of the pilot and the final test and survey).
- Monitored using National Research Coordinators' reports regarding the testing process.
- Referenced to an analysis plan guided by policy questions and IEA principles.
- Finalized through the International Steering Committee's review of Rasch scaling for test and survey items and choice of scales to be reported.

NOTES

1 The general procedure followed closely the one adopted for the Third International Mathematics and Science Study (TIMSS) as described in Foy, Rust and Schleicher (1996). In some countries, the sample for the Civic Education Study was linked to the TIMSS-R (Repeat), which was done in the same year as the Civic Education Study.

2 Originally sampled schools that refused to participate could be replaced by additionally sampled schools. In Denmark no replacement schools were sampled.

3 We have included some classical psychometric indices in appendices because some readers may be more familiar with them.