1. Assessment

1.1 Defining assessment

A definition of assessment offered by Rowntree (1987) focuses upon getting to know the person:

> ... assessment can be thought of as occurring whenever one person, in some kind of interaction, direct or indirect, with another, is conscious of obtaining and interpreting information about the knowledge and understanding, or abilities and attitudes of that other person. To some extent or other it is an attempt to know that person.

A definition of assessment offered by Sadler (in Joughin (2010)) has three key parts:

> The act of assessment consists of appraising the quality of what students have done in response to a set task so that we can infer what students can do, from which we can draw an inference about what students know.

A definition of assessment offered by Joughin (2010):

> To assess is to make judgements about students' work, inferring from this what they have the capacity to do in the assessed domain, and thus what they know, value, or are capable of doing.

1.1.1 Some questions

1. Discuss each of these definitions. How are they similar and how do they differ?
2. Do you agree or disagree with these definitions? Give reasons?
3. How is assessment defined in Nepal? Where is it defined?

4. What is your definition of assessment? Write it and share it with a colleague. Solicit critical feedback: what are the implications of your definition for teaching and learning.

1.2 Assessment of, as and for learning

The New South Wales Board of Studies Teaching and Standards offers the following statements:

> Assessment for learning involves teachers using evidence about students' knowledge, understanding and skills to inform their teaching. Sometimes referred to as 'formative assessment', it usually occurs throughout the teaching and learning process to clarify student learning and understanding.
>
> Assessment for learning:
>
> - reflects a view of learning in which assessment helps students learn better, rather than just achieve a better mark
> - involves formal and informal assessment activities as part of learning and to inform the planning of future learning
> - includes clear goals for the learning activity
> - provides effective feedback that motivates the learner and can lead to improvement
> - reflects a belief that all students can improve
> - encourages self-assessment and peer assessment as part of the regular classroom routines
> - involves teachers, students and parents reflecting on evidence is inclusive of all learners.
>
> Assessment as learning occurs when students are their own assessors. Students monitor their own learning, ask questions and use a range of strategies to decide what they know and can do, and how to use assessment for new learning.
>
> Assessment as learning:
>
> - encourages students to take responsibility for their own learning
> - requires students to ask questions about their learning
> - involves teachers and students creating learning goals to encourage growth and development
> - provides ways for students to use formal and informal feedback and self-assessment to help them understand the next steps in learning
> - encourages peer assessment, self-assessment and reflection.
>
> Assessment of learning assists teachers in using evidence of student learning to assess achievement against outcomes and standards. Sometimes referred to as 'summative assessment', it usually occurs at defined key points during a unit of work or at the end of a unit, term or semester, and may be used to rank or grade students. The effectiveness of assessment of learning for grading or ranking depends on the validity and reliability of activities. Its effectiveness as an opportunity for learning depends on the nature and quality of the feedback.
>
> Assessment of learning:
>
> - is used to plan future learning goals and pathways for students

- provides evidence of achievement to the wider community, including parents, educators, the students themselves and outside groups
- provides a transparent interpretation across all audiences.

1.2.1 Some questions

1. Discuss each of these types of assessment. How are they similar and how do they differ?
2. Which of these types of assessment would you observe in a Nepal classroom?
3. What training do teachers get in each of these types of assessments? When do they get it and how effective is the training?

Table 1.1: Assessment for and of learning: selected differences from Chappuis (2002) and Stiggins (2007)

	Assessment for learning	Assessment of learning
Reasons for assessing		
Audience for results		
Focus of assessment - learning targets		
Place in time		
Primary users		
Typical uses		
Teacher's role		
Students' role		
Primary motivator for students		
Examples		

Table 1.2: Assessment for and of learning: selected differences

	Assessment for learning	Assessment of learning
Reasons for assessing	Promote increases in achievement to help students meet more standards; support ongoing student growth; improvement	Document individual or group achievement or mastery of standards; measure achievement status at a point in time for purposes of reporting; accountability
Audience for results	Students about themselves	Others about students
Focus of assessment - learning targets	Specific achievement targets selected by teachers that enable students to build toward standards	Achievement standards for which schools, teachers, and students are held accountable
Place in time	A process during learning	An event after learning
Primary users	Students, teachers, parents	Policy makers, program planners, supervisors, teachers, students, parents
Typical uses	Provide students with insight to improve achievement; help teachers diagnose and respond to student needs; help parents see progress over time; help parents support learning	Certify student competence; sort students according to achievement; promotion and graduation decisions; grading
Teacher's role	Transform standards into classroom targets; inform students of targets; build assessments; adjust instruction based on results; offer descriptive feedback to students; involve students in assessment	Administer the test carefully to ensure accuracy and comparability of results; use results to help students meet standards; interpret results for parents; build assessments for report card grading
Students' role	Self assess and keep track of progress; contribute to setting goals; act on classroom assessment results to be able to do better next time	Study to meet standards; take the test; strive for the highest possible score; avoid failure
Primary motivator for students	Belief that success in learning is achievable	Threat of punishment; promise of rewards
Examples	Using rubrics with students; student self-assessment; descriptive feedback to students	Achievement tests; final exams; placement tests; short cycle assessments

1.3 Assessment washback effect

Biggs and Tang (2011) note that "students learn what they think they will be tested on" (p. 182).

Gipps (2012) writes:

> 'Washback' on teaching and the curriculum are long-established consequences of assessment, particularly high-stakes testing. These consequences may be sound educationally, such as encouraging the teaching of a broader range of skills, or negative as in a narrowing tendency to teach to the test.

1.3.1 Some questions

1. Have you noticed any negative washback effects in Nepal? If so, describe them and their effects?
2. Have you noticed any positive washback effects in Nepal? If so, describe them and their effects?
3. How might the positive effects be accentuated and the negative effects diminished?

1.4 Frequency of assessments

Marzano (2007) argues that a subject curriculum should consist of 15-20 topics and teachers should conduct assessments at least twice a week.

Table 1.3: Gain Associated with Number of Assessments over 15 Weeks

Number of assessments	Effect size	Percentile point gain
0	0	0
1	.34	13.5
5	.53	20.0
10	.60	22.5
15	.66	24.5
20	.71	26.0
25	.78	28.5
30	.80	29.0

1.4.1 Some questions

1. How frequently do Nepalese teacher assess their students?
2. Would the above results hold for Nepal? For all grades and subjects? Why? Or why not?

1.5 Large scale assessment policy

The British Columbian government has policies, practices and guidelines for much of education. One such policy is about large scale assessments.

Background on Large-Scale Assessments

The term large scale assessment refers to any provincial, national or international assessment, examination or test the Ministry directs boards of education to administer.

A primary purpose of large-scale provincial, national and international assessments and examinations is to obtain information for the purposes of public accountability, improving programs and certifying that students meet graduation requirements.

Large-scale assessments and examinations determine students' knowledge and skills in particular areas of learning. Assessment information enables educational decision-makers at the classroom, district and provincial levels to make informed choices related to improving student achievement.

British Columbia students, teachers, principals, and other school officials must participate in the Foundation Skills Assessment, Graduation Program Examinations, and national and international assessments as required by the Minister.

The Ministry of Education provides test specifications and sample test questions for all provincial assessments and examinations. The Ministry may designate as secure specific assessments and examinations, or certain parts of the assessments and examinations.

1.5.1 Some questions

1. How does the British Columbian government define large-scale assessment? How are large-scale assessment defined in Nepal? Give some examples of large-scale Nepalese assessments.
2. According to the British Columbian government, are examinations large-scale assessments or are they different from large scale assessments?
3. What are the purposes of these large-scale assessments according to the British Columbian government? What are the purposes of the Nepalese large-scale assessments?
4. Does the British Columbian government compell people to participate in these assessments and, if so, how does the government help people prepare for the assessments?
5. Are people compelled to participate in Nepal's large-scale assessments and, if so, how are they prepared for the assessments?
6. What is the legal basis for large scale assessments in Nepal?

1.6 Large-scale assessments and examinations

While some experts would argue that examinations are a specific form of large-scale assessment, Greaney and Kellaghan (2007) argue that examinations and large-scale assessment provide different information.

> Sometimes, public examinations are thought to provide the same information as a national assessment, thus appearing to eliminate the need for a national assessment system in a country that has a public examination system. However, public examinations cannot provide the kind of information that a national assessment seeks to provide. (p. 14)

They then go onto to describe the differents between national large-scale assessments and examinations (p. 15)

Table 1.4: Differences between National Assessments and Public Examinations

	National assessments	Public examinations
Purpose	To provide feedback to policy makers.	To certify and select students.
Frequency	For individual subjects offered on a regular basis (such as every four years).	Annually and more often where the system allows for repeats.
Duration	One or two days.	Can extend over a few weeks.
Who is tested?	Usually a sample of students at a particular grade or age level.	All students who wish to take this examination at the examination grade level.
Format	Usually multiple choice and short answer.	Usually essay and multiple choice.
Stakes: importance for students, teachers, and others	Low importance.	Great importance.
Coverage of curriculum	Generally confi ned to one or two subjects.	Covers main subject areas.
Effect on teaching	Very little direct effect.	Major effect: teacher tendency to teach what is expected on the examination.
Additional tuition sought for students	Very unlikely.	Frequently.
Do students get results?	Seldom.	Yes.
Is additional information collected from students?	Frequently, in student questionnaires.	Seldom.
Scoring	Usually involves statistically sophisticated techniques.	Usually a simple process that is based on a predetermined marking scheme.
Effect on level of student attainment	Unlikely to have an effect.	Poor results or the prospect of failure, which can lead to early dropout.
Usefulness for monitoring trends in achievement levels over time	Appropriate if tests are designed with monitoring in mind.	Not appropriate because examination questions and candidate populations change from year to year.

1.6.1 Some questions

1. What information is provided by Nepal's examinations? How is it used?
2. What information is provided by Nepal's large-scale assessments? How is it used?

1.7 India, national assessments and examinations

1.7.1 Limitations of examinations

The traditional Indian examination structure does not suffice to track learning outcomes on a systemic level, because:

- The purpose of internal school assessment is to evaluate the achievement of individual students and not the system as a whole.
- The focus, design and difficulty of these examinations varies greatly and does not take into account background factors that may impact learning.
- Common board examinations are conducted only in Class 10 and Class 12, which is the end of a child's schooling career. The results are 'high-stakes' for the students because they determine future course of study or employment. Hence, examinations are designed to allow the maximum number of students to qualify, and not specifically to distinguish between them.
- Students take a different set of question papers each year, with no unifying rubric to allow comparison of scores across years.

1.7.2 Reasons for national assessments

There are three primary objectives of a large-scale assessment:

- Evaluation – Large-scale assessments are often a major monitoring mechanism for a system. Monitoring and evaluation refers to collecting and analysing data to check performance against goals and to take remedial actions if needed.
- Accountability – Where assessment is used to hold any part of the system accountable, there need to be clear consequences of the evaluation.
- Improvement – Countries utilize assessment results for formative purposes, providing feedback to teachers on specific student performance. In this case, the results must be presented to teachers, school leaders and government officials in a meaningful manner, such that they can be readily utilized.

1.7.3 Some questions

1. Why are internal assessments unlikely to provide information for systemic improvement?
2. What information is required for systemic improvement?
3. Do you agree that examinations do not provide the information required for systemic improvement?
4. What objectives should a Nepalese national assessment meet? Why?

1.8 Conditions to improve learning using large scale assessments

From Ungerleider (2006):

> Any jurisdiction contemplating the use of large-scale student assessment to improve student achievement:
>
> - establish broad agreement about what school outcomes are essential for all students;
> - ensure that these areas are clearly articulated in the curriculum and are supported with appropriate instructional material;
> - hold students, parents, and teachers accountable for those outcomes;

- assess student progress in the areas of importance at different times over their school careers;
- prepare teachers and encourage them to use teaching strategies that increase learning outcomes for all students;
- encourage mixed-ability grouping and discourage grouping, tracking, or streaming students by socio-economic background or in ways that increase differentiation among students of different ethno-cultural backgrounds;
- assess schools on the basis of student growth in learning outcomes, taking into account their individual socio-economic backgrounds, the socio-economic context of the school community, as well as school policies and practices known to influence the achievement of the valued outcomes;
- examine rates of student progress as well as gradients in student progress associated with such background factors as socioeconomic standing, gender, and ethnicity;
- ensure that teachers and administrators are well prepared for their responsibilities;
- counter misuse of the results of large-scale assessments in the media and elsewhere; and
- provide teachers with adequate time to individually and collectively interpret data for the purpose of improving instruction.

1.8.1 Some questions

1. Do you agree or disagree with these conditions?
2. How are Nepal's large-scale assessments used to improve learning?
3. What would be your conditions for using large-scale assessments to improve learning in Nepal?

1.9 Main elements in a national assessment

Greaney and Kellaghan (2007) present the main elements in a national assessment on pages 12-14. This is high-level view and more detail is provided by the authors elsewhere. These elements are:

Ministry of Education elements

- The ministry of education (MOE) appoints either an implementing agency within the ministry or an independent external body (for example, a university department or a research organization), and it provides funding.
- The MOE determines policy needs to be addressed in the assessment, sometimes in consultation with key education stakeholders (for example, teachers' representatives, curriculum specialists, business people, and parents).
- The MOE, or a steering committee nominated by it, identifi es the population to be assessed (for example, fourth grade students).
- The MOE determines the area of achievement to be assessed (for example, literacy or numeracy).

Implementing agency elements

- The implementing agency defines the area of achievement and describes it in terms of content and cognitive skills.
- The implementing agency prepares achievement tests and supporting questionnaires and administration manuals, and it takes steps to ensure their validity.

- The tests and supporting documents are pilot-tested by the implementing agency and subsequently are reviewed by the steering committee and other competent bodies to (a) determine curriculum appropriateness and (b) ensure that items reflect gender, ethnic, and cultural sensitivities.
- The implementing agency selects the targeted sample (or population) of schools or students, arranges for printing of materials, and establishes communication with selected schools.
- The implementing agency trains test administrators (for example, classroom teachers, school inspectors, or graduate university students).

Schools

- The survey instruments (tests and questionnaires) are administered in schools on a specified date under the overall direction of the implementing agency.

Implementing agency elements

- The implementing agency takes responsibility for collecting survey instruments, for scoring, and for cleaning and preparing data for analysis.
- The implementing agency establishes the reliability of the assessment instruments and procedures.
- The implementing agency carries out the data analysis.
- The draft reports are prepared by the implementing agency and reviewed by the steering committee.
- The final reports are prepared by the implementing agency and are disseminated by the appropriate authority.

Ministry of Education elements

- The MOE and other relevant stakeholders review the results in light of the policy needs that they are meant to address and determine an appropriate course of action.
- And there is typically press releases and media presentations.

1.9.1 Some questions

1. Nepal has conducted a number of national assessments. What parts of the above hold true, what needs to be deleted and what should be changed?
2. Looking back at the need to improve assessment in schools, how might the use of school personnel be maximimzed in these steps?

1.10 Developing an assessment framework

The PISA draft mathematics assessment framework has the following statement:

> The PISA 2015 draft mathematics framework explains the theoretical underpinnings of the PISA mathematics assessment, including the formal definition of mathematical literacy continued from 2012, the mathematical processes which students undertake when using mathematical literacy and the fundamental mathematical capabilities which underlie those processes. The draft framework describes how mathematical content knowledge is organised into four content categories and outlines the content knowledge that is relevant to an assessment of 15-year-old students. It describes four categories of contexts in which students will face mathematical challenges. The draft framework recommends

> the proportions of items from each of the four content and context categories, each response format and each process to be used in the 2015 instrument. The categorisations are illustrated with seven units used in PISA surveys and field trials. The PISA assessment will measure how effectively countries are preparing students to use mathematics in every aspect of their personal, civic and professional lives, as part of their constructive, engaged and reflective citizenship.

This draft framework is almost 30 pages in length. You will notice the following elements in the above paragraph:

1. The framework explains the theoretical underpinnings of the assessment.
2. The framework has a definition of the construct being assessed.
3. An explanation of the content and cognitive domensions, and the context within which students are expected to be able to demonstrate their knowledge will be in the framework.
4. The framework presents the proportion of items for each content and cognitive category.
5. Finally, the framework describes the purpose of measuring the construct.

1.10.1 Some questions

There are many ways to develop an assessment framework. However, as research studies, they all tend to have the above components. Nepal has conducted a number of national assessments. Read the Nepalese assessment frameworks, focusing specifically on identifying the above components.

1. Are all components present?
2. Are there additional components?
3. Are there components missing and, if so, why?

1.11 Defining the construct

The PISA draft mathematics literacy framework has the following definition of mathematics literacy:

> Mathematical literacy is an individual's capacity to formulate, employ, and interpret mathematics in a variety of contexts. It includes reasoning mathematically and using mathematical concepts, procedures, facts and tools to describe, explain and predict phenomena. It assists individuals to recognise the role that mathematics plays in the world and to make the well-founded judgments and decisions needed by constructive, engaged and reflective citizens.

Note the components of the definition including the content, skills and context in which that content knowledge will be applied.

1.11.1 Some questions

1. Look at the Nepal national assessments and identify the construct definition.
2. Working as a group, select one of the national assessment subjects and write a one paragraph definition describing at least the content and skills.

1.12 Performance bands

The PISA studies divide the reporting continuum into six proficiency bands. Each band is then described in terms of what the students can do, drawing upon the content, skills and context.

Table 1.5: PISA proficiency scale descriptions for mathematics (2003-2009)

Level	Description
6	At Level 6 students can conceptualise, generalise and utilise information based on their investigations and modelling of complex problem situations. They can link different information sources and representations and flexibly translate among them. Students at this level are capable of advanced mathematical thinking and reasoning. These students can apply their insight and understandings along with a mastery of symbolic and formal mathematical operations and relationships to develop new approaches and strategies for attacking novel situations. Students at this level can formulate and precisely communicate their actions and reflections regarding their findings, interpretations, arguments and the appropriateness of these to the original situations.
5	At Level 5 students can develop and work with models for complex situations, identifying constraints and specifying assumptions. They can select, compare and evaluate appropriate problem-solving strategies for dealing with complex problems related to these models. Students at this level can work strategically using broad, well-developed thinking and reasoning skills, appropriate linked representations, symbolic and formal characterisations and insight pertaining to these situations. They can reflect on their actions and formulate and communicate their interpretations and reasoning.
4	At Level 4 students can work effectively with explicit models for complex concrete situations that may involve constraints or call for making assumptions. They can select and integrate different representations, including symbolic, linking them directly to aspects of real-world situations. Students at this level can utilise well-developed skills and reason flexibly, with some insight, in these contexts. They can construct and communicate explanations and arguments based on their interpretations, arguments and actions.
3	At Level 3 students can execute clearly described procedures, including those that require sequential decisions. They can select and apply simple problem-solving strategies. Students at this level can interpret and use representations based on different information sources and reason directly from them. They can develop short communications when reporting their interpretations, results and reasoning.
2	At Level 2 students can interpret and recognise situations in contexts that require no more than direct inference. They can extract relevant information from a single source and make use of a single representational mode. Students at this level can employ basic algorithms, formulae, procedures, or conventions. They are capable of direct reasoning and making literal interpretations of the results.
1	At Level 1 students can answer questions involving familiar contexts where all relevant information is present and the questions are clearly defined. They are able to identify information and to carry out routine procedures according to direct instructions in explicit situations. They can perform actions that are obvious and follow immediately from the given stimuli.

1.12.1 Some questions

TIMSS uses four levels while the USA's National Assessment of Educational Progress (NAEP) uses three. Focusing on PISA, examine the PISA level descriptions.

1. Did you notice how the size of the statements descriptions change as you move from level 1 to level 6?
2. Extract the first sentence from each level description. How does this sentence change from one level to the next?
3. What proficiency or performance levels are used in Nepal?
4. How are these proficiency levels used?

2. Writing assessment items

2.1 Test Specifications

We typically develop two to three times as many items as we need. The AERA standards for test development state the following:

> The test specifications should be documented, along with its rationale and the process by which it was developed. The test specifications should define the content of the test, the proposed number of items, and item formats, the desired psychometric properties of the items, and the item and section arrangement. It should also specify the amount of time for testing, directions to the test makers, procedures to be used for test administration and scoring and other relevant information.

The PISA mathematics literacy framework specifies the following:

Table 2.1: Approximate distribution of score pointsby process category

Process category	Percentage of score points
Formulating situations mathematically	25
Employing mathematical concepts, facts, procedures, and reasoning	50
Interpreting, applying and evaluating mathematical outcomes	25
TOTAL	100

Table 2.2: Approximate distribution of score pointsby process category

Content category	Percentage of score points
Change and relationships	25
Space and shape	25
Quantity	25
Uncertainty and data	25
TOTAL	100

Table 2.3: Approximate distribution of score points by context category

Context category	Percentage of score points
Personal	25
Occupational	25
Societal	25
Scientific	25
TOTAL	100

2.1.1 Some questions

Focusing on a subject of your choice, and working in a team with national education resources:

1. How many content areas will you use? Write a brief description of each area. Show how the areas are related to the curriculum. Allocate a percentage of items to each area.
2. How many cognitive areas will you use? Write a brief description of each area. Show how the areas are related to the curriculum. Allocate a percentage of items to each area.
3. How many content areas will you use? Write a brief description of each area. Show how the areas are related to the curriculum. Allocate a percentage of items to each area.(You may decide not to use content areas but if you are assessing literacy you really should.)
4. How would you communicate and justify your decision to stakeholders (ministry officals, teachers, parents, university professors)?

2.2 The item development process

Haladyna presents the following item development process:

1. Make a plan for how items will be developed.
2. Create a schedule for item development.
3. Conduct an inventory of items in the item bank.
4. Identify the number of items needed in each of these areas.
5. Identify and recruit qualified subject matter experts for developing new items.
6. Develop an item-writing guide.
7. Distribute the guide to the item writers.
8. Conduct item-writing training for these item writers.
9. Make assignments to item writers based on the inventory and the evaluation of needs.
10. Conduct item reviews leading to one of three decisions: keep, revise, retire.
11. Field test surviving items.

12. Evaluate the performance of items.
13. Place surviving items in the operational item bank.

2.2.1 Some questions

Assume you are planning a national assessment.

1. Make a plan for the development of an item pool for field testing.
2. Include the training of teachers in each of the development regions.
3. Develop an item submission protocol.

2.3 TIMSS item writing guidelines

The TIMSS 2011 item-writing will be conducted in four major areas, with approximately a quarter of the participants working in each area – mathematics fourth and eighth grades and science fourth and eighth grades. Typically, participants will work in groups of two or three. Each group will be assigned specific content areas. Participants will be writing items in English and saving them as Microsoft Word files that will be collected at the end of each day.

When writing items, PLEASE:

1. Address the TIMSS 2011 Framework. Write questions that match the topics in each content domain, and pay particular attention to writing questions that cover the range of the three cognitive domains. In accordance with the TIMSS 2011 Frameworks, write questions that address the applying and reasoning domains, as well as the knowing domain.
2. Consider the best item format for the question. About half of the items you develop should be multiple-choice and the other half should be constructed-response items worth 1 or 2 score points.
3. For the content domain(s) you are assigned, write at least 10 to 12 items each day.
4. For each item, consider the timing, grade appropriateness, difficulty level, potential sources of bias (cultural, gender, or geographical), and ease of translation. Make sure that item validity is not affected by factors that unnecessarily increase the difficulty of the item, such as unfamiliar or overly difficult vocabulary, grammar, directions, contexts, or stimulus materials.
5. For multiple-choice items, keep the guidelines for writing multiple-choice questions in mind. In particular—ask a direct question, make sure there is one and only one correct answer, and provide plausible distracters.
6. For constructed-response questions, write a fullcredit answer to the question in terms of the language, knowledge, and skills that a good fourth- or eighth-grade student could be expected to possess. This tests the clarity of the question and also provides guidance about whether to allocate 1 or 2 score points to the item.
7. Develop a specific scoring guide for each constructedresponse item.

2.3.1 General Issues in Writing Items for TIMSS

This section is copied verbatim from the TIMSS item writing guide.

Item writing is a task that requires imagination and creativity, but at the same time demands considerable discipline in working within the assessment framework and following the guidelines for item construction provided in this manual. These guidelines pertain to good item and test development practices in general, and have been collected from a number of sources. They are designed to help produce items that measure achievement in mathematics and science fairly and reliably, and that enhance the validity of the TIMSS

tests. All of the following issues must be considered in judging the quality and suitability of an item for TIMSS 2011. Items must meet these guidelines to be considered for inclusion in the field test for TIMSS in 2010.

Alignment with the Frameworks

The TIMSS assessment frameworks in mathematics and science describe those outcomes generally regarded as important at the fourth and eighth grades. It is fundamental that every item written for mathematics or science measures one of the content topics and one of the cognitive domains described in the TIMSS 2011 frameworks. In preparing to produce an item for either fourth or eighth grade, the first step is to focus on the content topic to be assessed. In writing each item, remember that it also contributes to a measure of proficiency in a cognitive domain. Keep in mind that TIMSS is assessing student learning in particular topics. Think:

- What should the student know?
- What should the student be able to do?

That is, what knowledge does this item allow a student to show? What cognitive processes does this item require a student to demonstrate?

Types of Items

TIMSS includes two types of items: multiple-choice items where the student chooses the correct answer from four response options, and constructed-response items where the student is required to provide a written response. PLEASE keep item format in mind. About half of the items you develop should be multiple-choice and half should be constructed-response.

- Multiple-choice items allow valid, reliable, and economical measurement of a wide range of content in a relatively short testing time.
- Constructed-response items allow students to provide explanations, support an answer with reasons or numerical evidence, draw diagrams, or display data.

If you think of another item type, it may be used as long as it provides valid measurement and is feasible to administer and to score reliably.

2.3.2 Some questions

Focusing on a subject area of your choice, and working in a team with curriculum and other resources, develop at least 5 multiple choice items and 5 short answer, constructed repsonse items.

1. For each item explain why it is located in a specific content by cognition cell.
2. For each multiple choice item, explain why the distractors are plausible and what they mean if a student selects them.

2.4 Marzano's approach

Marzano (2007) suggests the following:

- There should only be 15 to 20 measurement topics in each year in each subject.
- That means, on average, a measurement topic lasts 2 to 3 weeks.
- Teachers should assess the students at least twice a week in each topic, and they should design assessments that test accumulated knowledge.
- That three types of questions be asked of the students:
  - Type I are simple recall, understanding and application type questions. These questions for the basic building blocks of topic knowledge.

- Type II are more complex recall, understanding and application questions. These questions integrate topic knowledge but still draw upon knowledge and skills that have been explicitly taught.
- Type III questions extend the student by asking questions based upon the topic but which have not been explicitly taught.

2.4.1 Some questions

1. How is Nepal's curriculum organized?
2. Could an assessment framework document be written that explicitly writes about measurement topics that are faithful to the curriculum?
3. What positive and adverse effects might occur if such measurement topics were written?

2.4.2 Marzano and progress

He also suggests that the following scale be used for representing progress on a topic.

Table 2.4: Scoring scale representing progress on a measurement topic

Topic score on scale	Description of place on scale
4.0	In addition to Score 3.0 performance, in-depth inferences and applications that go beyond what was taught
3.5	In addition to Score 3.0 performance, partial success at inferences and applications that go beyond what was taught
3.0	No major errors or omissions regarding any of the information and/or processes (simple or complex) that were explicitly taught
2.5	No major errors or omissions regarding the simpler details and processes and partial knowledge of the more complex ideas and processes
2.0	No major errors or omissions regarding the simpler details and processes but major errors or omissions regarding the more complex ideas and processes
1.5	Partial knowledge of the simpler details and processes but major errors or omissions regarding the more complex ideas and processes
1.0	With help, a partial understanding of some of the simpler details and processes and some of the more complex ideas and processes
0.5	With help, a partial understanding of some of the simpler details and processes but not the more complex ideas and processes
0.0	Even with help, no understanding or skill demonstrated

2.4.3 An example Marzano test

An example test based upon Marzano's ideas is as follows: Section I:

Table 2.5: Scoring scale representing progress on a measurement topic

	Red-Bird	Rental Easy	Rental Reliable	Rental MandA Rental
Daily Rate	\$43.00	\$27.50	\$40.00	\$35.25
Free Mileage	1,200	500	900	800
Cost per Mile	\$0.22/mile	\$0.32/mile	\$0.25/mile	\$0.20/mile

1. Which company has the highest daily rate?
Answer

2. Which company has the most free mileage?
Answer

3. If each company had the same daily rate and the same amount of free mileage, which would be the cheapest?
Answer

4. If each company had the same amount of free mileage and the same cost per mile, which company would be the most expensive?
Answer

5. Once you've used up your free mileage, which company would cost the least amount of money to travel 100 miles in a single day?
Answer

Section II:
6. If you travel 100 miles per day, which company is the least expensive for
5 days: Answer

10 days: Answer

15 days: Answer

20 days: Answer

Create a table or a graph that shows how expensive each company is for each of the four options above (5 days, 10 days, 15 days, 20 days), and explain how you calculated your answers.

Section III:
7. Each of the four companies could be considered the "best deal" under certain conditions. For each company, describe the situation under which it would be the best selection. In your answer and explanation, use the daily rate, free mileage, and the rate per mile after free mileage.

2.4.4 Some questions

1. Could you design a national assessment tha used Marzano's ideas?
2. How could ou use such an assessment to inform policy and also teaching?
3. What positive and adverse effects might occur if such measurement topics were written?

2.5 A look at cognition using the revised Bloom taxonomy

Large scale assessments typically use a cognitive taxonomy which is specifically written for the construct (the assessed domain). This section will describe one of the most commonly used taxonomies which serves as a foundation for their descriptions.

Table 2.6: Bloom's revised Taxonomy - Cognitive Domain

Category or 'level'	Behavior descriptions	Examples of activity to be trained, or demonstration and evidence to be measured
Remembering	Recall or recognize information	Multiple-choice test, recount facts or statistics, recall a process, rules, definitions; quote law or procedure
Understanding	Understand meaning, re-state data in one's own words, interpret, extrapolate, translate	Explain or interpret meaning from a given scenario or statement, suggest treatment, reaction or solution to given problem, create examples or metaphors
Applying	Use or apply knowledge, put theory into practice, use knowledge in response to real circumstances	Put a theory into practical effect, demonstrate, solve a problem, manage an activity
Analyzing	Interpret elements, organizational principles, structure, construction, internal relationships; quality, reliability of individual components	Identify constituent parts and functions of a process or concept, or de-construct a methodology or process, making qualitative assessment of elements, relationships, values and effects; measure requirements or needs
Evaluating	Assess effectiveness of whole concepts, in relation to values, outputs, efficacy, viability; critical thinking, strategic comparison and review; judgment relating to external criteria	Review strategic options or plans in terms of efficacy, return on investment or cost-effectiveness, practicability; assess sustainability; perform a SWOT analysis in relation to alternatives; produce a financial justification for a proposition or venture, calculate the effects of a plan or strategy; perform a detailed risk analysis with recommendations and justifications
Creating	Develop new unique structures, systems, models, approaches, ideas; creative thinking, operations	Develop plans or procedures, design solutions, integrate methods, resources, ideas, parts; create teams or new approaches, write protocols and contingencies

2.5.1 Some questions

Sometimes the lower three levels of Bloom's taxomony (remembering, understanding, and applying) are grouped together and called "lower order thinking skills". The other three levels, analyzing, evaluating and creating are called "higher order thinking skills" or "HOTS".

1. Look back over the PISA proficiency levels. Can you identify examples of Bloom's taxonomy being applied?
2. Look over the Nepalese national assessment tests. Can you assign each item to one of the Bloom's taxonomy categories?

2.6 Some resources

The USA's National Assessment of Educational Progress has banks of released items:
https://nces.ed.gov/nationsreportcard/about/naeptools.aspx

The Trends in International Mathematics and Science Studies (TIMSS) has released items:
http://isc.bc.edu/isc/publications.html

The Programme for International Student Assessment (PISA) also has released items:
https://nces.ed.gov/surveys/pisa/educators.asp

3. A Workshop Introducing ConQuest

3.1 Objectives of This Workshop

This workshop seeks to introduce you to a large number of concepts in psychometrics, and specifically in a technically challenging and advanced area of item response theory (IRT). The workshop takes a generally "hands-on" approach and "skims the surface" of the field, glossing over much of the mathematics. It should be treated for what it is: an introduction. Nevertheless, there are a number of important objectives that shape the workshop. So, by the end of the workshop you should be able to:

- Identify and use a variety of resources through the internet.
- Understand some of the challenges in a "typical" test score.
  - Develop and defend a scoring system
- Use the ConQuest program and simulated data to develop and understanding of the following concepts:
  - Rasch model
  - Logit scale
  - Fit statistics
  - Item and test characteristics curves
  - Item and test information curves
- Apply the concepts of Rasch measurement to test design and analysis
- Understand the concept of "link items" and be able to apply this concept to test design.
- Calculate and interpret the item difficulty of test items

Finally, much of material presented here was initially developed during my work in the USA, Australia, Fiji, and especially while I was working at Flinders University. However, the work has been scaled back to reflect the two hours in the schedule and the very introductory nature of the presentation.

3.2 Introducing Item Response Theory

Modern test theory was developed partially to develop the address some of the limitations of classical test theory. Two specific issues that modern test theory seeks to address include:

1. Can you imagine what would happen if you gave the same test to a class of low achievers, a mixed class, and a class or high achievers? How would the questions on the test be perceived by each group?
2. Can you imagine how an estimate of a student's proficiency in a subject might change if the test was very easy, very hard, or had a mixture of easy, medium and hard items?

For these two reasons, and other reasons, measurement experts started developing a new theory of measurement: item response theory.
For the remainder of this workshop we will focus on this theory, using:

1. Mainly dichotomous items
2. Applying the simplest form of Item response theory to artificial test data
3. And use one of the major software program, developed initially for international studies of student achievement in mathematics and science.

In keeping with the "hands on" approach, we will avoid theory and mathematics as much as possible.
A word of caution, though:"Item response theory (IRT) is sometimes presented as an alternative to the traditional true score theory. However, it is more realistic in applications involving large scale test development and scoring to consider the traditional theory and IRT to be complementary: Practical test construction using IRT typically includes reference to traditional statistics as well as the item parameters and goodness-of-fit statistics of IRT." (Thissen and Orland, 2001, p. 73)

This workshop skips over classical or traditional test theory but you almost certainly need to develop your understanding of that theory in your work with educational assessments.

3.3 The Rasch Model

"Rasch (1960/1980, pp. 74–75) specified that a person should be characterized by degree of ability ξ and an item should be characterized by a degree of difficulty δ. Both ξ and δ are assumed to be greater than zero. Then, following a metaphor with physical laws, Rasch specified that if a second person has twice the ability of the first, 2ξ, and a second item is twice as difficult as the first, 2δ, the second person should have the same probability of solving the second item as the first person has of solving the first item.
Because the probability of a correct response is a function of the ratio of the proficiency of the person to the difficulty of the item, the item parameters cancel for ratios of probability correct for two persons, leaving an item-free comparison of their proficiencies. Thus, the model makes objective or itemfree statements about the relative likelihood that two persons will respond correctly to an item or a set of items, without any reference to the items themselves. Some theorists regard this property of the model to be highly desirable (see Andrich, 1988, cha P. 2, for advocacy of this position); others, including the authors, consider this property less important than other aspects of IRT models." (Thissen and Orlando, 2001, pp. 74-75)

3.3.1 Some questions

In keeping with my desire to avoid deep, technical issues, I will try to avoid unpacking too much of the above quote. But you should note the following:

1. Note the term "item-free". This is central to the Rasch model. It is the feature that enables the Rasch model to claim "objective measurement".
2. There are two broad fields of IRT models
   (a) Measurement or theory driven models (Rasch models)
   (b) Data-summary models (non-Rasch and Rasch models)
3. All IRT models are "strong models" in that they make strong assumptions:
   (a) If the assumptions are not met, then you should not use the model
   (b) The Rasch models make the strongest assumptions
   (c) The assumptions or IRT are hardly ever met in reality
4. Psychometricians are passionate people. There have been many bitter, personal fights, disguised as professional debates, centring on which models to use.

3.4 Introducing Simulated Data

Statisticians and psychometricians often use simulated data when they are develop their mathematics models. We will also use simulated data throughout the workshop. This lets us control the complexity of the data and, hopefully, make learning the core concepts much easier.

However, if we were using "real data" then we would also focus on (construct) validity and have a lot more information and theory to draw upon.

To simulate the data, I used the Rasch model, between 50 to 3000 students randomly drawn from a normal distribution and a set of binary (dichotomous) items. Binary or dichotmous items are coded as zero (wrong) and one (right or correct)

The simulated data is located on the USB (flash) drive in "X:\NEQMAP_IRT_Sim_Data" folder, where X is the drive name. The simulated data has been saved as text files (.TXT) and you can open them in MS Word, Word Pad and Note Pad.

I have used a file naming convention that indicates how many binary (dichotomous) items and how many fictitious students there are in the data set. For example, the file "NEQMAP_Sim_Data_01_40_50.TXT" has 40 items and 50 students while "NEQMAP_Sim_Data_01_100_500.TXT" has 100 items and 500 students.

If you open one or more of these files, you will see that they all have a common structure.The first column is simply an identification number. You will see that this column increments by one in most files. The next column has positive and negative numbers which are random numbers. We will come back to these later – they are simulations of student ability. The next column is the raw scores for the test. And the final block are "scored" answers.

3.5 Introducing ConQuest

Item Response Theory requires specialist programs. There are a large number of programs available, each with a set of features. The purpose of this workshop is not to advocate any particular program. However, and only for expediency reasons, this workshop with use ConQuest.

There are tutorials located at:
https://www.acer.edu.au/conquest/notes-tutorials
You can download a trial version from this site by following the links to:
http://conquest-sales.acer.edu.au/index.php?cmd=toTrial

3.6 Running ConQuest

Assuming you have downloaded ann installed a trial version of ConQuest, the next step is to open the program.
Then, selecting the "File" command, you should navigate to the NEQMAP folder. In that folder you will notice a large number of files with the "cqc" tag. These are ConQuest command files.
And finally, select one of the command files. We will focus the rest of this workshop on the file that has 40 items and 500 students.
When you open the file, you will notice that two boxes appear in ConQuest. One is the command or input box, and the other is an output box.
The opened command file may appear confusing and even daunting. However, keep in mind that it is actually very highly structured and the primary goal of this workshop is not to learn how to use ConQuest but is to learn about IRT. That is, you don't need to learn the command language.
Since you are not expected to learn the ConQuest command language, I have provided you with all the command files that we need. However, there is one thing that you may need to change. The second line of the ConQuest command file that you opened has the following line: "datafile C:\NEQMAP_IRT_Sim_Data\NEQMAP_Sim_Data_01_40_500.TXT;" This is telling the program where to find the data file. You may need to change the letter "C" to refer to another drive location. If you do, then you should make more changes further down in the command file before you "run" the command file.

3.7 Running ConQuest

Assuming you have made any necessary changes to the command files (that is the file locations), you are now ready to run ConQuest. Select the "Run" command from the menu and select "Run All" You will immediately see a few changes (unless there is an error in your program). An important one is shown in the figure below. This is showing you that the program has read in the student data and is now performing a large number of calculations. We won't go into detail what the program is doing but it is important to learn about later. You will also notice changes in the "output" window. If all goes as planned, you will see a large number of graphs produced as the ConQuest generates outputs we have asked for in the command file.
We will now explore these graphs.

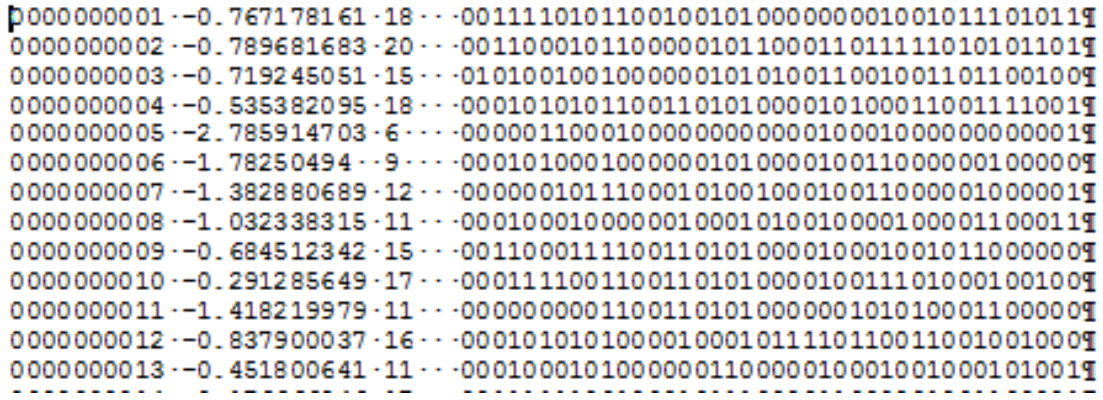

Figure 3.1: Screen capture of the simulated data file

3.8 The "Test Information Graph"

The last graph that appeared should have been the "Test Information" graph. We will explore feature of this graph. The x-axis (shown in figure below with the black circle; also called the abscissa or horizontal axis) extends from -5 to 6. This is labelled "Latent Trait (logits)". We will come to "logits" later, but at this stage we simply note that the "Latent Trait" is what we are trying to measure. That is, it is the construct underpinning the assessment. For simplicity, we make the following observations:

1. Negative numbers indicate low levels of the construct, or better still, less of the construct or ability.
2. Around zero is "average" levels
3. And positive numbers indicate above average levels of the construct, or more accurately more of the construct or ability.

Keep in mind that this is a very simple introduction. The meaning of the x-axis will become a lot clearer when I explain what "logits" are (and certainly we want to move away from using the term "average") The y-axis (shown in Figure 8 with the blue circle, also called the vertical or ordinate axis) is labelled "Information". The numbers are more or less arbitrary and should be interpreted as relative numbers. (In fact they come from a Hessian matrix computed for each test question which is mathematically a little challenging). The curve itself is telling us something very important about the test. It tells us where the test scores are most accurate and were they are least accurate. So, if you look at Figure 9, you will see that the curve is:

1. Rather peaked. That is, it looks like a mountain.
2. The peak occurs at around the "0.0" point on the "Latent Trait" scale, where the information is approximately 6.5. This is region is indicated by the black circle. The test is very suited to measuring students with ability in this region.
3. The curve trails off at -5 and +5 on the "Latent Trait" scale to information values less than one. These regions are indicated by the blue circles. The test is not well suited for measuring ability in these regions.

This test would be good if our main interest was within the middle region. It would not be so good if we were interested in other regions of the score or ability range. If we had a pass/fail cut-point, we would design the test so that the test information graph peaked at that cut-point. If we have many regions in which we are interested in, we would design the test so that the information graph was flatter (like a mesa or plateau).

3.8.1 Some questions

1. How might you use the information graph in the evaluation and future design of your assessments?
2. What questions do you have so far? What is clear and starting to make sense?

3.9 Saving and Closing the Test Information Graph

There are a number of options for saving the graph. I typically use the "BMP" option which saves the graph as a picture file. I also typically give the saved picture a name which

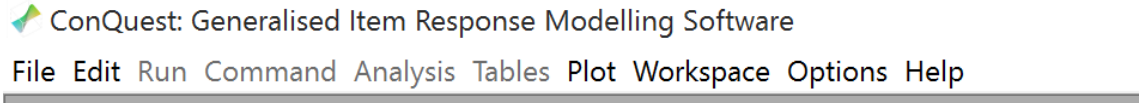

Figure 3.2: ConQuest commands

links it back to the ConQuest command file.
Once you have saved the graph, you might wish to close it. In this case, you should click on the greyed "x". If you accidently click on the large "x", all of the graphs will close. Note that in some computers the large "x" is also red (just to make it more confusing) but you might notice that I am trying to tell you something with the purple rectangle.
If you click on the wrong "x" and all the graphs close, simply do the following. Go back to the ConQuest program, highlight the last line and then choose the "Run" and "Run Selection".

3.9.1 Some questions

You might have noticed that the first word on each line is actually a command word. It is a special word that tells ConQuest what to do. You could use the ConQuest command dictionary to learn more about each line.

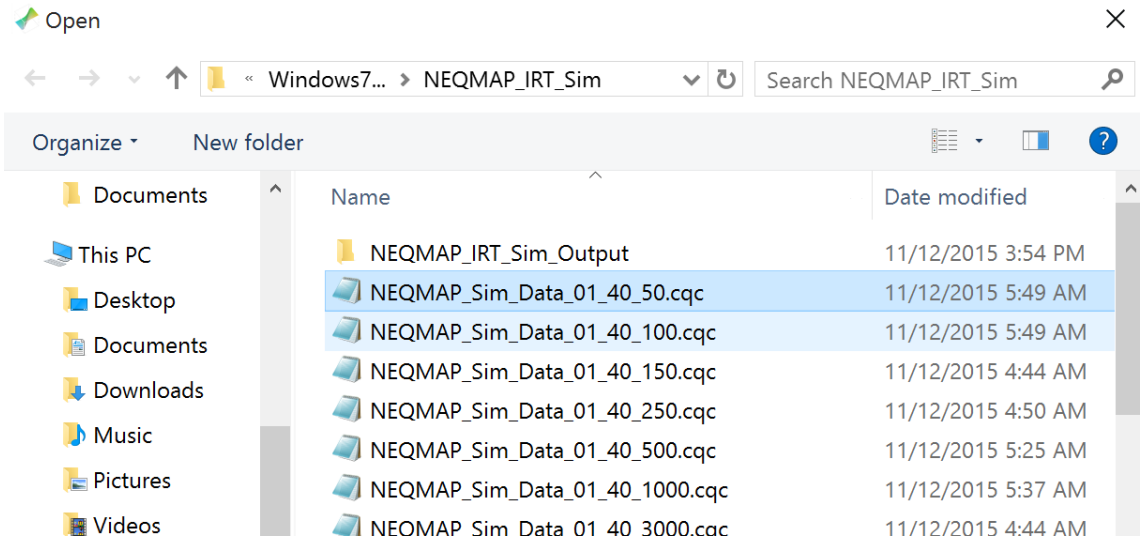

Figure 3.3: ConQuest command files in the NEQMAP folder

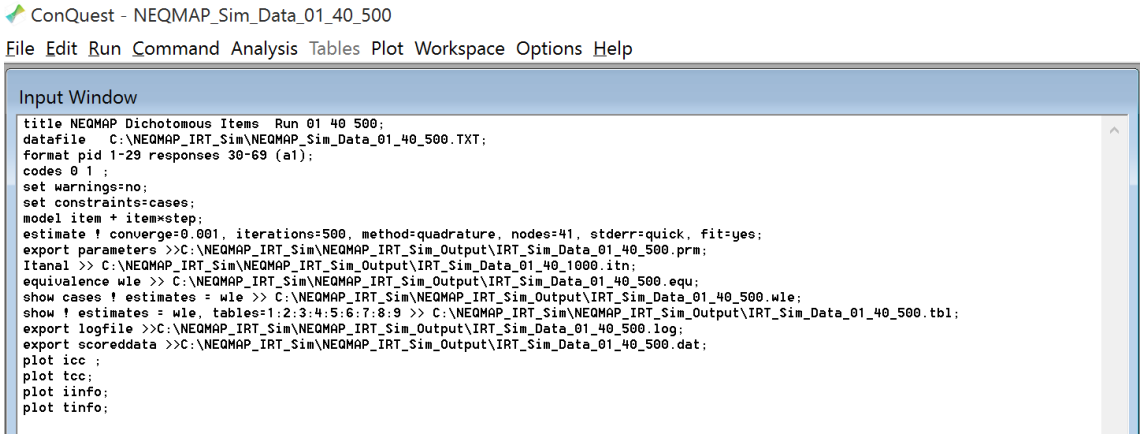

Figure 3.4: ConQuest command file NEQMAP_Sim_Data_01_40_500.cqc

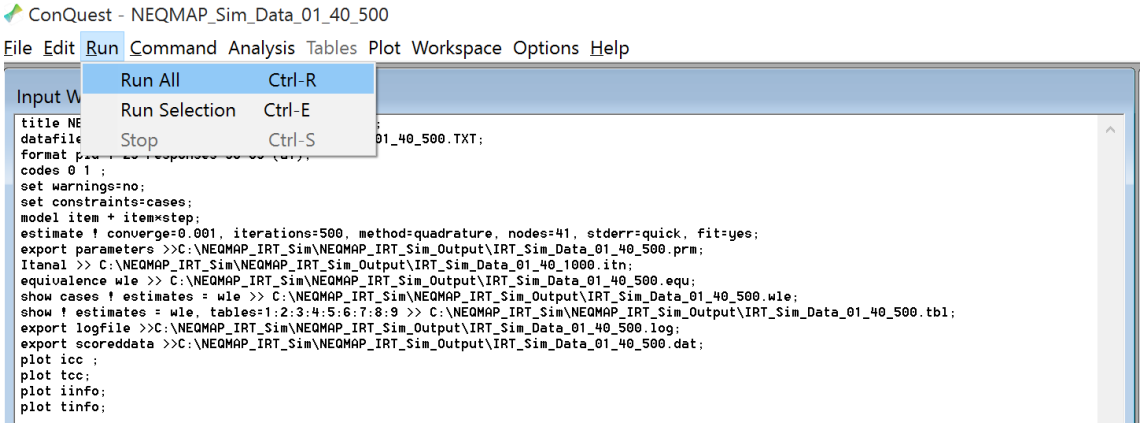

Figure 3.5: Running all the ConQuest commands

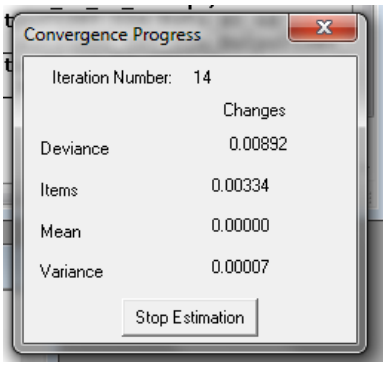

Figure 3.6: ConQuest estimating item parameters

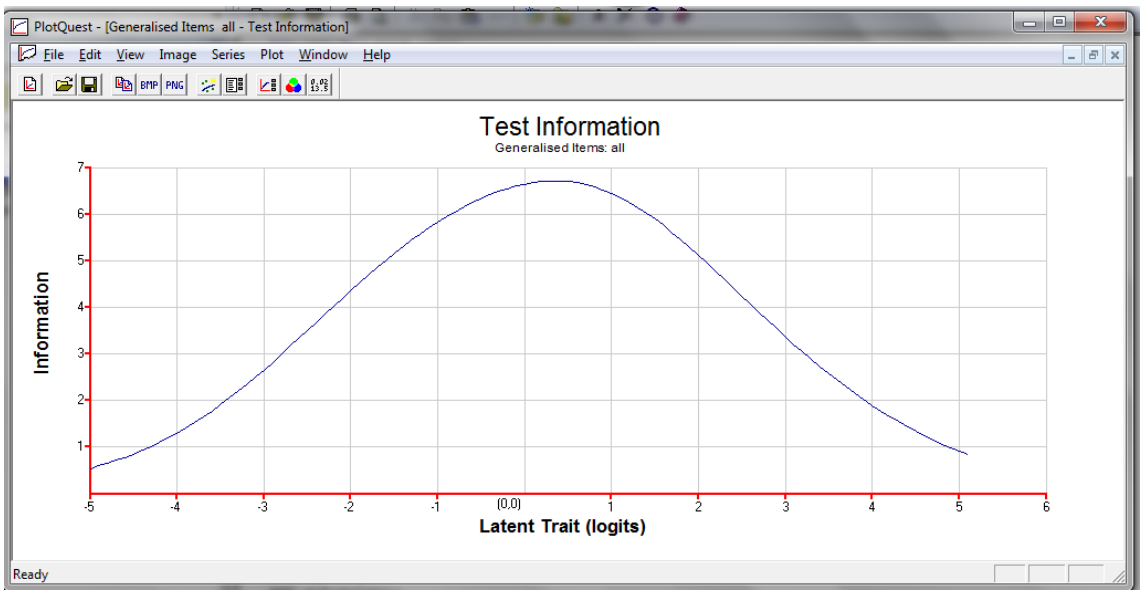

Figure 3.7: The test information graph

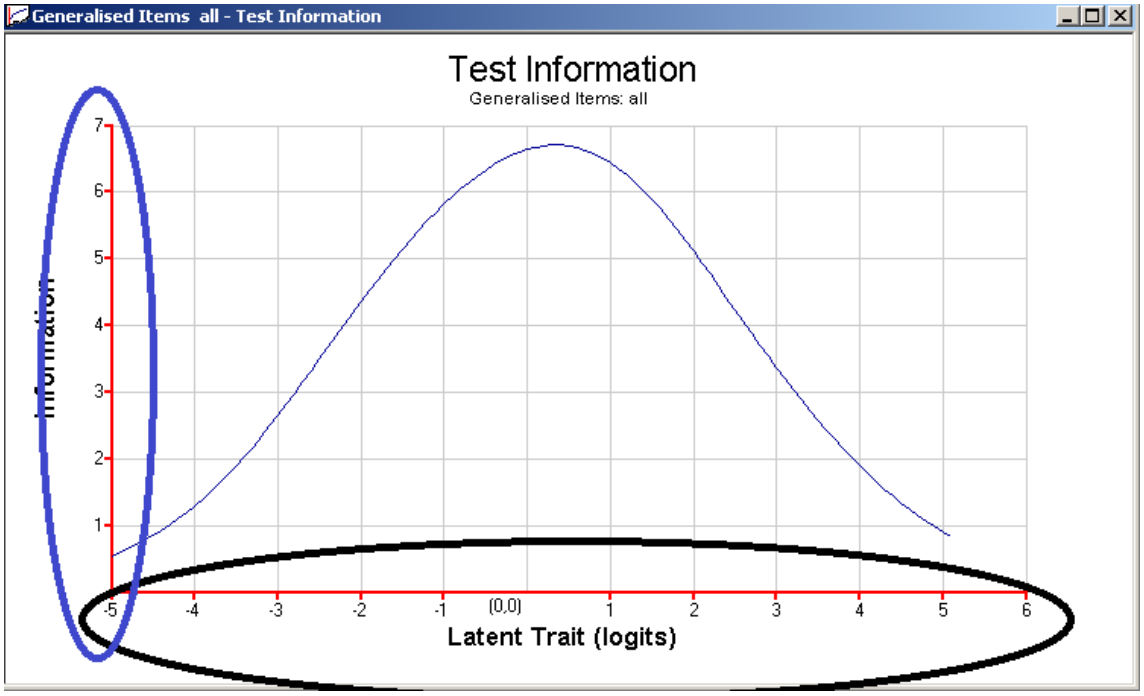

Figure 3.8: The test information graph explained

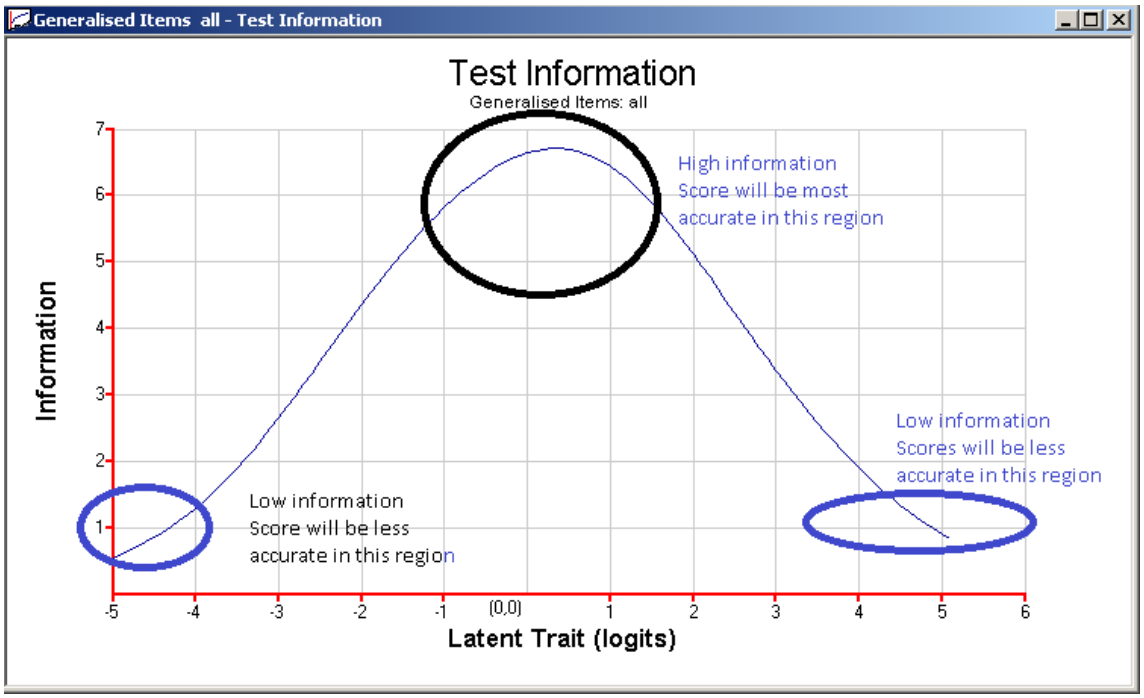

Figure 3.9: Interpreting regions of the test information graph

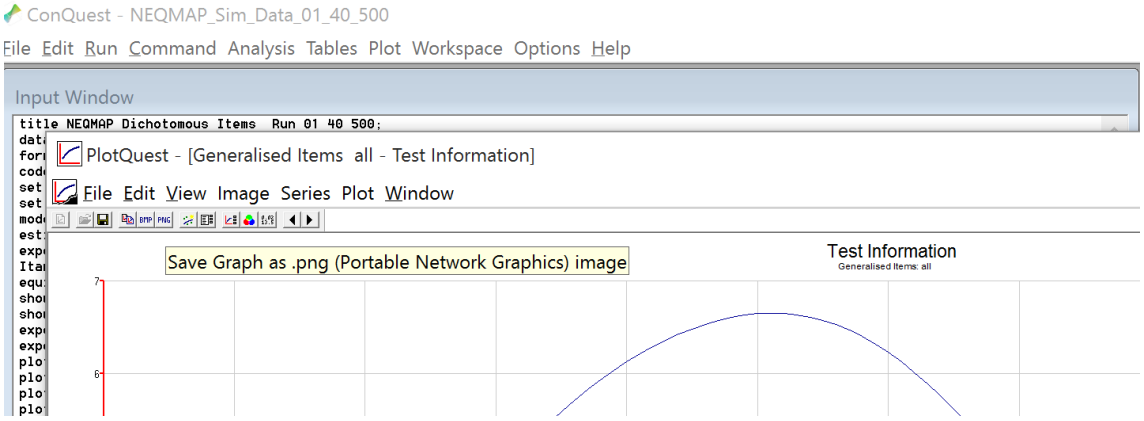

Figure 3.10: Regions of the test information curve

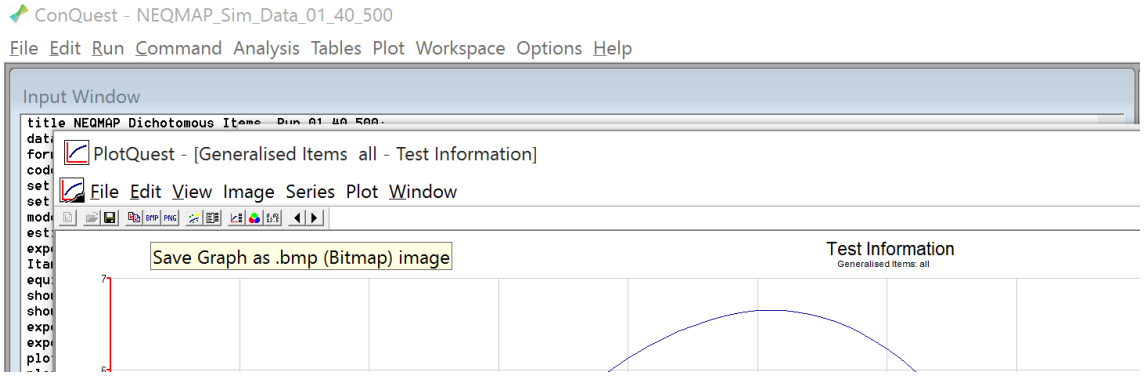

Figure 3.11: Saving a graph as a portable network graphic file

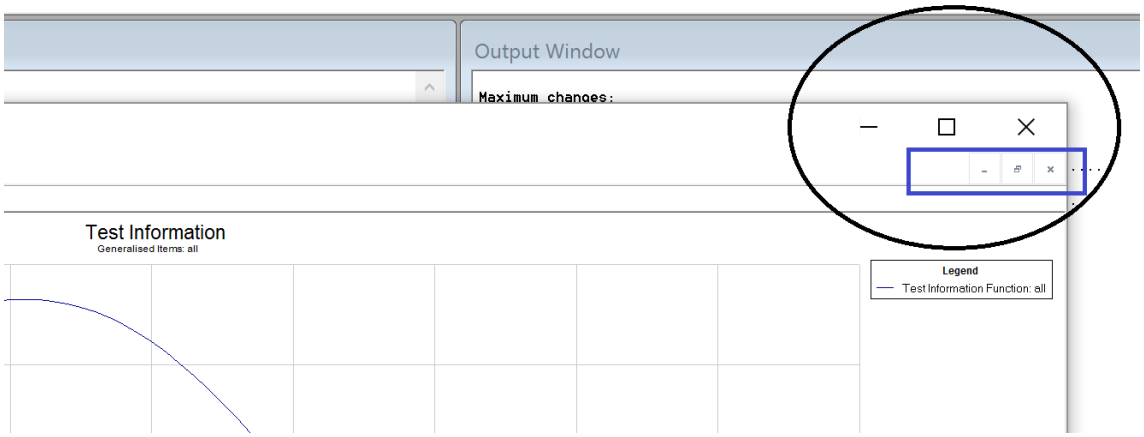

Figure 3.12: Saving a graph as a bitmap file

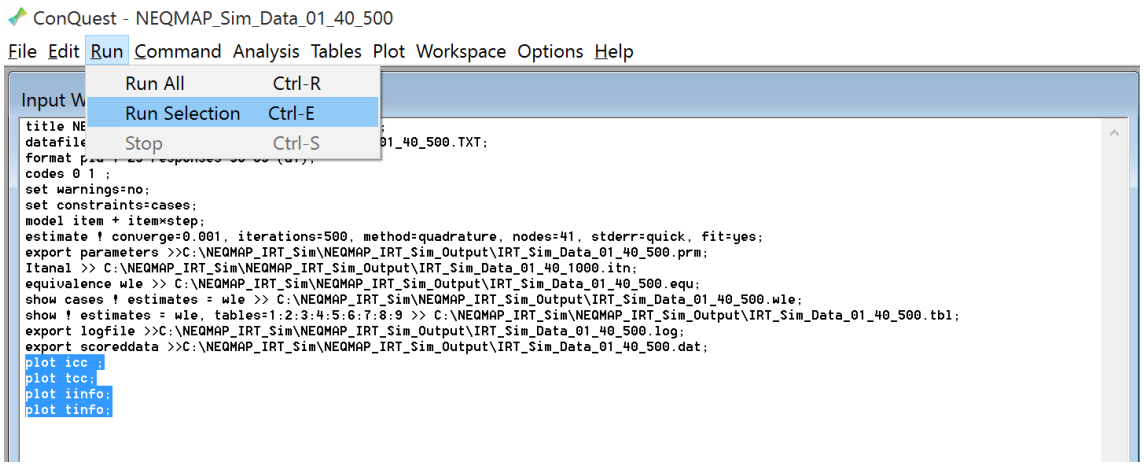

Figure 3.13: Closing the graphs by choosing the “x” in the blue rectangle

3.10 Item Information Graphs

If you closed just the test information graph, you should now be looking at an "Item Information Graph". You should notice the following points:

1. The item (question being graphed is identified in the top blue bar. In figure 13, the blue bar contains "PlotQuest –[item40 (4) – Item Information]"
2. The x-axis is labelled in exactly the same way as the Test Information graph.
3. The y-axis is now called "Item Information" and the scale is somewhat smaller than before.

The interpretation of the graph is much the same way as for the Test Information graph, only at the question level. For the item shown in Figure 3.14, we can say this item works best when the Latent Trait has a score of -1. It would generally be interpreted as an easy item.

However, before doing so we may wish to change the axes a little. The next few figures show you how to edit a graph in ConQuest.

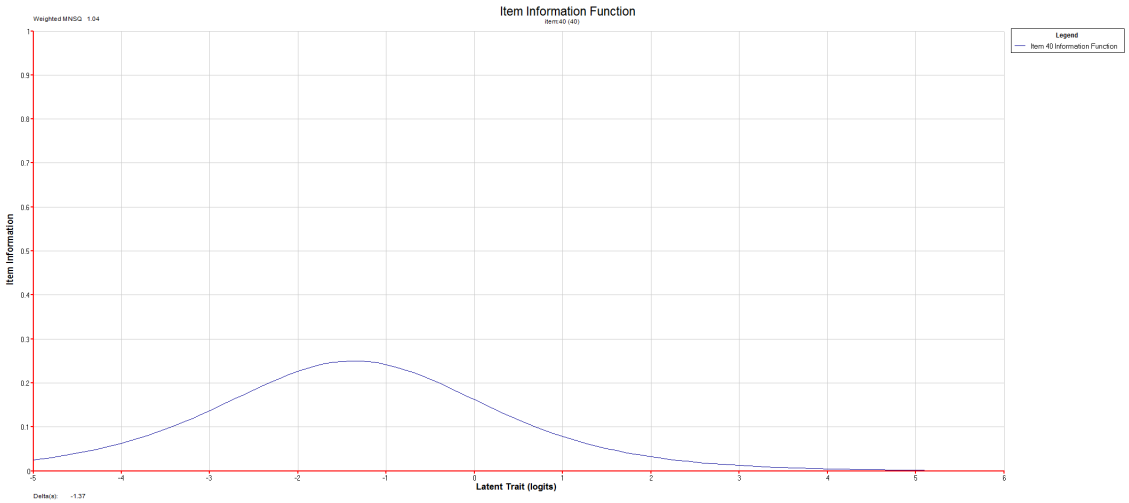

Figure 3.14: Saving a graph as a portable network graphic file

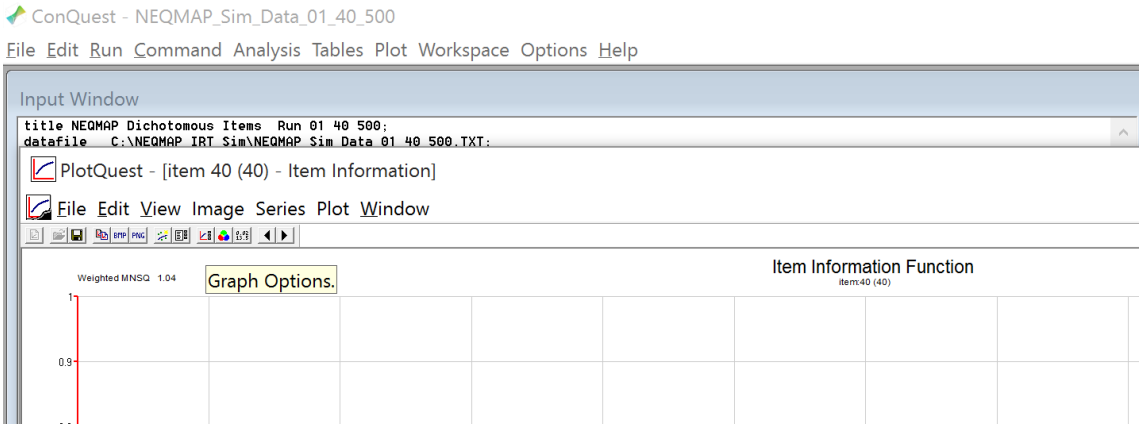

Figure 3.15: Editing the graph by choosing graph options

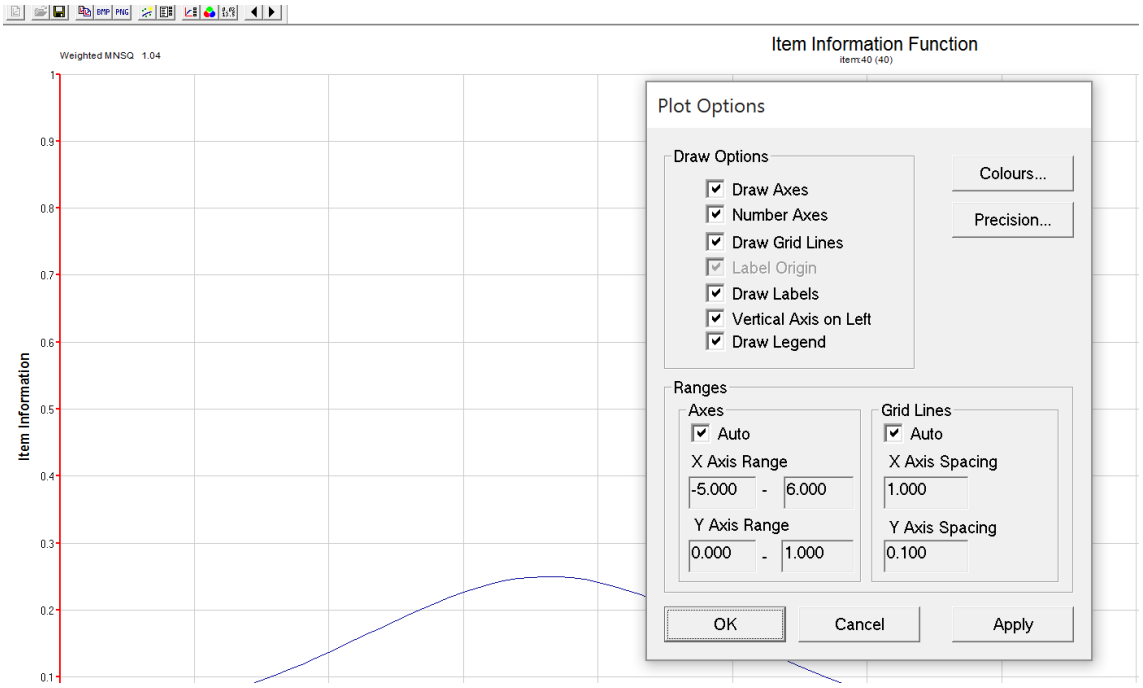

Figure 3.16: The graph options

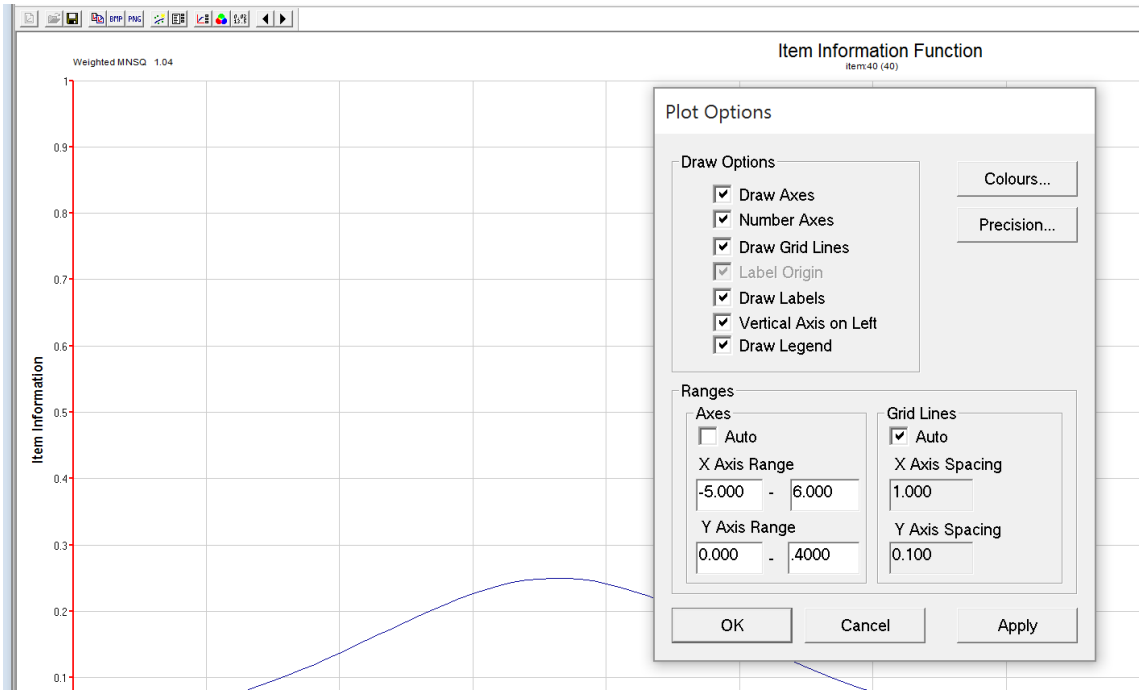

Figure 3.17: Changing the y-axis range

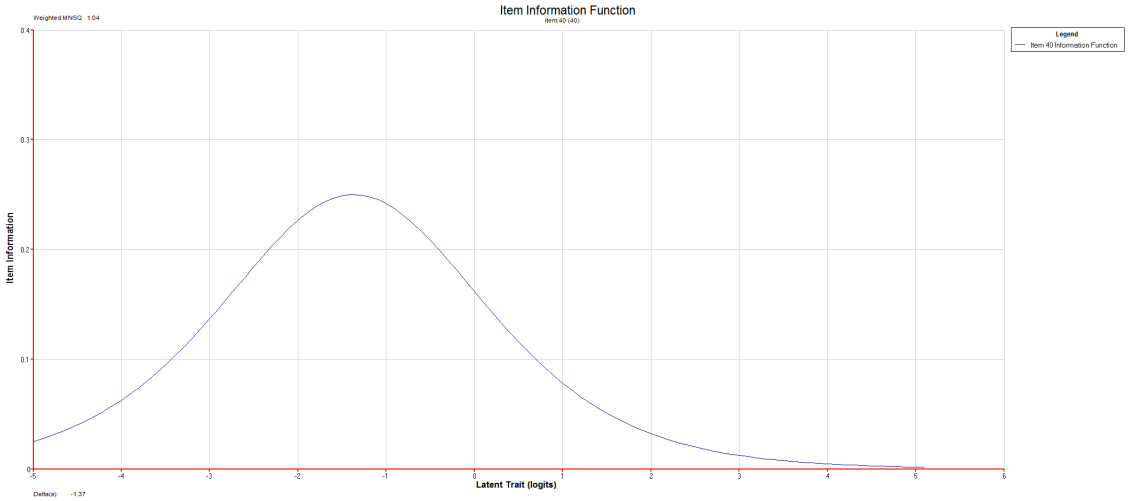

Figure 3.18: The revised item information graph

3.11 Table output from ConQuest

I now want to explore some of the text output files form ConQuest. These files were put into the "NEQMAP_IRT_Sim_Output" folder. You can see that ConQuest was asked to produce a number of files using the "export" command. Specially, the program was asked to produce text files for:

1. The item parameters (the deltas or item difficulty estimates)
2. A log file which has information about how the program performed
3. And a scored data file.

However, the text file that we should look at first has the file type "tbl" for table. It was produced by the command line:

```
show ! estimates = wle, tables=1:2:3:4:5:6:7:8:9 » C:\NEQMAP_IRT_Sim\NEQMAP_IRT_Sim_Output\IRT_Sim_Data_01_40_500.tbl;
```

This file contains a lot of information, all of it important but only some of which we will explore here.

The file can be opened with MS Word, Word Pad or Note Pad.

3.11.1 Opening in MS Word

Sometimes the table file opens up and all the lines have wrapped around, making the tables hard to read. For example, as shown Figure 3.19 , the lines have wrapped around. Also shown in that figure are the MS Word's font and font size settings. The New Courier font is best for viewing ConQuest output. Changing the font size to 8 for the entire document will get rid of the wrap-around (seee Figure 3.20).

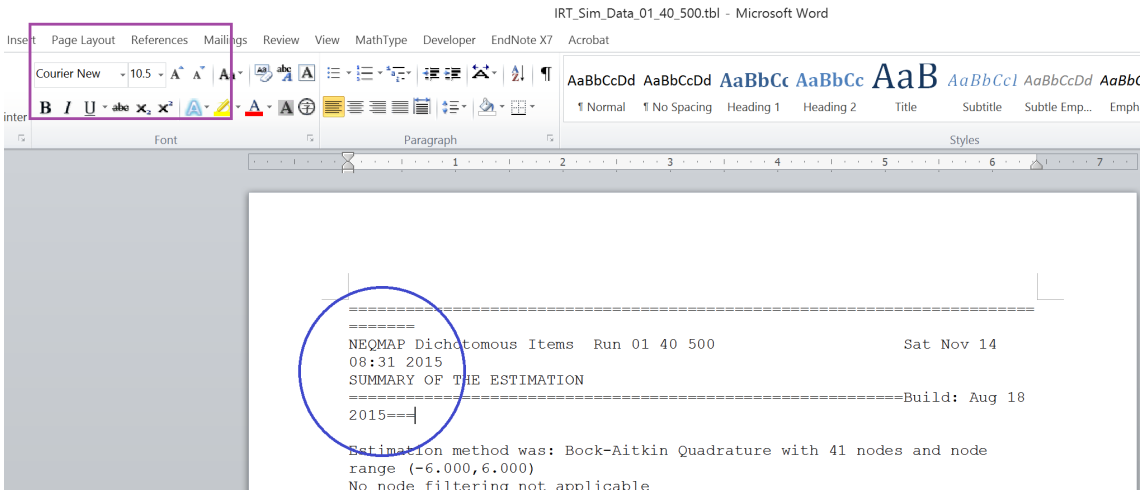

Figure 3.19: Wrapped around text in Microsoft word

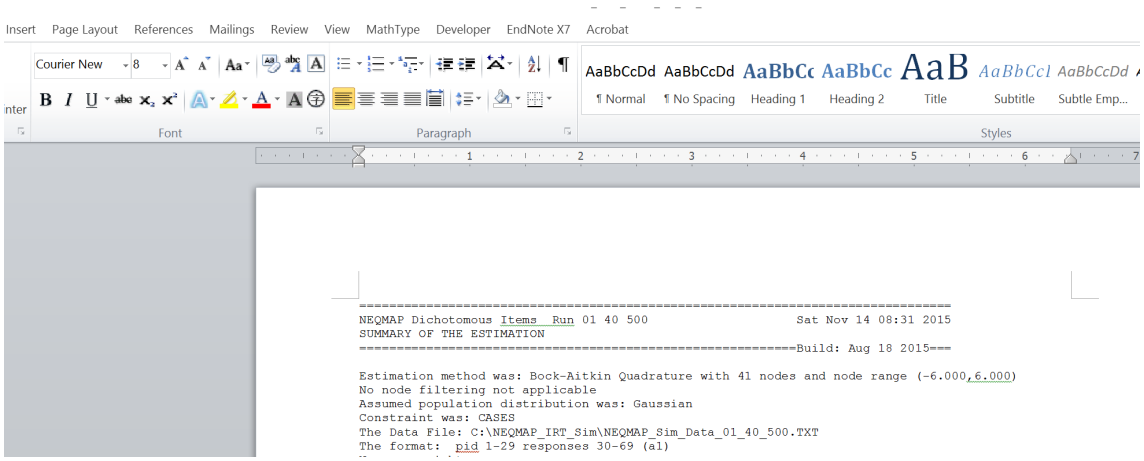

Figure 3.20: Reformed Microsoft word document

3.11.2 The first page

The first page of the table file has important information about the program and its settings (see 3.11.2). You should check that page, paying attention to the name of the data file, the format of that file and the other ConQuest specifications.

IRT_Sim_Data_01_40_500.tbl - Notepad

File Edit Format View Help

```
================================================================================
NEQMAP Dichotomous Items  Run 01 40 500                    Sat Nov 14 08:31 2015
SUMMARY OF THE ESTIMATION
============================================================Build: Aug 18 2015===

Estimation method was: Bock-Aitkin Quadrature with 41 nodes and node range (-6.000,6.000)
No node filtering not applicable
Assumed population distribution was: Gaussian
Constraint was: CASES
The Data File: C:\NEQMAP_IRT_Sim\NEQMAP_Sim_Data_01_40_500.TXT
The format:  pid 1-29 responses 30-69 (a1)
No case weights
The regression model:
Grouping Variables:
The item model: item+item*step
Slopes are fixed
Cases in file: 500  Cases in estimation: 500
Final Deviance:    21085.08503
Akaike Information Criterion (AIC): 21167.08503
Total number of estimated parameters: 41
The number of iterations: 9
Termination criteria:  Max iterations=500, Parameter Change= 0.00100
                       Deviance Change= 0.00010
Iterations terminated because the convergence criteria were reached
Random number generation seed:    1.00000
Number of nodes used when drawing PVs: 2000
Number of nodes used when computing fit: 200
Number of plausible values to draw: 5
```

Figure 3.21: ConQuest settings for the analyses

3.11.3 The item table

The item table has the item number (and its label if we had given the items one), an estimate of the Rasch item difficulty, and the standard error associated with that difficulty (Figure 3.22). Next there are two sets of fit statistics. Recall that the Rasch model is a strong model; these fit statistics are used to evaluate whether the data fit the model. More about these statistics will be provided later.

IRT_Sim_Data_01_40_500.tbl - Notepad

File Edit Format View Help

```
================================================================================
NEQMAP Dichotomous Items  Run 01 40 500                    Sat Nov 14 08:31 2015
TABLES OF RESPONSE MODEL PARAMETER ESTIMATES
============================================================Build: Aug 18 2015===
TERM 1: item
--------------------------------------------------------------------------------------------------------------------
TERM 1: item

   VARIABLES                            UNWEIGHTED FIT             WEIGHTED FIT
---------------                      -----------------------   -----------------------
     item           ESTIMATE  ERROR^   MNSQ       CI        T    MNSQ       CI        T
--------------------------------------------------------------------------------------------------------------------
 1   1               2.351   0.146    0.98 ( 0.88, 1.12) -0.4   1.01 ( 0.81, 1.19)  0.1
 2   2               2.394   0.148    0.96 ( 0.88, 1.12) -0.7   1.02 ( 0.80, 1.20)  0.2
 3   3               0.505   0.101    0.94 ( 0.88, 1.12) -0.9   0.98 ( 0.92, 1.08) -0.5
 4   4              -1.743   0.123    0.99 ( 0.88, 1.12) -0.2   1.02 ( 0.86, 1.14)  0.3
 5   5               0.607   0.101    0.94 ( 0.88, 1.12) -0.9   0.97 ( 0.92, 1.08) -0.7
 6   6              -1.804   0.125    0.98 ( 0.88, 1.12) -0.4   0.97 ( 0.86, 1.14) -0.3
 7   7              -0.664   0.102    0.99 ( 0.88, 1.12) -0.2   0.97 ( 0.92, 1.08) -0.7
 8   8              -1.208   0.110    0.86 ( 0.88, 1.12) -2.3   0.93 ( 0.90, 1.10) -1.4
 9   9               0.618   0.101    1.06 ( 0.88, 1.12)  1.0   1.04 ( 0.92, 1.08)  1.0
```

Figure 3.22: Item parameter information

3.11.4 The item-person map

The item-person map, also called the Wright map after Ben Wright (Figure 3.23), shows the position of each of the assessment items on the right-handside. The numbers are the item numbers and you could cross-check these with those in the previous table. On the left-handside you will see a lot of "X" letters. The key at the bottom of the figure states "Each 'X' represents 3.3 cases". That is, each X represents 3.3 simulated students.
This figure lets us assess how well the test items are distributed or targetted. For example, we can see that items 5, 9, 14, 21 and 33 all target the same point on the scale.

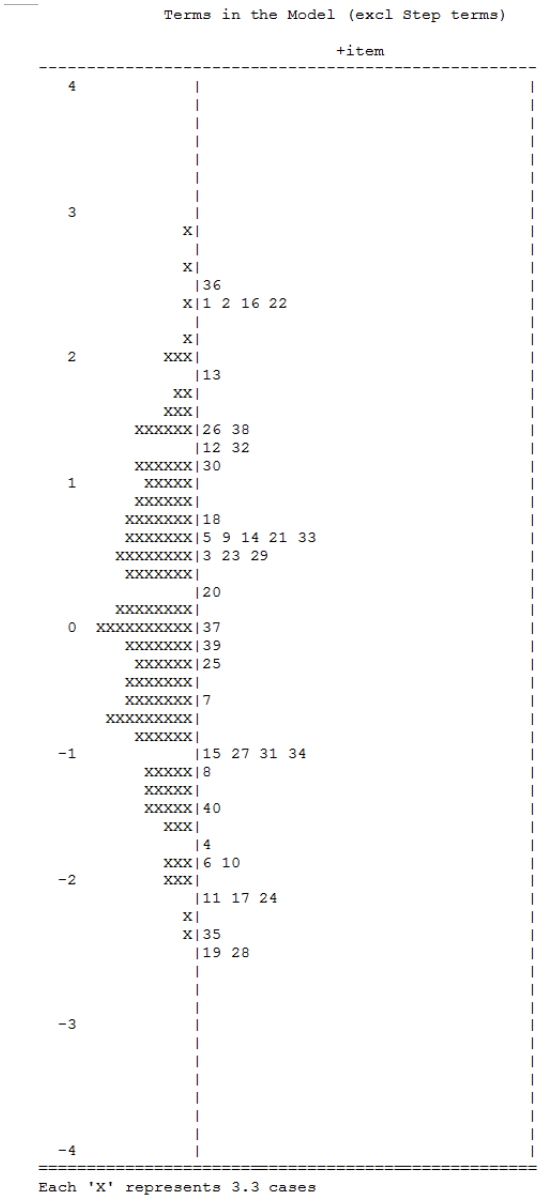

Figure 3.23: The item-person map

3.12 A Potential Resource: National Assessment of Educational Progress

The National Assessment of Educational Progress (NAEP) was the forerunner for many of the current international and national assessments. NAEP is a large scale, sample-based study of student achievement in the United States of America. The main website is:
http://nces.ed.gov/nationsreportcard/
If you explore the website will find a tool to access some very high quality questions from a number of curriculum areas and grades.
http://nces.ed.gov/nationsreportcard/itmrlsx/landing.aspx

You can select and download questions, answer keys and performance data from this site.

3.13 Books

Biggs, J. & Tang, C. (2011, November). *Teaching for quality learning at university (society for research into higher education)* (4th ed.). Open University Press.

Chappuis, J. (2002). *Understanding school assessment: a parent and community guide to helping students learn*. Assessment Training Institute.

Gipps, C. (2012, February). *Beyond testing (classic edition): towards a theory of educational assessment (routledge education classic edition)* (1st ed.). Routledge.

Greaney, V. & Kellaghan, T. (2007, November). *Assessing national achievement levels in education (national assessments of educational achievement)*. World Bank Publications.

Joughin, G. (Ed.). (2010, December). *Assessment, learning and judgement in higher education* (Softcover reprint of hardcover 1st ed. 2009). Springer.

Marzano, R. J. (2007, October). *Classroom assessment & grading that work* (1st ed.). Association for Supervision & Curriculum Deve.

Rowntree, D. (1987, April). *Assessing students: how shall we know them?* (2 Revised). Routledge.

Stiggins, R. J. (2007, July). *Introduction to student-involved assessment for learning, an (5th edition)* (5th ed.). Prentice Hall.

3.14 Articles

Ungerleider, C. (2006). Eflections on the use of large-scale student assessment for improving student success. *Canadian Journal of Education*, *29*(3), 873–888.