

*Professional Manual*

P E O P L E I M P R O V E  
P E R F O R M A N C E



Utrecht, November 2008

# Connector Ability 1.1

## Professional Manual

*Annette Maij-de Meij, PhD*

*Lolle Schakel, MSc*

*Nico Smid, PhD*

*Noortje Verstappen, MSc*

*Anela Jaganjac, MSc*



# Connector Ability 1.1

## Professional Manual

### Table of Contents

<i>Chapter 0</i>	<i>Introduction</i>
<i>Chapter 1</i>	<i>Theoretical background</i>
1.1	Testing and intelligence
1.2	Equal opportunity
1.3	Connector Ability 1.1
<i>Chapter 2</i>	<i>Contents</i>
2.1	Communication and instruction
2.1.1	Candidate brochure
2.1.2	Practice test
2.1.3	General and subtest instructions
2.1.4	Report
2.2	Characteristics of subtests and items
2.3	Administration
2.3.1	Administrative conditions
2.3.2	Time
2.3.3	Technical specifications
2.3.4	Adaptive process
2.3.5	Stop criterion
2.4	Scoring
2.4.1	T-score computation
2.4.2	Score interpretation
2.5	Use
<i>Chapter 3</i>	<i>Construction</i>
3.1	Instructions
3.2	Items
3.2.1	Item pool construction
3.2.2	Parameter estimation
3.2.3	Item parameters
3.2.4	Item information
3.2.5	DIF analyses

- 3.3 Group differences
  - 3.3.1 Group differences construction sample
  - 3.3.2 Group differences selection sample
- 3.4 (Sub)test correlations
- 3.5 Norm development
  - 3.5.1 Norm Sample
  - 3.5.2 Norms

*Chapter 4 Psychometrics*

- 4.1 Reliability
  - 4.1.1 IRT-based reliability
  - 4.1.2 Test-retest reliability
- 4.2 Validity
  - 4.2.1 Construct validity
  - 4.2.2 Criterion-related validity
  - 4.2.3 Discriminant validity
- 4.3 Adverse impact

*References*

*Appendices*

- A Candidate Brochure
- B Best Practice Guidelines
- C FAQ Top 10
- D Example of Test Report
- E Online Testing Process; an illustrative example

# Chapter 0

## Introduction

### Test use and the need to manage diversity

Managing 'diversity' is increasingly becoming a priority for government and business. The fact that the labour market is becoming tighter is an additional reason for making the most of available talent. Generating and safeguarding equal chances when comparing candidates and employees for access to jobs and/or opportunities for personal development is an important aspect of this. This makes it vitally important for organizations to select efficiently and fairly.

A growing number of employers are making use of assessment centres and psychological tests when selecting staff. The recruitment pool is becoming ever more diverse in terms of age, gender, national and cultural background. Multinational companies are increasingly recruiting from the international labour market and tests often have to provide the decisive answer in cases where diplomas are difficult to compare. Besides this, there is an increase in cultural diversity among the professional population in many countries worldwide. To be able to guarantee that selection has taken place fairly, tests have to measure purely what they have been designed to measure, without benefiting certain groups or putting others at a disadvantage.

Much attention is being given to the use of tests and to cultural diversity. In this context, the Netherlands National Bureau against Racial Discrimination (LBR) recently issued two publications in conjunction with the Netherlands Institute of Psychologists (NIP) that offer "Guidelines for the use of diagnostic instruments among ethnic minorities" and an insight into the extent to which current tests are applicable in this context (Bochhah, Kort, & Seddik, 2005a; 2005b).





# Chapter 1

## Theoretical background

### 1.1 Testing and Intelligence

#### Measuring the G-factor

Referring to the Schmidt and Hunter (1998, 2004) and Gottfredson (2002) research, the general factor underlying a broad range of intelligence tests (the G-factor or just G for short) not only generalizes over a large heterogeneous number of jobs and work contexts, but it might be expected also to be culturally independent. Even the heated discussion around the much-debated book of Herrnstein and Murray 'The Bell Curve' (1994) does not lead to conclusive evidence. It amounts to at most a weak conjecture of possible small racial differences in G. Therefore one may safely keep oneself to the hypothesis of only differences between cultures in culturally bound substance in tests which is *not* related to G. An important implication of the foregoing is that the *criterion* whether an intelligence test purely measures G actually amounts to showing that it is culturally unbiased. Below it will be argued that the intelligence test described in this report (Connector Ability 1.1) does not show practically relevant cultural differences, and may therefore be regarded an adequate measure of G.

Restricting oneself to G as the general variable in using intelligence as a predictor in organizational contexts has a well supported empirical basis. Kline (1992), in a summary of research up to then which is still the generally accepted view, distinguishes G in two subcomponents:

- Fluid (F), referring to so called 'pure' intelligence not disturbed by cultural differences.
- Crystallized (C) which measures components that are partly influenced by a person's cultural specific knowledge and skills.

F generally is measured in so-called 'culture-reduced' tests. Still, also crystallized intelligence will be attended here in order to cover G in both aspects mentioned by Kline. But then subtests will be selected which are as little as possible dependent on specific cultural or language knowledge. Therefore the subtests to be chosen for measuring aspects of crystallized intelligence will not be based on language or vocabulary directly (e.g. Kowall, Watson, & Madak, 1990; Naglieri, & Ronning, 2000), which is almost always the case by most intelligence tests (Mackintosh (1998): "They appear to be measures of knowledge, not ability..." p.280).

### **The role of language and culture**

The concept of intelligence refers to differences between individuals in the speed and accuracy with which they are able to solve new problems in new situations. To be able to measure that abstract difference, however, people will have to be presented with real problems. Just as you have to get someone to run on a real track to be able to estimate his or her running capability.

A condition for measuring intelligence is therefore that you create an equal starting position for test persons on three aspects: present them an equal set of problems (that is the test); take care they have an equal prior knowledge of substantive content that is required for making the test, and give them the same amount of time with this equal set of problems. If, under such equalized conditions, two people differ in the number of items they answer correctly, that difference is by definition attributed to differences in intelligence.

### **Three categories of tests**

Concrete problems in daily life generally appear to be formulated and solved by way of three 'channels': abstract-spatial symbols (drawings), numerical symbols (numbers), and verbal symbols (words / texts). That is why an intelligence test usually includes these three elements. However, if you want to measure G, then you don't necessarily have to use all three of these channels. Particularly when you want to compare two test candidates who speak different languages, it is wise to limit the test to abstract-spatial and numerical. You then only have items that require no knowledge whatsoever of a specific language.

A comparable argument can be made referring to knowledge of a specific culture. If everyone in a certain culture knows that you have to stop at a red traffic light, you can safely make an item in which that knowledge is assumed. Respondents from that culture are then equalized on that point. However, when you want to compare two people from different cultures, you will have to check very precisely that there are no calls on culture-specific knowledge that one person knows and the other does not.

### **Difference between instructions and test items**

The test score is determined in every test on the basis of the answers to the items within the time allotted. The time you need to be instructed in what to do during the actual test has no influence on the score.

In a test free of language or cultural bias, therefore, you can safely give instructions in the various native languages of the different respondents. As long as you make sure in the procedure that each respondent knows exactly what to do at the moment he or she starts working on the actual test items. In such cases it may be expected that systematic differences between language and cultural groups are hardly or no longer present. After all, evolution does not select for intelligence along national borders.

### **Intelligence and competency**

Knowing many words or analogies between words is not in itself intelligence. Neither is learning all the square numbers from 1 to 100. Both are competencies: you can do something. Organizations like to select people who can do things.

An intelligence test, however, does not aim at measuring whether you know or can do something specific, but whether you will be able to solve a new problem in a new situation. That 'new' aspect predicts whether you are able to learn.

It is not only sensible but also practical to make this distinction. Because an organization ought to ask itself whether it would rather bring in people with the potential to solve new problems in the future or to learn new skills, or only people who are able to do something now.

### **So why are there vocabulary tests or analogy tests in intelligence tests?**

That's because *within one and the same language community* the more intelligent people usually have a wider vocabulary. This is therefore a competency, which highly correlates with intelligence, and as such is useful in a test. However, when you want to compare the intelligence of two people with a different language background, a vocabulary test is not advisable. As stated above, you have to measure intelligence by means of solving concrete problems but before that, you first have to equalize the prior knowledge required to do so. If that is not possible in a particular channel, in particular the verbal channel, then you should not use that channel.

## **1.2 Equal opportunity**

### **Testing with ethnic minorities**

In the LBR-NIP publications (Bochhah et al., 2005a; 2005b) it is argued on the basis of extensive research that the following three aspects in particular should be controlled for when testing with ethnic minorities:

- It should be thoroughly known to test-persons what a testing situation in general is all about, the way the specific test is to be made and what exactly is expected from them. Test persons from ethnic minority groups often have little or no experience with a testing situation, feel themselves therefore unsure which may negatively affect their scores.
- Of course, the instruction what to do in making the test items will have to be given in a specific language. However, the level of language competency required (when it is not possible to do the test in the test persons own native language) should not be higher than the level needed to have simple everyday conversations. Especially, the used vocabulary should restrict itself to the most common words, as well as avoid words with a culturally specific meaning. Furthermore, care should be taken that any test person who has to be instructed in another language than his mother tongue, is allowed to choose himself the instruction language which suits him best.
- Any G-test should restrict itself to subtests which not directly demand more than elementary vocabulary knowledge.

### **One test for minorities and majorities**

One problem is that although a few tests have been specifically developed for (ethnic) minorities, test producers generally have no tests available that are sufficiently free of cultural bias to enable both the (ethnic) majority and (ethnic) minorities to justifiably take the same version of the test. To be able to make a true comparison and provide fair chances, it must be possible to use one and the same test for both groups.

Conditions for using such a test justifiably and efficiently are:

- Reliable prediction
- Equal opportunities
- Equal test programs
- Selection that matches the candidate's level

### **Reliable prediction**

As stated earlier, being able to accurately estimate the general level of cognitive ability is the most important predictor as regards selection procedures and for predicting career development compared to other predictors such as work experience and personality (Schmidt & Hunter, 1998; 2004; Gottfredson, 2002).

It is important to observe here that the generalized predictive power of cognitive ability relates in particular to the general cognitive level and not so much to the separate sub-capacities such as figural, arithmetic and verbal ability. Limiting them to the areas not necessarily linked to language such as spatial, numerical and logical/abstract ability can still lead to a good estimate of the general level if the test is long enough and the test construction sufficiently accurate.

### **Equal opportunities for applicants and employees**

Tests used to identify cognitive ability must give each participant an equal chance. Both differences in cultural background in general and differences in language skills in particular should not seriously affect the estimation of the general level of cognitive ability. However, ethnic minorities are readily put at a disadvantage when they take standard tests in a non-native language. There is therefore an urgent need for tests free from cultural bias.

Furthermore, research into the discriminatory aspects of ability tests shows that actually mainly the aspects that relate to language, whether this involves the instructions or the actual content, might have a serious biasing effect on test scores (see also the LBR-NIP publications previously referred to).

### **Comparability based on equal test programs**

Against this background, it is also important that specific tests should not be used for specific groups, but that *all* people being tested, regardless of their ethnic or cultural background, are given the *same* test. Tests specific to subgroups, intended to prevent discrimination, often

subsequently create their own comparison problems, due to confounding differences between subgroups with differences in test content. Besides this, tests in selection contexts must be efficient and quick to take. If they are to be used regularly, they will soon involve large-scale testing procedures.

#### **Efficient selection specific to the candidate's level**

An important condition for efficiency is that the person taking the test is only presented with items that are neither too difficult nor too easy. Items that are too difficult or too easy not only produce little information, but also lead to unnecessary confusion in the mind of the person taking the test as to its relevance.

When a test procedure like the one just described is chosen, each person being tested is given his or her own set of items taken from a large collection of all available items. This also minimizes the risk of items becoming generally known and maximizes the ability to compare results.

#### **Adaptive testing is the answer**

To be able to measure cognitive ability quickly and efficiently as part of the selection or pre-selection for jobs or development processes, the maxim always should be 'the right person in the right place' regardless of differences in cultural and other backgrounds. Being able to compare general cognitive ability individually is then a first requirement.

If an adaptive online test is available, this demand can be met efficiently, where necessary on a grand scale, and flexibly. Former limitations in terms of the time and place the test is taken, travelling time and/or availability of a test room are no longer an impediment. This makes an adaptive online test for measuring the general level of cognitive ability an extremely suitable instrument for both fair and non-discriminatory selection and pre-selection .

There are models generally available for so-called 'adaptive' tests that can be used to construct a test for general cognitive ability that meets the conditions described. These are tests constructed on the basis of what is known as 'item response theory', IRT for short. The reader is referred to general summaries for their background and ways of working. One example is Van der Linden and Glas (2000).

#### **Which practical benefits for diversity management will adaptive tests produce?**

Characteristics that lead to an adaptive test meeting the conditions described above are:

- *Engendering trust in the test and a serious attitude to the test*

Each person being tested is given items that in each case provide the most information about his or her general level of ability at that moment, given his or her answers to test items up to now. That item is therefore neither too difficult nor too easy at that particular moment. This will engender trust and a serious attitude.

- *Independent of levels of education*  
There is no need for separate tests for different levels of education. Each subtest (for example, a numerical test like a Series of Numbers) can be compiled as one long set of similar items ranging from very easy (lower vocational education) to very difficult (university level). This is certainly an important advantage when comparing ethnic majorities and ethnic minorities in a non-discriminatory manner. When testing such groups comparing different levels of education is generally problematical.
- *Fast, continuous safeguarding of fair measurements*  
Dynamic adjustments can be made to the test while it is being used, by adding new items and removing old ones. Items already in the test can be monitored to assess the extent to which they seem to be non-discriminatory. Items that score less well on that point can be changed or removed. A person's general level can still be effectively estimated even if such an item is removed.

### **Qualification of users**

The qualification of the user of Connector Ability 1.1 depends on the context of use. 'User' is defined here as the individual who discusses the content of the report with the test person. At the standard level, a user should be able to explain to a test person the meaning of the report and the consequences of it for selection or development. To that end, besides knowledge of the context in which the test is used, the user should have relevant knowledge on both background and meaning of the test itself and structure and text of the report. Furthermore, he should have the interviewing skills for having proper feedback sessions. PiCompany demands certification on these knowledge and skills as a condition for an allowance to use the test. This certification is based on a successful completion of a certification training specifically focused on Connector Ability 1.1. Information on training form and content may be found in Section 2.5.

In principle, every person who has a role as a manager or an individual HR professional in an HR process like selection, training or career guidance is eligible for such a certification training, irrespective of earlier academic qualifications.

If Connector Ability 1.1 is used as a part of a more dedicated personal development context as, e.g., in assessment or development centres, certified knowledge of and skills in applying the rules of the International Test Commission should be established. The professional qualifications of a registered psychologist generally will be based on these rules.

## 1.3 Connector Ability 1.1

For the development of a language and culture fair test for G it is good practice to take as a starting point a test for measuring G that has already demonstrated its quality for measuring G in a specific cultural context. The intelligence domain has such a firm conceptual and empirical base already that one should use this base.

In the present context the already existing PiCompany test Connector C 3.1 (PiCompany, 2005) has been used as the basis for developing the new language and culture fair G-test. For the substantive and operational aspects of Connector C 3.1, reference is made here to the professional manual of Connector C 3.1(2005). The new test reported here is called Connector Ability 1.1.

Connector Ability 1.1 differs from Connector C 3.1 on the following aspects:

- It is based on a subset of the subtests from Connector C 3.1, especially the ones that make minimal use of crystallized intelligence aspects that are supposedly culture specific.
- It uses only symbols and words which are expected to be culturally universal.
- It is constructed and used in practice with IRT methodology (Van der Linden & Glas, 2000). So, the test is adaptive in a way as described above. The specific adaptive models chosen for constructing and using the test in practice, as well as the arguments for the choices made, are described in detail below.

### Choice of subtests

Considering the subtests of Connector C 3.1, the following subtest categories are chosen as a starting point for constructing subtests for Connector Ability 1.1 (for more details see Section 2.2).

#### *Fluid intelligence (F)*

- Matrices
- Series of Figures

#### *Crystallized intelligence (C)*

- Series of Numbers
- Diagrams

In order to minimize culturally specific content, a focus group of interculturally knowledgeable experts have reviewed test instructions and test content in the just mentioned subtests of Connector C 3.1 (especially as regards specific items) on potentially biasing content. In constructing Connector Ability the new items have all been screened thoroughly on the same aspects by the same experts before adding those to the set of items to be piloted in the trial version of the test. A number of criteria have been specified for the construction of items for which the cultural influences are minimized.

**Target group and context**

Connector Ability 1.1 is meant to be applicable for all educational levels in principle. Because it is adaptive, eventually all items of each subtest will span one single underlying ability dimension for the whole human population.

The present version Connector Ability 1.1 restricts itself to measuring the subpopulations of persons who are comparable to persons with a mid-level education, bachelor or master level as far educational background is concerned. The test will be applicable in its first version to three norm populations: mid-level education (ME), bachelor (BA) and master (MA). The primary application domain is selection in organizational contexts.



# Chapter 2

## Contents

This chapter describes the contents of Connector Ability 1.1. The design of instructions, subtests and items is described. All administrative conditions are described, including the adaptive procedure. Furthermore, information with respect to scoring and training of users is given.

### 2.1 Communication and instruction

All information that is communicated to the candidate is described. A candidate receives a candidate brochure and may take the practice test. Connector Ability includes general and subtest instructions and results in a test report.

#### 2.1.1 Candidate brochure

A candidate brochure is available for each candidate, see Appendix A. In the candidate brochure, first the purpose of the test is explained. Second, the brochure contains the instructions of the test: the general instruction and the instructions for each of the subtests, including a set of sample items for each subtest. The candidate brochure is available in both a paper version and an online digital version. In the communication that precedes the actual test, it is made sure that each candidate is being sent or has access to the brochure and is referred to the online practice test (see Section 2.1.2).

The candidate brochure provides the candidate with the opportunity to calmly get acquainted to the content and nature of the test and to what is to be expected, without immediately being confronted with the online application. This adds to the opportunity to be able to prepare for the test and to practice beforehand.

#### 2.1.2 Practice test

An online practice test is available for each candidate. The online practice test contains for a large part the same information as does the candidate brochure (explanation of the purpose of the test, the general instruction and the instructions for each of the subtests, including sample items), but also a number of items are added that may be answered by the candidate as a real trial. Thus the candidate actually takes a 'mini-version' of the test. Also, a brief report is made,

based upon the answers the candidate has given on the real trial. This report is sent to the candidate.

In the communication that precedes the actual test, it is made sure that each candidate is being sent or has access to the candidate brochure (see Section 2.1.1) and is referred to the online practice test. This test can be found via a web link on the PiCompany website ([www.picompany.nl](http://www.picompany.nl)).

The practice test provides the candidate with the opportunity to get acquainted in a relaxed pace to the content and nature of the test and to experience what it is like to take the test on the computer and answer items in the actual application. This adds to the opportunity to be able to prepare for the test and to practice beforehand and also to get acquainted with and experience more of the look and feel of the actual test on the computer.

### **2.1.3 General and subtest instructions**

The instructions of the test contain both a general part and a specific part for each subtest.

#### **General instruction**

Each candidate first receives the general instruction. In the general instruction it is explained that the test consists of several different parts and an illustration is given of what the screens look like and how the test works, by both text and visual examples. It is also explained in the general instruction that for the actual test items a limited amount of time will be available, in which the candidate will have to choose an alternative. With this information, a visual image is shown of how to recognize on the screen (when taking the actual test) when the allotted time to choose an alternative is nearly used up. It is also stressed that the candidate can take as much time (s)he needs for the general and subtest instructions.

The general instruction provides each candidate with the opportunity to get accustomed in a relaxed pace to the way the test works and to what will be asked later on in the actual test items and how to deal with this. After having exited the general instruction, the instruction of the first subtest can be started.

#### **Subtest instructions**

For each of the four subtests, a specific subtest instruction is offered. Each of these specific subtest instructions consists of:

- an explanation;
- two sets of sample items; each set consisting of three sample items.

The candidate can go through the subtest explanation and sample items at his/her own pace. In the subtest explanation, the character and content of the specific subtest as well as each of the item types of the subtest, are explained by both text and visual images. The subtest explanation is followed by a first set of three sample items, which each candidate has to

answer. The candidate's answer is followed by feedback. The feedback consists of: a visual image of the item and the correct alternative, accompanied by a textual explanation of why this is the correct alternative. A correct answer is followed by the (above-mentioned) feedback/explanation and is subsequently followed by the next sample item. An incorrect answer is followed by the same feedback, as is the correct answer and then by the same sample item being offered again and subsequently being explained again.

Once the first set of sample items is completed, it is verified whether the candidate understands what is expected and is ready to start with the actual test items. At this point, there are two possible routes:

1. The first route is for the candidate to now go directly to the actual test items of this subtest.
2. The second route is for the candidate to first go to the second set of sample items and then go to the actual test items of this subtest.

The sample items of the second set are similar to the sample items of the first set in the sense that they both contain the same item types, but they differ in the sense that they are slightly easier to solve than the sample items of the first set. Candidates who have answered two or more of the sample items of the first set *incorrectly*, will automatically go to the second set of sample items (they are directed to the second route). This is because these candidates did not yet arrive at the desired starting position and are likely to benefit from extra practice. Candidates who have answered two or more of the sample items of the first set *correctly*, are given a choice. They can either choose to go directly to the actual test items of this subtest (follow the first route) or they can choose to answer the second set of sample items before going to the actual test items of this subtest (take the second route). Thus the candidate decides him-/herself whether (s)he feels ready to start or prefers to practice some more. This choice specifically contributes to the feeling of control and security of candidates who suffer from test anxiety or learn more slowly, and of course also to others who prefer to have more time and practice. The second set of sample items has the same structure as the first set: a correct answer of the candidate is followed by feedback/explanation and the next sample item, an incorrect answer is followed by feedback, the same sample items being offered again and subsequently being explained again.

Once the sample items are completed (either one or two sets), the candidate can start the actual test items of the specific subtest. The time will not start to run until the candidate starts the first item. During the test, the 'Help' button can be activated at any given time. This prompts a screen with a short explanation of the specific subtest. This short explanation is a summary of the elaborated subtest instruction the candidate has been offered before. The time keeps on running when requesting this short explanation. The candidate taking the test knows the time is still running because this is indicated in the short explanation. The short explanation serves the purpose of quickly triggering the subtest information that was learned before which may help to recognize the patterns and answer the items correctly.

### 2.1.4 Report

In Appendix D, a dummy report of a fictive candidate, the so-called 'Bert Smith', is shown. The textual information in the report is explained in non-technical straightforward language that can be understood and explained to the candidate by the intended user. The intended user is a person who has successfully completed the certification program (see Section 2.5). The computation of T-scores as well as score interpretation is described in more detail in Section 2.4.

The primary use of Connector Ability is in a selection setting. A selection decision needs to be based on the score of the candidate on the G-factor. This point is stressed in the certification program. Nevertheless, candidates often value knowing how they performed on the different subtests. Therefore, the scores on the four subtests are given to provide more detailed feedback to the candidate.

The report is a fixed format. The only flexible component is the norm group which is chosen beforehand and the T-score that is reflected in the report, which is based upon the comparison to this norm group.

## 2.2 Characteristics of subtests and items

Measurement of the G-factor is based on the scores achieved in the subtests: Series of Figures, Matrices, Series of Numbers, and Diagrams. Each subtest measures the ease with which someone can:

<i>Series of Figures</i>	Complete logical reasoning;
<i>Matrices</i>	Analyse and continue complicated relationships;
<i>Series of Numbers</i>	Analyse and continue the relationship between numbers;
<i>Diagrams</i>	Make connections between concepts.

### Item format

The question that is asked within one subtest remains the same for all items in the subtest.

For the four subtests these questions are respectively:

<i>Series of Figures</i>	Which figure most logically continues this series?
<i>Matrices</i>	Which figure most logically continues this matrix (bottom right)?
<i>Series of Numbers</i>	Which number most logically continues this series?
<i>Diagrams</i>	Which figure best describes the relationship between these three concepts?

Each item consists of a problem. The candidate has to solve this problem by choosing one out of four alternatives. The chosen alternative can always be changed by clicking on a different

alternative within the time limit (see 2.3.2). It is not possible to return to an item at a later stage.

For each subtest, characteristics can be defined that may be varied across items. Thus, each item is constructed by combining a specific number of these characteristics. The construction of the items is furthermore restricted by a number of additional criteria. These criteria were formulated prior to and during the construction of the test and are defined in such a way that cultural bias is minimized. The final set of criteria that characterizes the item pool is described below, for each subtest separately. The practice test on the internet provides examples of items that are included in Connector Ability 1.1 (see [www.picompany.nl](http://www.picompany.nl) for access to the practice test).

### **Series of Figures**

The items in this test consist of a series of four figures. A systematic change takes place in each subsequent figure in this series. The candidate has to choose the (fifth) figure from one of the four alternatives that most logically continues the series of four figures.

With respect to figures and transformations, the following criteria were defined:

- Each figure in the series can be regarded as one cell, or can be divided into nine cells. This characteristic applies to all figures in the series of one item.
- Only basic geometrical figures that are known worldwide have been used as construction elements.
- Only basic transformation rules have been used to construct the different items:
  - Rotation.
  - Size.
  - Colour (black, gray, white).
  - Type of figure.
  - Contents (stripes, dots, empty).
  - Location.
  - Line thickness.
  - Combinations of the aforementioned transformation rules.

The items were constructed in a systematic way, so as to represent a various set of all kinds of figures and transformation rules.

### **Matrices**

In this subtest a matrix containing eight images is presented in each item. A regular change takes place in these eight images, both horizontally and vertically. The candidate is asked to complete the matrix with a ninth figure that follows logically from the other figures both horizontally and vertically. The candidate can choose from four alternatives.

Criteria for the construction of matrix items were:

- Each one of the (nine) cells of the matrix containing an image can be regarded as one cell or can be divided into four cells.
- Only basic geometrical figures known worldwide have been used as construction elements.
- Only basic transformation rules have been used to construct the different items:
  - Rotation.
  - Size.
  - Colour (black, gray, white).
  - Type of figure.
  - Contents (stripes, dots, empty).
  - Location.
  - Line thickness.
  - Number (addition/subtraction of figures).
  - Three different figures may alternate by row or column.
  - Combinations of the aforementioned transformation rules.
- Solving the item is done by discovering the logic in the matrix. Items in which the candidate has to count lots of stripes unnecessarily, for example, are avoided.
- The matrix forms a logical series both horizontally and vertically in each item.

Items were generated systematically by varying the different operations and transformation rules.

### **Series of Numbers**

In this subtest a Series of Numbers is shown in each item. The numbers succeed each other logically. The candidate has to choose the number from one of the four alternatives that most logically continues the Series of Numbers.

Possible changes in numbers are: addition, subtraction, multiplication, and division. One change may take place in a Series of Numbers, from one number to the next. Also, two changes may take place, where one change occurs from the first to the third number, and another from the second to the fourth number. So two changes take place. A Series of Numbers may contain either four or six numbers in an item.

Criteria that were taken into account during test construction are:

- Only simple arithmetic skills at ground school level are required.
- Multiplication by zero does not appear in the series.
- Very large, difficult numbers are avoided.
- No more than two calculations are included at each stage.

Items were generated systematically by varying the different operations and transformation rules.

## Diagrams

Three concepts (words) are presented in the items in this subtest. Each concept represents a set. Candidates have to decide whether the three sets overlap or are completely separate. The alternatives show the relationship between three concepts by means of three circles (Venn-diagrams). Neither the relative size of the circles nor the order of the words that are given is important. Candidates have to choose the alternative in which the positions of the circles best portray the connection between the three concepts. The candidate, again, can choose from four alternatives.

Criteria for the words that are used in the items are:

- Only concepts and words are used which have supposedly worldwide the same empirical reference. This has been checked by the intercultural focus group referred to earlier. E.g. Culture-dependent concepts or relationships between concepts (such as 'dress – female' or 'winter – snow') were to be avoided.
- All words that are used are unequivocally translatable between any two languages. This has also been checked by the mentioned intercultural focus group.
- Words are singular nouns or adjectives that state a property without a norm or scale.
- Difficult words whose meaning is not clear to everyone are avoided.
- Professions are not mentioned in combination with the concepts 'men' and 'women'.
- Verbs or adverbs are not used.

The construction of unambiguous items has proven to be difficult. Only one alternative should be the correct alternative, there should not be any debate possible. Nevertheless, the items also have to vary in difficulty, where specifically difficult items have shown to be hard to make. To guide the construction and review of the items, two types of relationships among the concepts were defined to be allowed:

- 1 A word is an attribute or a component of another word. Attributes are for example colour, material, size etc. Examples of components are 'minute' as a part of an hour, where 'stairs' may be part of a building.
- 2 A word is 'a kind/type of' another word which is a broader concept. One can think off a 'cow' as a kind of animal, for example, or a 'dress' as a kind of clothing.

The eleven possible combinations of circles are drawn. For all combinations of circles, sets of concepts were written. Above, several criteria are given for selecting the concepts. Also, several criteria were set that show properties that are *not* allowed in an item.

- The type of relationship may not be a word representing an object that may be 'inside' another object (for example, a chair may be in a house, but this is not a permitted relationship).
- For some combinations of words it is not possible to draw their relationship using the circles. For example, the words 'house' and 'roof'. A roof is part of a house, furthermore, each house has a roof. However, roofs may also be elsewhere. Therefore, this is a combination of words for which the relationship cannot be drawn.

These conditions drastically narrow down the amount of possible combinations of words. However, the restrained conditions are necessary for the construction of unambiguous items. Adjustments of the criteria and conditions were made during the several (pilot) studies. The criteria mentioned above are the result of all test construction experiences. This means that at the start of the data-collection, not all items met these criteria. Of course, these items are not included in Connector Ability 1.1.

## **2.3 Administration**

Administration comprises the administrative conditions, time and technical specifications. Furthermore, information is given with respect to the adaptive process.

### **2.3.1 Administrative conditions**

Connector Ability 1.1 is administered online under supervision of a test assistant. With a candidate specific login and password, the test assistant logs on to the computer, after which the candidate can start the test. The items appear on screen, after which one has to choose from four possible alternatives. This is done with the computer mouse. Pen and paper are available. These are also the only aids that one may use during the test and which are to hand in at the end.

Before someone starts with the test, one is asked to indicate whether his or her personal data are correct. These data refer to their name and date of birth. Next, the person is asked to provide some more data on their personal background, for example on education, gender and on their own and their parents' country of birth. This information is used for research purposes only. These background data are processed anonymously and are not in any way used in reporting or in interpreting test results. Protection of personal details and other personal and test information is guaranteed. Further information can be found in Candidate Brochure (Appendix A) and Best Practice Guidelines (Appendix B).

### **2.3.2 Time**

The number of items a candidate receives in each subtest depends on the answers given. The computer program offers items until it has been able to estimate the problem-solving ability based on the given answers. For each item one has to choose an alternative within a time limit of 90 seconds for the subtests Series of Figures, Matrices and Series of Numbers. For the subtest Diagrams, the time limit is set to 45 seconds. A time bar is shown at the top of the screen, which will start running when the final 20 seconds of the given item start running.



The exact amount of time needed to complete the test depends on the number of items a candidate has to respond to. Each subtest consists of a maximum of 15 items. This results in a maximum duration of almost 80 minutes. The time span needed beforehand for the instruction (to read the explanation and answer the sample items) is excluded from this calculation.

### 2.3.3 Technical specifications

Some technical specifications need to be met with respect to the computer that is used to successfully administer Connector Ability 1.1. These technical specifications are stated below under 'computer requirements'. Also, a number of specifications are formulated concerning the internet connection.

Minimum computer requirements are:

- Processor: 1,5 GHz or higher
- Memory: 512mb or higher
- OS: Windows XP or Vista
- Resolution: 1280x1024 pixels

Minimum requirements concerning the internet connection:

- |              |             |                  |
|--------------|-------------|------------------|
| – 1-5 users: | 0.1 megabit | 300 connection   |
| 5-10 users:  | 0.2 megabit | 600 connections  |
| 10-20 users: | 0.4 megabit | 1000 connections |
| 30 users:    | 0.5 megabit | 1500 connections |
- Permanent and uninterrupted connection is required.
  - Use of content scanning or https has negative effects on the connections and speed, and is therefore dissuaded.

### 2.3.4 Adaptive testing

Connector Ability measures cognitive ability in an adaptive way. This means that each individual receives an item, which at that moment will be the most informative with regard to this individual's cognitive ability taking into account his/her previous answers on the test. Therefore, at that moment this item will neither be too difficult nor too easy for this individual.

In computerized adaptive testing, the so-called theta value of an individual is estimated. At each stage of the administration of the (sub)test, an item is selected from the item bank based on specified criteria. A theta estimate is computed based on the item responses on the items that have been administered so far. The theta estimate and the item responses are used in

the selection of the next item, and so on, until some pre-specified criterion is met (Van der Linden & Glas, 2000).

Below, first the estimation of theta values is explained. Subsequently, the procedure of item selection is described. In the next section, the stop-criteria are described.

### Estimation of theta values

An item response theory (IRT) model needs to be specified to estimate theta values. The item parameters of this IRT model are also used for item selection, which will be discussed below. The IRT model specified for Connector Ability is the Two-Parameter Logistic (2PL) Model. The probability of a correct response  $x_i$  to item  $i$  given the theta ( $\theta$ ) value of a person is given by;

$$P(x_i = 1 | \theta) = \frac{1}{1 + \exp[-\alpha_i(\theta - \beta_i)]}$$

where  $\alpha_i$  is the item discrimination parameter and  $\beta_i$  is the item difficulty parameter (Van der Linden & Glas, 2000). Given a number of  $k$  items, the likelihood function for theta can be written as;

$$L_k(\theta) = \prod_{i=1}^k P_i^{x_i} Q_i^{1-x_i}$$

where  $Q_i^{1-x_i} = 1 - P_i$  is the probability of a wrong response. The maximum likelihood estimate (MLE) of theta is the value of theta that maximizes the likelihood function for a particular item response pattern. This can be computed by fixing the first derivative of the likelihood function equal to zero.

The estimation of theta values is based on the item responses of a person. With each administration of an item, the estimate of the theta value is updated by estimating it on all available responses. The accuracy of the theta estimate can be inspected by investigating the standard error of estimation, as will be explained below.

### Item selection

In each subtest, the first item is selected at random from a set of three items with an average difficulty level. When an item response pattern consists of only correct or incorrect responses, it is not possible to estimate a theta value, as it goes to plus or minus infinity. This means that after administering a first item, it is not possible to estimate a theta value. Therefore, a theta value of plus (correct answer) or minus (incorrect answer) 0.70 is used to be able to base item selection on maximum information, as is described below. This procedure is derived from research by Dodd, Koch, and De Ayala (1989), with the addition of simulations as their research is based on an item pool that has an ideal distribution of items across the theta continuum.

These so called 'step sizes' of 0.70 are repeated, until a minimum of one correct and one incorrect answer has been given. At that point, theta may be estimated according to MLE as described above, based on all responses available. Simulations showed that smaller step sizes decrease efficiency, whereas larger step sizes increase the probability that the item is far too difficult or too easy for a candidate.

Generally, item selection is based on maximum information. An item is selected that maximizes the Fisher information (see for example Van der Linden & Glas, 2000) for a given value of theta. For the 2PL model, the information function for item  $i$  and a theta value,  $\theta$ , is given by;

$$(1) \quad I_i(\theta) = a_i^2 P_i(\theta)[1 - P_i(\theta)]$$

where  $a_i$  is the discrimination parameter of item  $i$ ,  $P_i(\theta)$  is the probability of a correct response under the model, and  $[1 - P_i(\theta)]$  is the probability of an incorrect response. The item information expresses the contribution an item can make to the accuracy of the measurement of an individual as a function of his/her ability.

The test information of a test with  $k$  items is equal to the sum of the item information of these  $k$  items;

$$(2) \quad I(\theta) = \sum_{i=1}^k I_i(\theta)$$

This means that the more items there are in a test, the greater the amount of information. The amount of test information can be translated into a standard error of estimation:

$$(3) \quad SE(\theta) = \frac{1}{\sqrt{I(\theta)}}$$

The standard error of estimation gives information about the precision of the estimate of theta. It quantifies the variance in the estimated theta value, that would be expected when a measure is administered repeatedly to a candidate, without the candidate remembering his/her previous administrations. The greater the amount of information, and thus the smaller SE, the greater the precision of the theta estimate. This characteristic is used as a stop criterion, as will be described in Section 2.3.5.

Note that after each item administration, the theta value is updated, which is then used to compute the standard error of estimation. The subtest is ended when the stop-criterion is met, or a next item is selected based on maximum information given this new value of theta. All computations of the program have been checked independently, see also Appendix E.

### 2.3.5 Stop criteria

Two criteria were specified to determine when the subtest would be terminated:

- A subtest will be terminated when the standard error of estimation is below 0.54. The standard error of estimation is a function of the test information as shown in Equation 3. When the standard error is below this value, the estimate of theta is made with enough precision. A standard error value of 0.54 is equal to a reliability of 0.7 (see also Section 4.1.1). This value guarantees a sufficient subtest reliability to be able to estimate G as a function of all four subtests at an acceptable level.
- In particular around the mean of the theta scale, the standard error of estimation may quickly reach a value below the specified value of 0.54 after a person has responded to only 4 or 5 items. However, the minimum number of items that will be administered during a test is set at 10 items for each subtest. The reason for this being that a person may give a wrong response to one of the first items while this does not reflect his/her position on the theta scale. Then, the person needs to answer enough items more to make up for this 'mistake', in order to end with an accurate estimate of theta. To prevent too long test sessions, both as regards items and time, the maximum number of items was restricted to 15 for each subtest. However, for the greater part of the theta scale, less than 15 items will be sufficient to meet the specified accuracy of measurement.

## 2.4 Scoring

Above, the procedure of theta estimation is described. However, not theta estimates but T-scores are reported to a candidate. The procedure of T-score computation is described below as well as the interpretation of these T-scores.

### 2.4.1 T-score computation

A theta value is estimated for each candidate based on the item responses, as explained in Section 2.3.4. However, the scores are reported as T-scores to the candidates. The scale of T-scores is reported into groups of 5 T-score points intervals: to 30, from 31 to 35 etcetera. The distribution of the T-scores for each norm-group has a mean value of 50 and a standard deviation of 10. The computation from estimates of theta values to T-scores is given below, for the subtests and G-factor separately. More information concerning the norm groups that are used in the computation of T-scores is given in Section 3.4.

### T-score subtest

The T-score for each of the subtests  $t$  is computed by;

$$(4) \quad T_t = \left( \left( \frac{\hat{\theta}_t - \bar{\theta}_t}{\bar{\sigma}_t} \right) * 10 \right) + 50$$

where  $\hat{\theta}_t$  denotes the estimated theta value of a candidate for subtest  $t$ ,  $\bar{\theta}_t$  represents the mean theta value in the norm group of the subtest and  $\bar{\sigma}_t$  denotes mean standard deviation in the norm group of the subtest.

### T-score G-factor

The T-score for the G-factor is obtained in two steps. If a mean T-score across subtests would be computed, the result would no longer be a T-score. After all, the standard deviation is reduced and no longer equal to 10.

To obtain an accurate distribution of T-scores on the G-factor, first a theta value for the G-factor  $\hat{\theta}_{tot}$  is computed. The subtests are considered to be equally important in determining the G-factor. The unweighted mean theta value across the four subtests could be used as an estimate for the G-factor theta value. However, theta estimates of one subtest may be estimated with more accuracy compared to theta estimates of another subtest. Therefore, the subtest theta estimates are weighted with the accuracy of their estimation in the computation of the G-factor theta value. A subtest theta value that is estimated with a small standard error will be given a higher weight compared to a subtest theta estimate with a larger standard error. This results in the computation of the G-factor theta estimate from;

$$(5) \quad \hat{\theta}_{tot} = \frac{\sum_{t=1}^4 I(\hat{\theta}_t) * \hat{\theta}_t}{\sum_{t=1}^4 I(\hat{\theta}_t)}$$

where  $\hat{\theta}_t$  is the estimated theta value for a candidate on a subtest  $t$ , and  $I(\hat{\theta}_t)$  is the amount of information for the corresponding estimated theta value on subtest  $t$ , which is obtained by;

$$(6) \quad I(\hat{\theta}_t) = \frac{1}{(SE(\hat{\theta}_t))^2}$$

where  $SE(\hat{\theta}_t)$  is equal to the standard error of the estimated theta value for this subtest  $t$ . Of course, it is important that the subtest theta estimates have a common metric, that is, they are on the same scale. This is ensured by fixing the theta scale according to a standard normal distribution.

The second step is to transform the theta value of the G-factor to a T-score on the G-factor.

$$(7) \quad T_{tot} = \left( \left( \frac{\hat{\theta}_{tot} - \bar{\theta}_{tot}}{\bar{\sigma}_{tot}} \right) * 10 \right) + 50,$$

where  $\bar{\theta}_{tot}$  denotes the mean theta value on the G-factor in the norm group and the mean standard deviation in the norm group of the theta for the G-factor is denoted by  $\bar{\sigma}_{tot}$ .

The norms for both the subtests and G-factor are given in Table 3.36.

## 2.4.2 Score interpretation

In Appendix D, a dummy report of a fictive candidate, the so-called 'Bert Smith', is shown. As can be seen in the dummy report, the scores on the subtests and the G-factor are given in so called T-scores. These are standard scores with a mean of 50 and a standard deviation of 10. The computation of T-scores was just explained in more detail in Section 2.4.1.

Within the text of the report of Connector Ability 1.1 the meaning of a T-score and the interpretation it warrants is explained in non-technical straightforward language that can be understood by the intended user who has successfully completed the certification program (see Section 2.5).

As indicated in 2.1.4, the scores on the four subtests are given to provide more detailed feedback to the candidate. However, the scores on subtest level may have been measured less reliable compared to the G-factor. Therefore, for each subtest score a bar around the score is given. The bar shows the margin around the score. In three-quarters of cases, the score will be within this margin if the candidate would take the test again.

## 2.5 Use

As explained earlier, 'user' is defined here as the individual who discusses the content of the report with the respondent. PiCompany demands certification on knowledge of the test itself and the context for its use as a condition for an allowance to use the test. This certification is based on a successful completion of a dedicated certification training. The certification training is offered in an e-learning environment.

## E-learning module

In principle, every person who has a role as a manager responsible for HR or as an HR professional in an HR process like selection, training or career guidance is eligible for such a certification training, irrespective of earlier academic qualifications. The training is also appropriate for everyone who has no psychometric background.

The e-learning module is offered online. Providing information is alternated by questions with respect to content and use in practice. The training provides many references and additional information for experienced test psychologists. At the end of the training, the user has sufficient knowledge to autonomously use Connector Ability 1.1 in his/her own setting. The e-learning module is terminated by a final test that has to be passed to obtain a certificate.

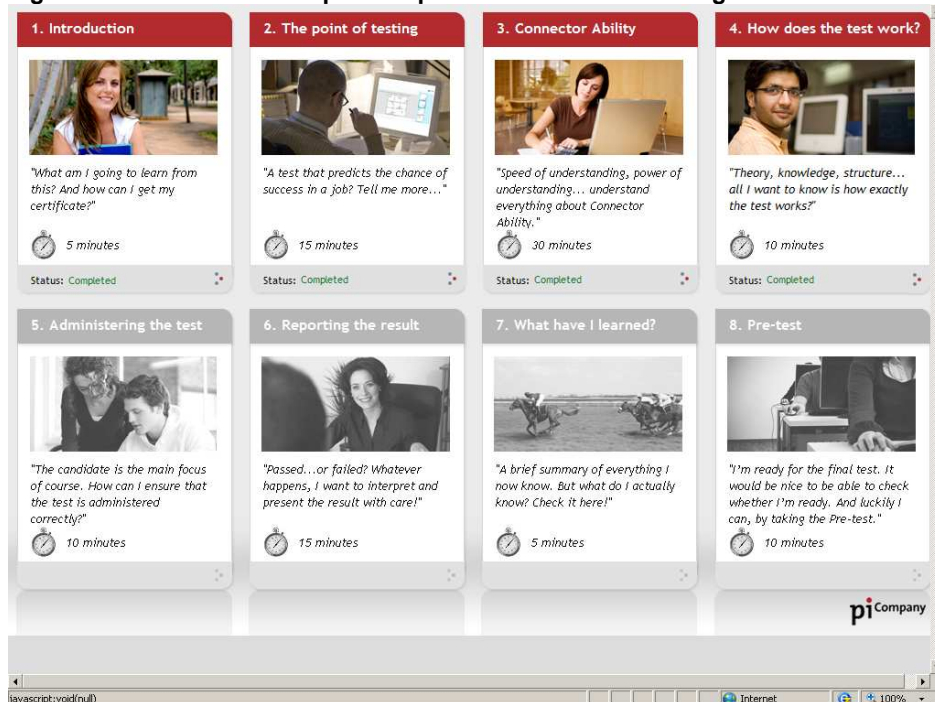
At the end of the e-learning module, the user knows the answers to the following questions:

- How do I set up a good testing procedure, and what is my role in that respect?
- When and why do I use Connector Ability 1.1?
- How does Connector Ability 1.1 work?
- How do I administer the test and how do I interpret the results?
- How do I associate with candidates in a professional and comfortable way?

## Chapters

The e-learning module comprises eight chapters. These chapters all need to be completed by the user. The user may proceed to a second chapter, only if the previous chapter has been finished. Figure 2.2 shows a screen dump with an overview of the chapters.

Figure 2.2 Screen dump of Chapter Overview in E-learning Module



A short description of the users knowledge after finishing the eight chapters;

- 1 Introduction
- 2 Point of testing  
The user knows why (s)he's setting up a test as a part of recruitment and selection procedures. Which instrument is best at predicting success in a job? What is validity, and what is 4TP? Why does an intelligence test need to be reliable, and how to measure that?
- 3 Connector Ability  
The user knows the most important features of Connector Ability 1.1. (S)he knows what intelligence and G-factor is. Knows what is meant by free of cultural bias and adaptive testing. Furthermore, it has been explained how the G-factor score is calculated and what a T-score is.
- 4 How does the test work?  
The user knows exactly how the test works. It is clear how different parts of the test are constructed and how much time there is to do each part of the test.
- 5 Administering the test  
The user knows how to administer the test with due care. It is clear that good preparation is important for both the candidate and the test administrator, as well as how you get the test ready to be started. The user knows the possible limitations for taking an intelligence test and how to deal with candidates ethically. Finally, the five rules for a good use of the test are well known.
- 6 Reporting the result  
The user knows how to conduct a feedback interview properly and how to deal with different candidates. It is clear what the user should pay attention to during a feedback interview. Finally, the user knows how to deal concisely and to the point with candidates irrespective of the result.
- 7 What have I learned?  
Provides a summary of what has been learned in the previous chapters. Users are asked if they remember all information and are provided the opportunity to go back to a previous chapter.
- 8 Pre-test  
Before the final test is administered, a pre-test is provided to the user. The pre-test is representative of the final test and can be taken as often as one likes.  
Feedback is provided concerning correct and incorrect responses.

Once all chapters have been finished, the user has to take the final test. This final test consists of 20 questions that are representative of the contents of the eight chapters. To pass the final test, the user has to respond correctly to at least 80 % of the questions. The user may take the final test twice at a maximum. When the exam has been passed, the user obtains a certificate, which is a condition for an allowance to use the test. Additional information is available to the user. There are best practice guidelines, an example of a test report, and a candidate brochure. Furthermore, a manual for the test assistant is provided and the top 10 of frequently asked questions (see Appendices).



# Chapter 3

## Construction

This chapter describes the process of test construction. First, the construction of the instructions is explained. Next, the process of item pool construction is described, including DIF analyses (see, e.g. Angoff, 1993). Group differences have been investigated and of course norms are given.

### 3.1 Instructions

The instructions of this test are constructed in such a manner that the usability and equal opportunity for all candidates to take the test to the best of their ability, is maximized. Obviously, language is a part of this test, especially with regard to the instructions. The influence of language, culture and other differences between candidates however is minimized in several ways:

- *Usability*

Much attention has been paid to the usability of the instructions, especially in the design of the screens. The screens display tranquil colours, and text and visual images are only shown when they are relevant to the actual explanation. This helps candidates to focus on what is important and to not be distracted by bright colours or non-relevant information. The screens are designed by a professional designer, emphasizing the above-mentioned aspects to maximize the usability.

- *Stepwise structure and practice opportunity*

In the instructions the emphasis lies on explaining, step by step, how the test works and on providing each candidate with the opportunity to practice before taking the actual test. Sample items are offered for this latter purpose of practicing. Each candidate is provided with the opportunity to go through the instructions (explanation and sample items) at his/her own pace. This way, each candidate has a reasonable chance of arriving at the same starting position for the subtest and therefore has a reasonable chance to subsequently recognize and solve the actual test items of the subtest to the best of his/her ability. For example, both candidates who do *not* have any test experience and candidates who *do* have test experience, are all provided with the opportunity to arrive at the same starting position before actually taking the test, by practicing and preparing for the test at their own pace. This equal opportunity also applies to both candidates who *do* and candidates who *do not* suffer from any test anxiety. A candidate whose pace of understanding is slower, for whatever reason, can take more time to go through the explanation and sample items.

Also, the (possible) impact of other differences between candidates on their level of understanding of the instructions is minimized as a result of the structure of these instructions and the practice opportunities they offer. So everyone has a fair chance of learning what is expected from him or her. This is in accordance with the demands mentioned in the LBR-NIP publications on testing with ethnic minorities (Chapter 1).

- *Candidate brochure and practice test*  
The instructions, including the sample items, are also presented in a candidate brochure and online practice test. This adds to the opportunity to be able to prepare for the test and practice beforehand (Chapter 2).
- *Language*  
The complete test, including the instructions, was made available in several languages. This provides candidates with an enhanced opportunity to understand and learn and it decreases the influence of language interfering with measuring the construct of intelligence. The candidate is allowed to choose the instructions language which suits him/her best. This is in accordance with the demands mentioned in the LBR-NIP publications on testing with ethnic minorities (Chapter 1). Currently, the instructions are available in Dutch and English. Furthermore, with regard to the technical system a flexible design is used. This makes it easy to add different languages quickly and easily.
- *Text and visual images*  
Text is combined with visual images in both the explanation and sample items. This helps candidates to understand the information and learn, even when the text might not be fully understood, for language or other reasons.

From the start, the instructions were designed and used as a combination of text and visual images. In all pilot studies in constructing the test, it was confirmed by participants that this and the other characteristics and content of the instructions contribute to their level of understanding and preparation for the test.

First, it was tested for the BA/MA-group whether the instructions were clear and adequately prepared candidates for the test. A group of 70 candidates was interviewed. The results indicated that they were very positive and valued the instructions, with the explanation and sample items, as being very clear and a good preparation for the actual test. Also, for the ME-group it was tested whether the instructions were clear and prepared candidates for the test in a sufficient way. The results for this ME-group showed that they were also very positive and valued the instructions, with the explanation and sample items, as being very clear and a good preparation for the actual test. The elaborate practice opportunity and the combination of text and visual examples were valued most highly by this specific group.

## 3.2 Item pool construction

First, the design of data collection and process of item pool construction will be described. Next, estimation procedures for item parameters are explained, and information with respect to the quality of the items is given. Furthermore, (sub)test correlations are given. Finally, the procedure of DIF analyses is described, including the results and consequences.

### 3.2.1 Design

In this section, the procedure of item pool construction is described. Different phases of data collection and data analyses are discussed. After a pre-test phase, two pilot studies were set up. These studies resulted in Connector Ability 1.0, which subsequently was applied in a selection context. Meanwhile, a practice test was constructed and put on the internet. All data collection and research activities have resulted in the release of *Connector Ability 1.1*.

#### Pre-test

For each subtest, a number of experts constructed items that met the criteria that were formulated at that point. Review by at least one other expert of many generated items resulted in approximately 200 items for each subtest. The pre-test was intended to analyse these items to quickly filter out poor functioning items.

Individuals from a heterogeneous set of working adults were asked to each select a number of an also heterogeneous set of acquaintances to participate in the pre-test. This procedure resulted in a total number of 586 participants for this pre-test. 24 % of the sample had a BA educational level and 76 % MA. The composition is diverse, students from various disciplines as well as a heterogeneous group of working adults.

Eight tests were constructed, each of them consisting of 25 items per subtest. These tests were administered online, unproctored and without any time limits. Each item was administered between 25 and 60 times.

It was reported by participants that it took them a lot of time to complete the test. Participants reported to have a hard time to solve the items. This resulted in the specification of time limits per subtest.

#### *Consequences for the item pool*

The p-values (probability of a correct response) of the administered items were inspected. For an item to be kept in the item pool, a minimum p-value of .25 was required, which is equal to the guessing probability. The maximum p-value was set at .9.

Higher p-values indicate that nearly all participants give the correct response irrespective of their position on the theta scale, which means that the item does not contribute any information. Also, incorrect alternatives were inspected for deviant items. For example, a wrong alternative that is never chosen may be altered.

The results mainly affected the subtests Series of Figures and Diagrams. For Series of Figures, the items were shown to be relatively easy. New items were written that were more difficult.

As a result of adjustment of the criteria for Diagram items, many items had to be discarded of the item pool. A selection of approximately 80 items conformed to the criteria. Some items that were discarded did show some consistent properties. The concepts in the items related to family and family relations, or items required knowledge of for example the classification of animals. New items were constructed under the more restricted conditions to administer in the first pilot study.

### **First pilot study**

The first pilot study was set up to estimate  $\alpha$  (discrimination) and  $\beta$  (difficulty) parameters that are needed for adaptive testing (see Section 2.3.4). Furthermore, data of this construction sample are used for data analysis with respect to reliability and differential item functioning.

In the first pilot study, time limits were implemented for individual items, based on the findings of the pre-test. The time limit was based on the mean response time of an item in the pre-test, plus one standard deviation.

As said earlier, data obtained in this pilot study were used to estimate item parameters. Therefore, the preliminary item pool was divided into booklets consisting of 14 items per subtest. The first booklet contained the first 14 items of a subtest. Within one booklet, the items increased in difficulty, as assessed in the pre-test. In this way, both easy and difficult items are administered in each booklet. For the second booklet, the second, fourth, sixth etc item were chosen from booklet one, after which new items were added to have a comparable range in difficulty of the items. This was repeated for each subsequent booklet.

This means each booklet has an overlap of seven items with the previous booklet and seven items overlap with the next booklet. The overlap of items across booklets is required to link the scales to have the same metric. A short overview of a comparable sampling design is given in Figure 3.1.

To obtain accurate estimates of the item parameters, 300 responses are needed for each item (Chuah, Drasgow, & Leucht, 2006). As each item is administered in two booklets, each booklet has to be administered to a minimum of 150 participants.

**Figure 3.1. Example of Sampling Design**

	Booklet items						
Booklet	1-7	8-14	15-21	22-28	...	...	133-140
1	1						
2		2					
3			3				
4				4			
...					...		
...						...	
19	19						19

Several conditions were varied to be able to study differences between groups in different conditions. The conditions are: proctored versus unproctored, differences in ethnic background (autochthon, member of a (non)-western ethnic minority group), and administration mode (online or paper-and-pencil). Within booklets the conditions were varied as much as possible to be able to compare groups.

*Sample*

The total sample consisted of 4811 participants. The vast majority of them were selected from a database from a market research agency. The sample is structured according to the above-mentioned conditions, balanced also with respect to gender and age.

**Table 3.1 Frequencies for sample of pilot study one**

Variable	Category	Frequencies	Percentage
Gender	Men	1997	42
	Women	2785	57
	Unknown	29	1
Age	< 30 years	1777	37
	30 - 45 years	1614	34
	> 45 years	1361	28
	Unknown	59	1
Educational Level	BA (HBO)	2314	48
	MA (WO)	2155	45
	Unknown	342	7
Ethnic Background	Autochthon	2688	56
	Western minority	1433	30
	Non-western minority	594	12
	Unknown	96	2

N = 4811

A sample of 3258 participants was composed, with an equal number of BA and MA participants. Both groups form a representative reflection of the population for which the test is developed (gender, age, ethnic background). These data were used for the analyses of group differences, allowing for an analysis of the sample as a whole, without differentiating in educational level.

#### *Data analysis and consequences for the item pool*

The data were used to estimate  $\alpha$  and  $\beta$  parameters, see for more information Section 2.2.1.

The procedure of estimating item parameters will be described in the next section.

The data of this construction sample were used to investigate some of the psychometric properties of the test. Groups were compared with respect to the p-values of the items. Items were removed which showed serious differences between groups of participants. Items of which one of the parameters was not estimable were removed. DIF analyses were performed wherever possible, which again resulted in the removal of some of the items (see Section 3.2.4).

The items that could be kept in the item pool after this pilot study were included in Connector Ability 1.0. The item pool of each subtest contained a number of between 110 and 114 items.

#### **Second pilot study**

In the second pilot study, the functioning of the computerized adaptive test was examined in practice. Two hundred participants received an extensive instruction for the test. Sample items were provided to exercise, which warranted that each participant had the same starting position. Before each subtest was started, it was verified whether the participant understood what was expected and knew what to do. The level of understanding of the instructions by the participants, test length in practice and adaptive item selection were evaluated.

Apart from testing the adaptive procedure in a real selection setting, data were obtained for the analyses of test-retest reliability and for validity studies, see Chapter 4.

#### **Connector Ability 1.0**

Connector Ability 1.0 is the computerized adaptive test resulting from all previous studies.

The test closely resembles the test administered in the second pilot study. Again, each participant received an extensive instruction of the test. Sample items were provided to exercise, which warranted that each participant has the same starting position. Before each subtest was started, it was verified whether the participant understood what was expected and knew what to do. The test was administered for selection purposes in different organizations.

Gathered data were used to compute norms that are based on data obtained in a setting that is equal to the setting for which Connector Ability 1.1 is intended. Furthermore, standard error of estimation in the aimed setting can be determined, and differences between groups can be examined.

### *Selection sample*

A total of 2095 candidates have been administered Connector Ability 1.0 between September 2007 and September 2008. The data were gathered at various industries as financial and insurance, transportation and storage, professional, scientific and technical industries. About 20% of all data were gathered at the Assessment Centre of PiCompany. Table 3.2 shows the characteristics of the selection sample.

**Table 3.2**      **Frequencies for selection sample**

Variable	Category	Frequencies	Percentage
Gender	Men	1305	62
	Women	737	35
	Unknown	53	3
Age	< 30 years	1603	77
	30 - 45 years	335	16
	> 45 years	152	7
	Unknown	5	0
Educational Level	MA (WO)	822	39
	BA (HBO)	679	32
	ME	159	8
	Other	433	21
	Unknown	2	0
Ethnic Background	Autochthon	1372	66
	Western minority	197	9
	Non-western minority	497	24
	Unknown	29	1

N = 2095

### **Practice test**

A practice test was constructed for people who want to practice in advance of an assessment procedure as well as for people who would like to take a test measuring intelligence. The test is available through the internet. Organizations that administered Connector Ability 1.0 give the advice to their candidates to visit the website of PiCompany to take the practice test as a preparation for their assessment. Also via other channels, possible participants were made aware of this possibility. Participants who completed the practice test got a short report in return including their G-factor score in one of five categories.

The practice test consists of 14 items for each subtest. A total of 21 responses, seven items for each subtest, were used to compute a reliable estimate of the G-factor. Apart from the seven items in each subtest to compute the G-factor, also a set of seven experimental items were administered. The 'G-factor items' are the same for all participants and function as an anchor. The experimental items were administered in booklets. That is, a fixed set of seven items for each subtest was administered to be able to compute  $\alpha$  and  $\beta$  parameters.

When a booklet had been administered at least 350 times it was replaced by a new booklet of items. These 350 participants have all indicated to have taken the test in a concentrated manner, and to have understood the purpose of the test. The minimum sample size of 350 participants is larger than the required 300. Participants were allowed to take the test as many times as they wanted. Based on name and date of birth, participants who made use of this possibility and took the test more than once were identified. Only the first test administrations were used for further analyses, which included the responses of a minimum of 300 participants for one booklet.

In a period of approximately six months, a total of 13 booklets were administered, which resulted in the administration of 91 experimental items for each subtest. These data were used to compute  $\alpha$  and  $\beta$  parameters of the experimental items, that are all on the same scale as the item parameters already estimated in the first pilot study. The procedure of estimating item parameters will be described in the next section.

The practice test, like the structure and instructions, is identical to the selection test, which means that only the items differ across the two tests. Of course, the participants selected themselves to do the test, and the setting was unproctored.

#### *Sample completing practice test*

Because people selected themselves to participate, there were various reasons and motives for participating. The objective of this test is to prepare candidates for their assessment. However, experimental items were included as well.

To estimate reliable  $\alpha$  and  $\beta$  parameters, concentration and knowledge of the purpose were evaluated at the end of the test. Only those participants were included in the sample that indicated to have worked in a concentrated manner and understood the purpose of the test. As these questions were asked at the end of the test, all of these respondents had finished all subtests. As described above, participants that participated multiple times were deleted, except for their first test administration.

At the start of the test participants were asked to provide some background data. It was stressed that this information will only be used for research and it will be treated anonymously. It was not obligatory to provide the data, though this was not specifically mentioned. The result was that some background variables show a lot of missing values. Table 3.3 shows the frequencies for the sample that completed the practice test.

The variable educational level was measured by asking the highest completed educational level. A relatively large number of respondents did not report ME, BA or MA. The majority of these respondents indicated to have obtained a Secondary Educational level. It is likely that many of them are students that are doing a BA or MA education, but have not obtained their degree yet.



**Table 3.3**      **Frequencies for sample that completed the practice test**

Variable	Category	Frequencies	Percentage
Gender	Men	299	3
	Women	323	4
	Unknown	8220	93
Age	< 30 years	4841	55
	30 - 45 years	2363	27
	> 45 years	966	11
	Unknown	672	8
Educational Level	ME	1091	12
	BA	2506	28
	MA	1887	21
	Other	2435	29
	Unknown	923	10
Ethnic Background	Autochthon	5502	62
	Western minority	1990	23
	Non-western minority	991	11
	Unknown	359	4

N = 8842

*Data analysis and consequences for the item pool*

The 91 experimental items administered in the practice test were *only* included in the final item pool if they met the following criteria;

- The item parameters are estimable and the standard error of the parameter estimate is not too high (i.e. it is a reliable estimate).
- All response categories have been selected one or more times during its administration.
- The item has a difficulty parameter above zero, irrespective of the value of the discrimination parameter.

OR

The item has a difficulty parameter below zero and a discrimination parameter above one. Items with low difficulty parameters as well as low discriminative power will hardly be selected, as there are a number of items that provide more information for the estimation of theta, and therefore will be selected. (Note, that there are items with lower discrimination parameter that are included after the pilot studies. These items will remain in the item pool until more items are included with higher discrimination parameters).

- In the subtests Series of Figures and Matrices of the practice test, new as well as clone items were included. Clone items are items that are basically identical to items administered in the first pilot study, but where the type of figure used is altered. For example, circles are replaced by triangles.

Candidates should not encounter two nearly identical items in one test administration, which is possible when clone items are included in an adaptive test. At this point, there are no content restrictions; therefore only one of the clone items may be included in the final item pool. Evidently, the best item is chosen, based on the item parameters, i.e. item with the highest discrimination parameter.

Below, the consequences and results for the final item pool are provided, for each subtest separately.

#### *Series of Numbers*

From the items that were gathered until pilot study two, 99 items could remain in the final item pool. These items vary in difficulty parameters, and have discrimination parameters below as well as above one. The practice test resulted in a set of 64 items that could be added to the item pool according to the above-described criteria. This results in a total of 163 items.

#### *Matrices*

A total of 94 items could remain in the final item pool from all items calibrated in the first pilot study. The practice test resulted in a set of 67 items that conformed to the criteria. A number of clone items were removed, resulting in an item pool containing 139 items.

#### *Series of Figures*

In the first pilot study, also items of a different type were included. Experience has taught, that by knowing some tricks these items were relatively easy solvable. The estimated high discriminative power of these items is therefore not sustainable, as it is not directly related to intelligence (knowing tricks) and is not what is intended to be measured by the test. Therefore, these items are eliminated from the preliminary item pool, resulting in an item pool of 76 items. The practice test resulted in another set of 57 items that could be included in the item pool. After removal of clone items, the final item pool includes 117 items

#### *Diagrams*

As the item criteria were restricted to the conditions described in Section 2.2, a large number of items from the preliminary item pool had to be removed. 47 items could remain in the final item pool. The practice test contributed 48 items to the item pool, resulting in a total of 97 items.

More information about this final set of items, like psychometric properties, is given in Section 3.2.3 and Section 3.2.4.

### **3.2.2 Parameter estimation**

The model underlying the computerized adaptive procedure is the two-parameter logistic model (see Section 2.3.4). This model contains item discrimination ( $\alpha$ ) as well as item difficulty ( $\beta$ ) parameters. Data to calibrate the item parameters have been obtained in two phases.

In the first pilot study, data were gathered in subsequent booklets. These data were analyzed in one go, by an adapted algorithm which takes into account the incomplete nature of the data (Glas, Twente University, internal report).

Next, for each subtest seven items were selected that were used in the practice test to compute a score on G-factor level as feedback to the participant. At the same time, this item set functions as an anchor for new booklets of experimental items that were administered in this practice test. The program Multilog (Thissen, Chen, & Bock, 2002) was used to estimate the item parameter for the experimental items, while the item parameters of the anchor items were fixed to their specific values estimated previously. This warrants that all item parameter estimates are on the same scale.

To estimate the item parameters, the responses were used of participants who responded to at least seven items for one of the subtests. Furthermore, only those participants were included who indicated to have worked in a concentrated manner during test administration. Information about the samples has been given above.

### 3.2.3 Item parameters

The characteristics of the estimated item parameters are given below. These items are included in the item pool of Connector Ability 1.1.

#### Discrimination parameters

The mean, standard deviation, minimum and maximum value of the discrimination parameters are given for each subtest separately in Table 3.4.

**Table 3.4 Descriptive statistics of discrimination parameters**

Subtest	n	Mean	SD	Minimum	Maximum	# items $\alpha > 1$
Series of Figures	117	1.44	0.72	0.25	3.87	89
Matrices	139	1.26	0.62	0.17	3.09	87
Series of Numbers	163	1.55	0.80	0.25	4.00	123
Diagrams	95	1.47	0.70	0.27	4.00	68

n = number of items

Items with higher discrimination parameters are preferred as they are more informative and thus increase the accuracy of the estimation of theta. Items with high discriminative power are selected more frequently, because item selection is based on the information function which is a function of the item parameters (see also Section 2.3.4). In the last column of Table 3.4, for each subtest the number of items with a discrimination parameter above 1 is given.

### Difficulty parameters

Table 3.5 shows the descriptive statistics of the difficulty parameters of the four subtests.

**Table 3.5** Descriptive statistics of difficulty parameters

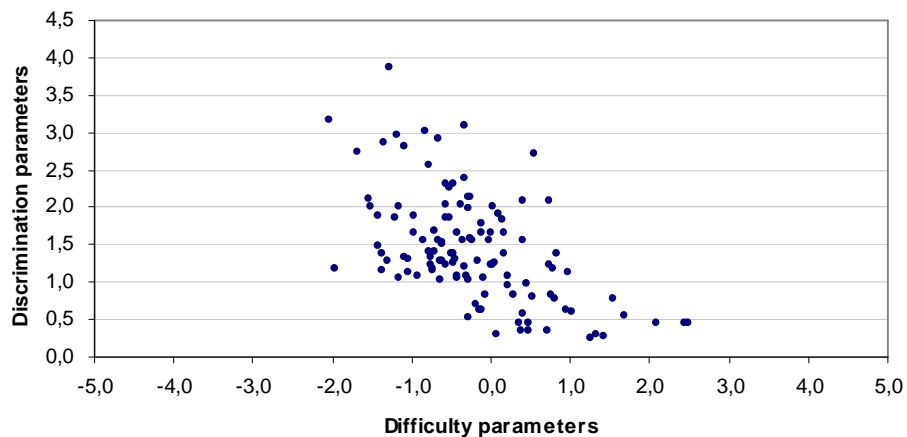
Subtest	n	Mean	SD	Minimum	Maximum
Series of Figures	117	-0.22	0.86	-2.03	2.49
Matrices	139	-0.14	0.95	-2.22	2.75
Series of Numbers	163	-0.67	1.31	-4.10	3.28
Diagrams	95	-0.58	0.93	-3.14	2.42

n = number of items

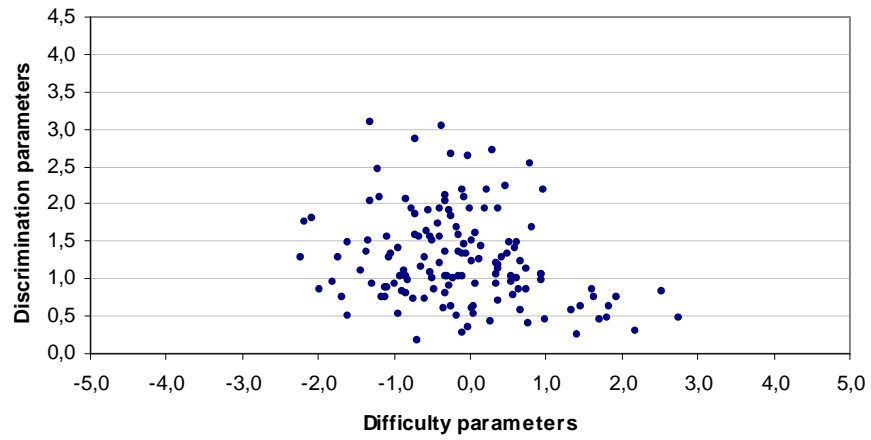
The theoretical minimum and maximum value of the difficulty parameters is between minus and plus infinity. The mean value is below zero, which indicated that there are more items with a lower difficulty parameters compared to higher values.

An important characteristic in IRT is that the difficulty parameters and theta are on a common scale. Connector Ability 1.1 will be used primarily in a selection setting. This means that a reliable estimate is particularly important around the cut-off score of theta. The cut-off score often lies between theta values of -1 and + 0.5. Therefore, it is important that in particular for this range a reliable estimate of theta can be given. This requires a sufficient number of items with a difficulty parameter between -1 and + 0.5, with preferably high discrimination parameters. This requirement is met, as will be explained below. The next four graphs depict on the X-axis the difficulty parameters, and on the Y-axis the discrimination parameter of the items for each of the subtests

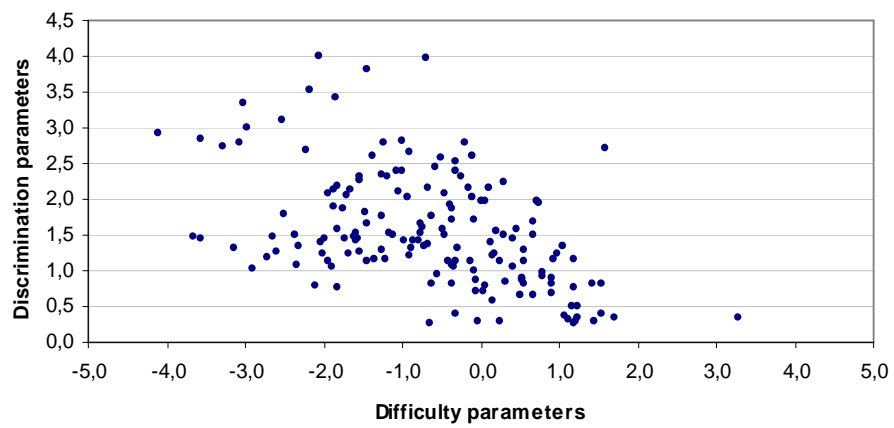
**Figure 3.2**  
Item parameters Series of Figures



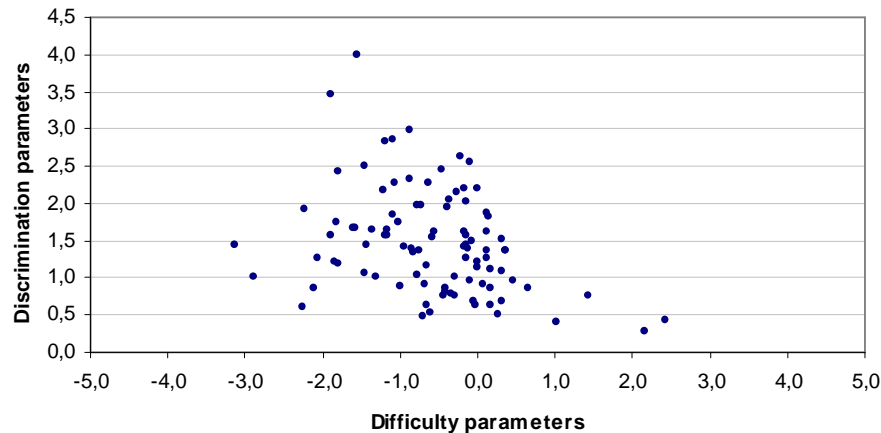
**Figure 3.3**  
Item parameters Matrices



**Figure 3.4**  
Item parameters Series of Numbers



**Figure 3.5**  
Item parameters Diagrams



It is seen that there are more easy compared to difficult items. The majority of the items have a difficulty parameter value in the range -2 through + 1. This means that for the range were accurate estimation is particularly necessary, a large number of items are available with sufficient discriminative power.

### 3.2.4 Item information

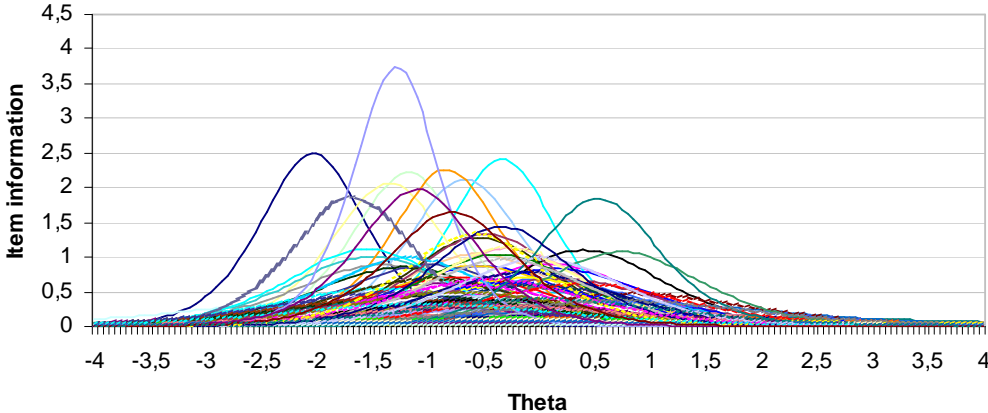
The quality of the items can be assessed by inspecting the information functions of the individual items. Information functions provide an overview of the information that an item contributes given the value of theta, see also Section 2.3.4. Thus, the item information depends on the value of theta.

The item information functions are used in the selection of items for a candidate. The item providing the largest contribution to obtain a reliable estimate of theta, given the theta value that is estimated for the candidate at a particular point during test administration, will be selected. The test information is equal to the sum of the item information given the value of theta. This means that the information value of one item may be relatively low, while with all items in a test combined the test information may be high.

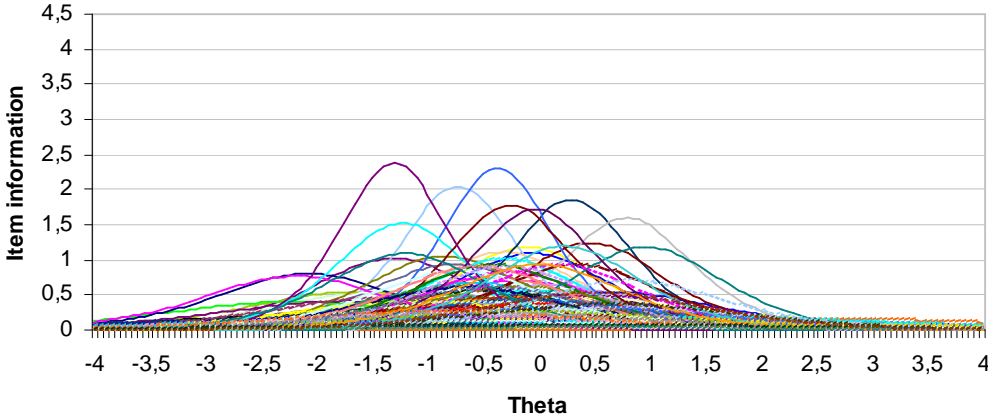
Because Connector Ability is an adaptive test and each candidate may respond to a different set of items, the test information differs among candidates. Furthermore, as for each candidate the items are selected that contribute most to the estimation of theta, the test information will be higher compared to administration of a fixed set of items. Higher test information is related to more reliable theta estimates, as will be discussed in Section 4.1.1.

Figure 3.6 through Figure 3.9 show for each subtest the information functions of all items of the subtest. It is seen that the items provide most information in the range of theta between -2 and + 1. As described earlier, this is the range where a reliable estimate of theta is needed, as this is the range where cut-off scores are set.

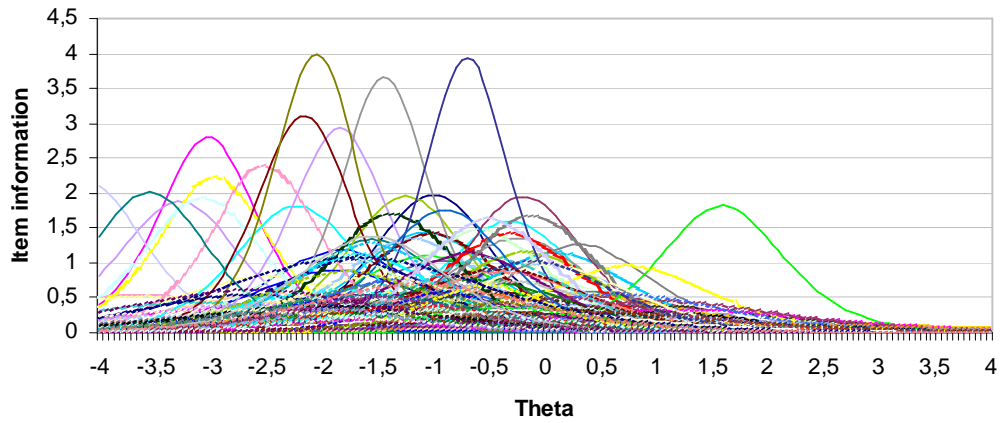
**Figure 3.6**  
Item informations curves Series of Figures



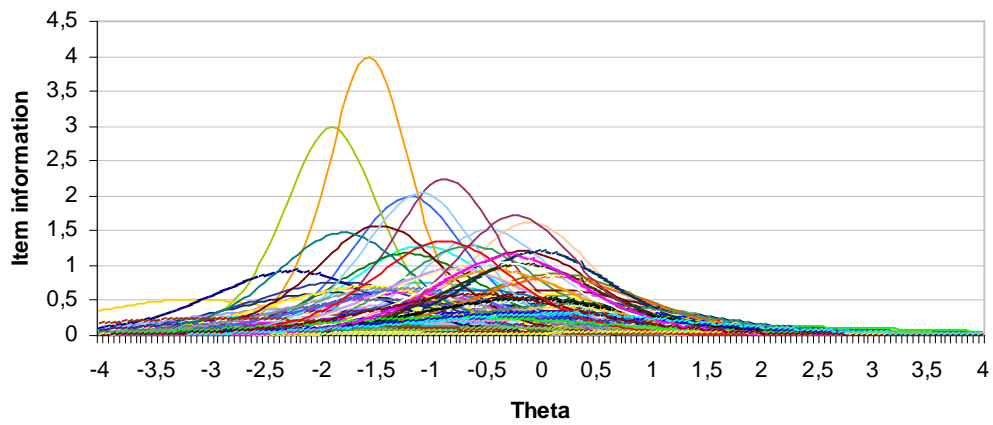
**Figure 3.7**  
Item informations curves Matrices



**Figure 3.8**  
Item informations curves Series of Numbers



**Figure 3.9**  
Item informations curves Diagrams





### 3.2.5 DIF analyses

An item is said to exhibit differential item functioning (DIF) when it has different response probabilities for different groups, after matching the groups with respect to their position on the theta scale (Angoff, 1993). DIF detection methods compare the functioning of an item across manifest groups.

An important characteristic of DIF detection methods that are based on IRT, is that the differential functioning of the item is inspected while conditioning on theta. The underlying distribution of theta does not need to be equal across the two groups, that is, groups may differ in their ability level.

#### DIF detection procedure

In Section 3.2.1, the procedure of data collection is described. In the first pilot study, booklets of items were administered for each subtest. Each booklet was administered a sufficient number of times to be able to accurately estimate the item parameters. The data collection was designed in such a way that some variables were varied to allow the study of differential item functioning. After data collection, 16 sets of seven items could be analyzed on the presence of DIF with respect to gender. DIF with respect to ethnic background as well as with respect to age could be studied for eight sets of seven items.

An IRT-based DIF detection method was chosen. Each subtest and each set of items was analyzed separately for differences in difficulty parameters across groups. The procedure will be described for one subtest and one set of seven items that is studied for the presence of DIF with respect to gender. The variable gender is of course dichotomous.

The DIF detection procedure will be explained stepwise. The models were all fitted with the program Multilog (Thissen, Chen, & Bock, 2002):

- 1 Model 1  
Model fitted with item parameters constrained to be equal across groups.
- 2 Model 2  
Model fitted where difficulty parameters of 1 item are allowed to vary across groups.
- 3 Fit of Model 1 - Model 2  
Difference in model fit (log-likelihood values;  $\Delta \log-L$ ) is chi-square distributed, with degrees of freedom (*df*) equal to the difference in the number of parameters estimated in each model.  
The critical chi-square value can be obtained for a chosen level of significance. To correct for multiple comparisons, a Bonferroni correction can be imposed. The correction involves the division of the alpha level of 0.05 by the number of items that are studied for DIF in a given set of items. This level of significance is used to set the critical chi-square value.

- 4 Does the item exhibit DIF?
  - $\Delta \log\text{-L} (\Delta df) < \chi^2$   
No DIF in the item.
  - $\Delta \log\text{-L} (\Delta df) > \chi^2$   
DIF item; difficulty parameters vary significantly across groups.
- 5 Further analyses needed?
  - No or one item identified as displaying DIF  
→ Stop DIF detection for this set of items.
  - Two or more items identified as displaying DIF for a given set of items  
→ Iterative DIF detection;
    - o Item with largest  $\Delta \log\text{-L} (\Delta df)$  value identified as first DIF item.
    - o Model 3 is the previous found Model 2 where the difficulty parameters of a second possible DIF item are also allowed to vary across groups.
    - o Fit of Model 2 - Model 3;  $\Delta \log\text{-L} (\Delta df)$ .
    - o Repeat step 4 and 5 until no more item can be identified as displaying DIF.

Once the DIF detection procedures are finished, a number of items are identified as displaying DIF. The items are inspected in more detail to study whether specific item characteristics are associated with the presence of DIF. All items that were shown to exhibit DIF are removed from the item pool.

### Results of DIF analyses

First, the results for DIF detection with respect to gender will be discussed, followed by DIF detection with respect to ethnic background and age. For each of the manifest variables and each subtest, no specific item characteristics could be identified that were associated with the presence of DIF. All items that were identified as displaying DIF were removed from the item pool.

Note that not all items could be inspected for the presence of DIF with respect to any or all of the manifest variables. The items that were calibrated based on data from the practice test suffered from background data that show a good balance of the groups involved. Groups were generally not large enough to be able to assess DIF with respect to gender, ethnicity and/or age. Data from the first pilot study did not have large enough groups to be able to compare them across all booklets.

### Overall results of DIF detection

Some items were detected to display DIF with respect to more than one manifest variable. Therefore, first the overall results are given in Table 3.6. The number of items that are studied is relatively low for the subtests Series of Figures and Diagrams. Many items of these subtests were removed based on adjustments of the criteria which items had to meet.

**Table 3.6**      **Frequencies and percentages of identified DIF items for each subtest**

Subtest	# studied items	# DIF items	Percentage
Series of Figures	85	5	5.9
Matrices	112	8	7.1
Series of Numbers	113	7	6.2
Diagrams	61	8	13.1
Total	371	28	7.5

Below, the results of DIF detection with respect to gender, ethnic background and age are given separately.

**DIF detection with respect to gender**

All items that remained in the item pool after the first pilot study, have been tested for the presence of DIF with respect to the variable gender, see Table 3.7. From the 371 items that have been tested, 20 items were found to exhibit DIF. This is 5.4 % of the items.

**Table 3.7**      **Frequencies and percentages of items identified as displaying DIF with respect to gender**

Subtest	# studied items	# DIF items	Percentage
Series of Figures	85	4	4.7
Matrices	112	5	4.5
Series of Numbers	113	4	3.5
Diagrams	61	7	11.5
Total	371	20	5.4

The subtests Series of Numbers, Matrices and Series of Figures contained 4 or 5 DIF items. There was no explanation found for the items to exhibit DIF.

The subtest Diagrams shows the largest number of DIF items, though the number of items that are investigated is the smallest of all subtests. Four of the seven DIF items concern items containing words that are associated with clothing. Three of those items are more difficult for men compared to women. However, other clothing items do not exhibit DIF with respect to gender. Nevertheless, it is important to keep this in mind during the construction and analysis of new items.

**DIF detection with respect to ethnic groups**

A total of 185 items for the four subtests have been investigated for the presence of DIF with respect to ethnic background. For the majority of those items, the responses of three groups could be compared; autochthon, western minority and non-western minority groups. For smaller set of items, the western and non-western minority groups had to be combined to be able to study the differences between autochthon and minority groups.

**Table 3.8**      **Frequencies and percentages of items identified as displaying DIF with respect to ethnic background**

Subtest	# studied items	# DIF items	Percentage
Series of Figures	42	1	2.4
Matrices	56	2	3.6
Series of Numbers	56	3	5.4
Diagrams	31	0	0
<b>Total</b>	<b>185</b>	<b>6</b>	<b>3.2</b>

In Table 3.8 the frequencies and percentages of items identified as displaying DIF with respect to ethnic background are shown. Just six items were identified as displaying DIF with respect to ethnic background, which is 3 % of the studied items. The six DIF items had no specific characteristics in common to explain the difference in difficulty for these items. Therefore, it was concluded that the DIF does not seem to be related to specific characteristics. It is seen that for the subtest Diagrams no items were identified to exhibit DIF with respect to ethnic background.

**DIF detection with respect to age**

DIF detection with respect to age focuses on the study of differences in difficulty parameters across three age groups; under 30 years old, between 30 and 45 years old, and older than 45 years. A total of 177 items have been studied for the presence of DIF with respect to age. There were 8 items identified as displaying DIF, which is 4.5 % of the studied items, see also Table 3.9.

**Table 3.9**      **Frequencies and percentages of items identified as displaying DIF with respect to age**

Subtest	# studied items	# DIF items	Percentage
Series of Figures	38	2	5.3
Matrices	56	2	3.6
Series of Numbers	56	1	1.8
Diagrams	27	3	11.1
<b>Total</b>	<b>177</b>	<b>8</b>	<b>4.5</b>

As for the other studied manifest variables, no specific characteristic could be found to explain the differences in difficulty for these items. The items were removed from the item pool.

### 3.3 Group differences

Some descriptive statistics of the theta estimates for the different subtests as well as for the G-factor will be given. Data of both the construction and selection sample are studied. It will be examined whether differences between groups based on gender, ethnic background and age matter.

To study whether the differences between groups are meaningful, effect sizes were computed. As groups may be quite large, differences between means are easily found to be statistically significant. Effect sizes help to determine whether the observed differences are differences that matter. Cohen's  $d$  (Cohen, 1988) is used to compute the effect size. Equation 10 describes the computation of Cohen's  $d$ ,

$$(10) \quad d = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}}$$

The mean values (of theta) for group 1 and 2 are denoted by  $\mu_1$  and  $\mu_2$ . Their standard deviations are denoted by  $\sigma_1$  and  $\sigma_2$  for group 1 and 2 respectively. Cohen (1988) defined an effect size of 0.2 as small. An effect size of 0.5 and 0.8 were considered medium and large.

First, results from the construction sample will be discussed, followed by the results of the selection sample.

#### 3.3.1 Group differences construction sample

The minimum number of answers required to estimate a theta value for a participant for one subtest was restricted to seven. To obtain a theta value for the G-factor, it is required that a participant has obtained a theta value for each subtest. Therefore, the construction sample is reduced to a sample of 3258 respondents who have answered at least seven items of each subtest. The characteristics of the sample reflect the composition of the total construction sample as given in Table 3.1.

In Table 3.10, the mean and standard deviation of theta are given for the BA and MA sample respectively. It is seen that the differences between the mean theta values of the BA and MA sample is approximately 0.5 SD.

**Table 3.10 Descriptive statistics of the theta values on the subtests and G-factor for the BA and MA sample**

Subtest	BA		MA	
	Mean	SD	Mean	SD
Series of Figures	0.087	0.962	0.545	1.176
Matrices	-0.190	0.675	0.086	0.742
Series of Numbers	-0.292	0.748	0.020	0.978
Diagrams	-0.627	1.081	-0.024	1.185
G-factor	-0.277	0.423	-0.021	0.491

N = 1669 (BA), N = 1589 (MA)

The composition of the construction sample was balanced as much as possible. Gender, ethnic background, and age are balanced across the BA and MA sample. Therefore, the differences between the groups are studied for the total construction sample.

The theta values of each subtest were computed with a maximum of 14 item responses. As the items were administered as experimental items, their quality and characteristics were not known at the time of administration. As a consequence, the theta values (estimated afterwards) may not always have been measured with sufficient reliability. Therefore, the differences between so-called plausible values are used to compute the effect sizes. A plausible value is a random draw from the estimated distribution of theta for a person; the posterior distribution (Mislevy, 1991). For computation of effect sizes regarding the G-factor, the estimated theta values are used. These theta values are estimated reliably, therefore it is not required to use the plausible values.

### Gender

In Table 3.11 the mean and standard deviation of the plausible values of the subtests and the theta value of the G-factor are given for men and women separately.

**Table 3.11 Descriptive statistics of the plausible values on the subtests and theta value of the G-factor for men and women**

Subtest	Men		Women	
	Mean	SD	Mean	SD
Series of Figures	0.350	1.29	0.261	1.29
Matrices	-0.020	0.92	-0.096	0.80
Series of Numbers	0.011	1.26	-0.305	0.93
Diagrams	-0.314	1.35	-0.389	1.41
G-factor	-0.089	0.51	-0.194	0.44

N = 1307 (Men), N = 1947 (Women)

The effect sizes for the subtest differences are computed using the plausible values. For the subtest Series of Figures  $d = -0.098$ , which is a very small effect size. The effect sizes for Matrices and Diagrams are small as well,  $d = -0.125$  and  $-0.076$  respectively. Series of numbers resulted in  $d = -0.402$ . This effect is somewhat larger, but still under a medium effect size. For the G-factor the effect size was computed based on the theta estimates, and resulted in  $d = -0.309$ . This is a small difference between men and women.

### Ethnic background

In Table 3.12, the mean and standard deviations of plausible values of the subtests and theta value of the G-factor are provided for the different ethnic groups.

**Table 3.12 Descriptive statistics of the plausible values on the subtests and theta value of the G-factor for the different ethnic groups**

Subtest	Autochthon		Western minority		Non-western minority	
	Mean	SD	Mean	SD	Mean	SD
Series of Figures	0.476	1.42	0.085	0.99	-0.028	1.20
Matrices	-0.041	0.88	-0.089	0.80	-0.142	0.88
Series of Numbers	-0.115	1.14	-0.240	0.93	-0.284	1.22
Diagrams	-0.192	1.40	-0.596	1.39	-0.537	1.23
G-factor	-0.090	0.50	-0.226	0.41	-0.267	0.43

N = 1886 (Autochthon), N = 958 (Western minority), N = 385 (Non-western minority)

It is seen in Table 3.12 that the differences in means for the subtests Matrices and Series of Numbers is relatively small. For the subtests Series of Figures and Diagrams the group of autochthon respondents have higher mean plausible values compared to both minority groups. As a result, also the G-factor shows some differences.

To determine which differences in means truly matter, effect sizes are computed as described above. The values of Cohen's  $d$  are given in Table 3.13 for the comparison of different ethnic groups.

**Table 3.13 Values of Cohen's  $d$  for the differences between different ethnic groups**

Subtest	Autochthon vs.	Western minority vs.	Autochthon vs.
	Western minority	Non-western minority	Non-western minority
Series of Figures	0.453	0.145	0.543
Matrices	0.081	0.090	0.163
Series of Numbers	0.170	0.058	0.203
Diagrams	0.409	-0.063	0.369
G-factor	0.419	0.135	0.534

N = 1886 (Autochthon), N = 958 (Western minority), N = 385 (Non-western minority)

The results in Table 3.13 show that in particular for the subtests Matrices and Series of Numbers the differences between all groups are very small. For Series of Figures, the differences between autochthon and non-western minority respondents have a medium effect size. As a consequence, these differences are of a medium size for the G-factor as well. The differences between autochthon and western minority respondents show an effect size of 0.45. The subtest Diagrams shows effect sizes for autochthon versus minority groups of 0.41 and 0.37, which is below a medium effect. For the construction sample, the estimation of theta values is based on all items available. This includes items that may have been found to exhibit DIF, or that are removed from the final item pool as a consequence of restricted item criteria. In particular, items have been removed from the subtests Series of Figures and Diagrams, which are exactly the subtest showing the weakest results. It will be investigated in a selection context whether these effects remain.

### Age

In Table 3.14 the mean and standard deviation of the plausible values of the subtests and theta value of the G-factor for different age groups are provided. In general, the oldest group of respondents shows relatively lower scores compared to the other two age groups. Again, Cohen's *d* is computed to determine the extent to which these differences matter. The effect sizes for the differences between the three age groups are provided in Table 3.15.

**Table 3.14 Descriptive statistics of the plausible values on the subtests and theta value of the G-factor for different age groups**

Subtest	Age < 30		Age 30-45		Age > 45	
	Mean	SD	Mean	SD	Mean	SD
Series of Figures	0.487	1.48	0.331	1.21	0.001	1.02
Matrices	0.029	0.91	-0.034	0.84	-0.246	0.74
Series of Numbers	-0.102	1.21	-0.180	0.99	-0.284	1.00
Diagrams	-0.267	1.36	-0.279	1.39	-0.575	1.41
G-factor	-0.081	0.50	-0.129	-0.46	-0.282	0.43

N = 1273 (< 30), N = 1082 (30-45), N = 886 (45<)

**Table 3.15 Values of Cohen's *d* for the differences between age groups**

Subtest	<30 vs. 30-45	30-45 vs. 45<	<30 vs. 45<
Series of Figures	0.163	0.417	0.541
Matrices	0.101	0.379	0.468
Series of Numbers	0.099	0.148	0.232
Diagrams	0.012	0.300	0.315
G-factor	0.144	0.484	0.611

N = 1273 (< 30), N = 1082 (30-45), N = 886 (45<)



The differences between the youngest and middle age group are very small. One effect size can be classified as a medium effect, which is the comparison of the youngest versus the oldest age group for the subtest Series of Figures. The other effect sizes are between small and medium.

### Administrative conditions

As described above, administrative conditions were varied where possible, to be able to study their influence on test scores. First, proctored versus unproctored administration will be discussed. Next, the online administrative condition is compared with a paper-and-pencil version of the test. Finally, the influence of concentration and understanding of the purpose of the test will be discussed.

**Table 3.16 Descriptive statistics of the plausible values on the subtests and theta value of the G-factor for proctored and unproctored setting as well as the corresponding effect sizes**

Subtest	Proctored		Unproctored		Effect size
	Mean	SD	Mean	SD	
Series of Figures	1.182	2.17	0.201	1.12	0.804
Matrices	0.373	1.02	-0.112	0.82	0.739
Series of Numbers	0.218	1.73	-0.214	1.00	0.431
Diagrams	0.598	1.24	-0.457	1.36	1.143
G-factor	0.293	0.56	-0.189	0.45	1.350

N = 249 (Proctored), N = 3009 (Unproctored)

Table 3.16 shows that participants in a proctored setting obtain higher plausible values compared to participants who have taken the test in an unproctored setting. The effect sizes indicate large differences.

A relatively small part of the sample has taken the test under supervision and controlled conditions. The persons who participated in this proctored setting have taken the test either online or in a paper version. Participants in the unproctored setting have all taken the online version of the test. Another important difference in characteristics of the two samples is that the participants in the proctored setting often reported to have a MA educational level. Evidently, MA participants tend to score higher compared to BA participants.

Next, the results for the comparison of paper-and-pencil to online administration are shown in Table 3.17. Again, the effect sizes are generally large, except for the subtest Series of Numbers that shows a medium effect size. The paper versions were all answered by participants with a MA educational level (students), and were all administered under proctored conditions. The online sample, though, containing a much more varied educational background, may be expected for that reason alone to show lower theta values.

**Table 3.17 Descriptive statistics of the plausible values on the subtests and theta value of the G-factor for paper and online administration as well as the effect sizes**

Subtest	Paper		Online		Effect size
	Mean	SD	Mean	SD	
Series of Figures	1.258	2.41	0.229	1.14	0.773
Matrices	0.369	1.08	-0.097	0.83	0.684
Series of Numbers	0.283	1.68	-0.205	1.04	0.494
Diagrams	0.670	1.32	-0.429	1.36	1.159
G-factor	0.295	0.53	-0.178	0.46	1.344

N = 175 (Paper), N = 3083 (Online)

At the end of the test, participants were asked whether they had worked in a concentrated manner during test administration. It may be expected that participants who indicated not to have worked in a concentrated manner will obtain lower scores compared to participants who did. Also, it was asked whether participants understood the purpose of the test and knew what to do. As for concentration, it is expected that participants who understand the purpose of the test obtain higher scores compared to participants who did not understand the purpose. Table 3.18 and 3.19 show the results for concentration and knowledge of purpose respectively.

**Table 3.18 Descriptive statistics of plausible values on subtests and theta values of the G-factor with respect to concentration as well as effect sizes**

Subtest	Not concentrated		Concentrated		Effect size
	Mean	SD	Mean	SD	
Series of Figures	-0.266	0.86	0.327	1.17	-0.817
Matrices	-0.461	0.65	-0.026	0.84	-0.820
Series of Numbers	-0.554	0.82	-0.115	1.06	-0.659
Diagrams	-0.822	1.52	-0.329	1.36	-0.484
G-factor	-0.474	0.38	-0.141	0.45	-1.132

N = 302 (Not concentrated), N = 2718 (Concentrated)

Table 3.18 and 3.19 show that working in a concentrated manner and knowing the purpose is associated with higher scores on both the subtests and G-factor. The effect sizes are medium to large. These results have supported the decision to only include data of those participants who have indicated to have worked in a concentrated manner and who understood the purpose of the test. Data from participants who did not meet these criteria were removed before further analyses were performed.

**Table 3.19 Descriptive statistics of plausible values on subtests and theta values of the G-factor with respect to purpose knowledge as well as effect sizes**

Subtest	Purpose unknown		Purpose known		Effect size
	Mean	SD	Mean	SD	
Series of Figures	-0.096	1.13	0.295	1.16	-0.484
Matrices	-0.312	0.68	-0.050	0.84	-0.483
Series of Numbers	-0.360	0.82	-0.143	1.06	-0.325
Diagrams	-0.931	1.65	-0.333	1.35	-0.562
G-factor	-0.3+2	0.42	-0.157	0.46	-0.757

N = 225 (Purpose unknown), N = 2796 (Purpose known)

### 3.3.2 Group differences selection sample

In Table 3.20, the mean and standard deviations of the theta estimates for the different subtests as well as for the G-factor are given, for the BA and MA sample respectively. These results are based on data of Connector Ability 1.0 sample in selection context. The characteristics of the sample are shown in Table 3.2.

**Table 3.20 Descriptive statistics of the theta values on the subtests and G-factor for the BA and MA sample**

Subtest	BA		MA	
	Mean	SD	Mean	SD
Series of Figures	0.292	0.751	0.724	0.998
Matrices	0.256	0.637	0.532	0.717
Series of Numbers	0.078	0.928	0.638	1.187
Diagrams	0.121	1.113	0.713	1.038
G-factor	0.010	0.395	0.311	0.461

N = 679 (BA), N = 822 (MA)

It is seen that the differences between the mean theta values of the BA and MA sample are approximately 0.5 SD.

The distribution of BA and MA educational level across samples of men and women, different ethnic groups as well as age groups, is not in all cases equivalent. As shown above, candidates with a MA educational level can be expected to have higher theta values compared to candidates with a BA educational level. Therefore, the comparison of gender, ethnic and age groups will be made for the BA and MA sample separately. The theta values are based on administration in a selection context. These estimates are sufficiently reliable on a subtest level, to compare the theta values (see also Section 4.1.1).

Below, the mean and standard deviations of theta values for various groups are given. Differences between gender, ethnicity and age groups are investigated computing effect sizes to determine its meaningfulness.

**Gender**

Table 3.21 and 3.22 show the mean theta values as well as the standard deviation for men and women, for the BA and MA sample respectively.

**Table 3.21 Descriptive statistics for the BA sample of the theta values on the subtests and G-factor for men and women**

Subtest	Men		Women	
	Mean	SD	Mean	SD
Series of Figures	0.293	0.73	0.255	0.75
Matrices	0.266	0.65	0.222	0.62
Series of Numbers	0.116	0.97	0.005	0.87
Diagrams	0.082	1.21	0.166	0.92
G-factor	0.002	0.39	0.007	0.41

N = 442 (Men), N = 217 (Women)

**Table 3.22 Descriptive statistics for the MA sample of the theta values of the subtests and G-factor for men and women**

Subtest	Men		Women	
	Mean	SD	Mean	SD
Series of Figures	0.746	1.01	0.680	0.96
Matrices	0.557	0.75	0.495	0.67
Series of Numbers	0.777	1.26	0.452	1.06
Diagrams	0.806	1.09	0.596	0.93
G-factor	0.348	0.48	0.259	0.43

N = 468 (Men), N = 334 (Women)

It is seen that the differences for the BA sample are relatively small. For the MA sample, it is seen that in particular for the subtest Series of Figures and Matrices, the differences in theta estimates for men and women are small. For the other two subtests as well as the G-factor there are some differences, where men tend to obtain higher values compared to women.

All effect sizes can be considered small, though the differences between men and women for the MA sample are somewhat larger compared to the BA sample. These results, shown in Table 3.23, are in agreement with the results based on the data of the construction sample.

**Table 3.23 Values of Cohen’s d for the differences between men and women for the BA and MA group**

Subtest	BA	MA
Series of Figures	-0.073	-0.095
Matrices	-0.098	-0.123
Series of Numbers	-0.170	-0.395
Diagrams	0.111	-0.293
G-factor	0.018	-0.276

**Ethnic background**

Next, the mean and standard deviation of the theta values for the different ethnic groups are given, again for the BA and MA sample separately in Tables 3.24 and 3.25

**Table 3.24 Descriptive statistics for the BA sample of the theta values of the subtests and G-factor for the different ethnic groups**

Subtest	Autochthon		Western minority		Non-western minority	
	Mean	SD	Mean	SD	Mean	SD
Series of Figures	0.335	0.74	0.514	0.86	0.201	0.66
Matrices	0.275	0.62	0.288	0.52	0.193	0.68
Series of Numbers	0.175	0.94	0.105	0.91	0.073	0.98
Diagrams	0.417	0.98	0.478	1.05	0.020	1.01
G-factor	0.092	0.38	0.082	0.32	-0.031	0.43

N = 366 (Autochthon), N = 65 (Western minority), N = 140 (Non-western minority)

**Table 3.25 Descriptive statistics for the MA sample of the theta values of the subtests and G-factor for the different ethnic groups**

Subtest	Autochthon		Western minority		Non-western minority	
	Mean	SD	Mean	SD	Mean	SD
Series of Figures	0.763	1.02	0.711	0.99	0.655	0.92
Matrices	0.557	0.74	0.564	0.71	0.455	0.66
Series of Numbers	0.661	1.16	0.657	1.30	0.597	1.27
Diagrams	0.815	1.02	0.710	1.08	0.387	0.92
G-factor	0.346	0.47	0.295	0.46	0.223	0.43

N = 582(Autochthon), N = 87 (Western minority), N = 133 (Non-western minority)

It is seen from Table 3.24 and Table 3.25 that the differences between autochthon and western minority candidates are small. For the BA sample, the differences between these two groups and the non-western minority group are somewhat larger for the subtests Series of Figures and Diagrams. This also has an effect on the values for the G-factor.

For the MA sample the differences are generally smaller. Only for the subtest Diagrams a larger difference can be observed between the non-western minority group and the other two groups. Whether the effects found are in fact meaningful will be studied by computation of effect sizes.

The values of Cohen's d computed for differences between theta values for different ethnic groups again are studied for the BA and MA separately. The results are shown in Table 3.26 and 3.27.

**Table 3.26 Values of Cohen's d for the differences between different ethnic groups for the BA sample**

Subtest	Autochthon vs. Western minority	Western minority vs. Non-western minority	Autochthon vs. Non-western minority
Series of Figures	-0.317	0.579	0.270
Matrices	-0.034	0.222	0.178
Series of Numbers	0.106	0.048	0.150
Diagrams	-0.085	0.628	0.563
G-factor	0.041	0.419	0.426

N = 366 (Autochthon), N = 65 (Western minority), N = 140 (Non-western minority)

**Table 3.27 Values of Cohen's d for the differences between different ethnic groups for the MA sample**

Subtest	Autochthon vs. Western minority	Western minority vs. Non-western minority	Autochthon vs. Non-western minority
Series of Figures	0.073	0.083	0.157
Matrices	-0.013	0.226	0.209
Series of Numbers	0.005	0.066	0.075
Diagrams	0.143	0.456	0.625
G-factor	0.154	0.229	0.387

N = 582(Autochthon), N = 87 (Western minority), N = 133 (Non-western minority)

The results for the BA sample show small differences for the subtests Matrices and Series of Numbers. For the subtests Series of Figures and Diagrams western minority candidates are found to have higher theta values compared to non-western minority candidates, there is a medium effect size. For the subtest Diagrams a medium effect has also been found for the differences between autochthon and non-western minority candidates. These results affect the found medium effect for differences in theta values of the G-factor.

The results for the MA sample show that nearly all differences between the groups can be considered to be small differences. Only the subtest Diagrams show differences with a medium effect size for the non-minority candidates compared to one of the other two groups of candidates.

## Age

In Table 3.28 and 3.29 the mean and standard deviation of theta for different age groups are provided, for both the BA and MA sample.

**Table 3.28 Descriptive statistics for the BA sample of the theta values of the subtests and G-factor for the different age groups**

Subtest	Age < 30		Age 30-45		Age > 45	
	Mean	SD	Mean	SD	Mean	SD
Series of Figures	0.318	0.77	0.290	0.68	0.137	0.77
Matrices	0.318	0.68	0.194	0.51	0.010	0.52
Series of Numbers	0.136	1.00	0.021	0.68	-0.153	0.83
Diagrams	0.002	1.10	0.421	1.07	0.274	1.18
G-factor	0.021	0.41	0.034	0.32	-0.098	0.39

N = 457 (< 30), N = 144 (30-45), N = 77 (45<)

**Table 3.29 Descriptive statistics for the MA sample of the theta values of the subtests and G-factor for the different age groups**

Subtest	Age < 30		Age 30-45		Age > 45	
	Mean	SD	Mean	SD	Mean	SD
Series of Figures	0.758	1.00	0.600	0.96	0.625	1.02
Matrices	0.564	0.73	0.392	0.65	0.427	0.74
Series of Numbers	0.727	1.22	0.213	0.94	0.482	1.11
Diagrams	0.670	1.01	0.924	1.13	0.784	1.16
G-factor	0.340	0.48	0.182	0.34	0.230	0.49

N = 655 (< 30), N = 128 (30-45), N = 35 (45<)

It is seen that in general the younger group of candidates (which is also the largest sample) shows higher theta values on both the subtests and G-factor compared to the other age groups. The subtest Diagrams is an exception for both the BA and MA sample, where the youngest age group shows the lowest theta values. The differences in sample size of the various age groups are quite large. The youngest age group is very large, and there are for example hardly any older candidates in the MA sample (N = 35).

In Table 3.30 and 3.31 the values of Cohen's d are given for the differences between age groups for the BA and MA sample respectively.

Age differences are found when measuring intelligence using Connector Ability, with older people scoring somewhat lower than young ones. However, effect sizes are small enough to warrant use of Connector Ability 1.1 in the general population.

**Table 3.30 Values of Cohen’s d for the differences between age groups for the BA sample**

Subtest	<30 vs. 30-45	30-45 vs. 45<	<30 vs. 45<
Series of Figures	0.055	0.298	0.332
Matrices	0.292	0.505	0.720
Series of Numbers	0.190	0.324	0.445
Diagrams	-0.546	0.185	-0.337
G-factor	-0.050	0.523	0.421

N = 457 (< 30), N = 144 (30-45), N = 77 (45<)

**Table 3.31 Values of Cohen’s d for the differences between age groups for the MA sample**

Subtest	<30 vs. 30-45	30-45 vs. 45<	<30 vs. 45<
Series of Figures	0.228	-0.036	0.186
Matrices	0.352	-0.071	0.264
Series of Numbers	0.667	-0.370	0.297
Diagrams	-0.335	0.173	-0.148
G-factor	0.537	-0.161	0.321

N = 655 (< 30), N = 128 (30-45), N = 35 (45<)

### 3.4 (Sub)test correlations

In Table 3.32, the correlations among the theta values of the subtests and G-factor are given. These correlations are based on theta values from the selection sample as described in Section 3.2.1.

**Table 3.32 Correlations among the theta values of the subtests and G-factor**

	Series of Figures	Matrices	Series of Numbers	Diagrams	G-factor
Series of Figures	1				
Matrices	.368**	1			
Series of Numbers	.369**	.364**	1		
Diagrams	.260**	.282**	.253**	1	
G-factor	.640**	.648**	.662**	.597**	1

N =2095 , \*\* p<.01 (2-tailed)

It is seen that the subtests have significant correlations. The correlations are moderate, which indicates that the subtests do not measure identical abilities, which would result in higher correlations given the level of reliabilities of the subtests. The correlations between the



subtests and the score on the G-factor are relatively high, which is to be expected as the subtest are a part of the G-factor and all are indicative of intelligence. Nevertheless, the correlations are not too close to 1, which means that each subtest contributes in its own way to the G-factor. No subtest is redundant or identical to the G-factor.

## 3.5 Norm development

### 3.5.1 Norm sample

The samples for the norm groups were composed by selecting candidates from the large selection sample, described in Section 3.2.1. These samples are as much as possible balanced for gender, age, ethnicity, organization size, line of business and occupation. Frequencies for the different samples are given in Table 3.33 through 3.35.

**Table 3.33**      **Frequencies for ME norm group**

Variable	Category	Frequencies	Percentages
Gender	Men	49	31
	Women	108	68
	Unknown	2	1
Age	< 30 years	101	64
	30 - 45 years	33	21
	> 45 years	20	13
	Unknown	5	3
Ethnic Background	Autochthon	115	72
	Western minority	10	6
	Non-western minority	33	21
	Unknown	1	1
Business	Wholesale and retail trade	2	1
	Financial and insurance activities	79	50
	Professional, scientific and technical activities	1	1
	Administrative and support service activities	7	4
	Human health and social work activities	4	3
	Other	66	42

N = 159

**Table 3.34**      **Frequencies for BA norm group**

Variable	Category	Frequencies	Percentages
Gender	Men	213	59
	Women	145	40
	Unknown	1	0
Age	< 30 years	139	40
	30 - 45 years	125	36
	> 45 years	84	24
Ethnic Background	Autochthon	256	71
	Western minority	41	11
	Non-western minority	59	16
	Unknown	3	1
Business	Manufacturing	7	2
	Wholesale and retail trade	2	1
	Financial and insurance activities	103	29
	Professional, scientific and technical activities	13	4
	Administrative and support service activities	15	4
	Human Health and social work activities	13	4
	Other	206	57

N = 359

**Table 3.35**      **Frequencies for MA norm group**

Variable	Category	Frequencies	Percentages
Gender	Men	205	57
	Women	153	42
	Unknown	4	1
Age	< 30 years	217	61
	30 - 45 years	102	29
	> 45 years	35	10
Ethnic Background	Autochthon	261	72
	Western minority	40	11
	Non-western minority	57	16
	Unknown	4	1
Business	Manufacturing	8	2
	Financial and insurance activities	126	35
	Professional, scientific and technical activities	54	15
	Administrative and support service activities	29	8
	Other	145	40

N = 362

### 3.5.2 Norms

The norms that are obtained from these data are given in Table 3.36. The norms are used to compute T-scores from the estimates theta values. See for more information Section 2.4.1.

**Table 3.36 ME, BA and MA norms**

Subtest	ME		BA		MA	
	Mean	SD	Mean	SD	Mean	SD
Series of Figures	0.053	0.723	0.276	0.723	0.633	0.723
Matrices	0.105	0.641	0.237	0.641	0.481	0.641
Series of Numbers	-0.334	0.858	0.051	0.858	0.447	0.858
Diagrams	0.017	1.041	0.363	1.041	0.671	1.041
G-factor	-0.206	0.386	0.036	0.386	0.260	0.386



# Chapter 4

## Psychometrics

Research concerning IRT based reliability analyses and test-retest reliability studies is reported. Furthermore, construct, criterion-related and discriminant validity studies are described. Finally, a study is reported concerning adverse impact.

### 4.1 Reliability

Usually, reliability is studied by inspecting measures of internal consistency such as values of Cronbach's alpha. However, Connector Ability 1.1 is an adaptive test and does not consist of a fixed set of items. Furthermore, IRT allows measurement precision to be determined at different levels of theta using the information function (see Equation 1) to compute the standard error of estimation (see Equation 3). Therefore, reliability will be reported here based on this standard error. Furthermore, the test-retest reliability will be reported as well.

#### 4.1.1 IRT-based reliability

The classical formulation of reliability has little relevance when measurement is based on IRT. The standard error of estimation is a function of ability (in this case theta). Generally, there will be relatively low standard error in the mid-range of theta, and relatively high standard error for low and high theta values. Of course the standard error depends on the applied stop criteria (see also Section 2.3.5).

An overall estimate of reliability may be obtained by;

$$(9) \quad r_t = \left( \sigma^2(\theta_t) - \frac{1}{\bar{I}(\theta_t)} \right) / \sigma^2(\theta_t)$$

where  $\sigma^2(\theta_t)$  denotes the variance of the theta scale of a given subtest. The subtest information  $I(\theta_t)$  is calculated from Equation 6 for each person, after which the mean information value  $\bar{I}(\theta_t)$ <sup>1</sup> for the given sample is computed. This calculation of reliability takes into account the variance of theta in the sample (for the derivation, see: Green, Bock, Humphreys, Linn, & Reckase, 1984).

---

<sup>1</sup> The bar denotes that that it concerns a mean value

For the reliability of the G-factor theta value, first for each individual,  $i$ , separately the test information  $I(\hat{\theta}_{i,tot})$  is computed by taking the sum of the information of the four subtests; that is,

$$(10) \quad I(\hat{\theta}_{i,tot}) = \frac{1}{(SE(\hat{\theta}_{i,tot}))^2} = \sum_{t=1}^4 I(\hat{\theta}_{it})$$

In accordance with the general procedure for computing the test information (see for example Embretson & Reise, 2000), the information values of the individual subtests are summed. To justify this procedure, in advance, the underlying assumption of unidimensionality has been examined. From a principal component factor analysis only one single factor was derived with an initial eigenvalue above one (eigenvalue = 1.9). These results support the assumption on which the additive procedure is based.

Next, the mean test information  $\bar{I}(\theta_{tot})^2$  for the given sample is computed. Subsequently, the reliability of the G-factor theta value  $r_{tot}$  is computed by;

$$(11) \quad r_{tot} = \left( \sigma^2(\theta_{tot}) - \frac{1}{\bar{I}(\theta_{tot})} \right) / \sigma^2(\theta_{tot})$$

Because the subtests are positively correlated and all have non-negative weights, the standard error of the composite (G-factor) will be smaller than that of any of the subtests (Wainer, 2000). Correspondingly, the true reliability of the G-factor will be higher, under the condition of equal variances.

### Construction sample

In this section, first the results for the construction sample will be given. It is important to note that these participants were not administered the adaptive version of the test, but responded to a fixed set of items. This means that the standard error and reliability are based on the responses of participants who responded to different sets of seven through 14 items. These items were not yet adjusted to the appropriate ability level of the participant as will be the case when the adaptive version of the test is administered. As a consequence, the mean standard error of the adaptive test can be expected to be lower compared to the mean standard error for these fixed sets of items.

The mean standard error of estimation as well as its standard deviation based on the data of the construction sample are given in Table 4.1. The reliabilities ( $r$ ) are computed as described in Equation 9 and 11.

---

<sup>2</sup> The bar denotes that that it concerns a mean value

**Table 4.1 Mean standard error and reliability of four subtests as well as the G-factor based on construction sample data**

Subtest	Mean [SE( $\theta$ )]	SD [SE( $\theta$ )]	r
Series of Figures	0.535	0.43	0.89
Matrices	0.428	0.23	0.75
Series of Numbers	0.476	0.38	0.82
Diagrams	0.632	0.38	0.72
G-factor	0.213	0.06	0.83

N=3258

The mean reliabilities are in the range of 0.72 and 0.89 for the individual subtests, and 0.83 for G-factor.

The results for the BA and MA sample respectively are given in Table 4.2.

**Table 4.2 Mean standard error and reliability of four subtests as well as the G-factor based on construction sample data for the BA and MA sample**

Subtest	BA			MA		
	Mean SE( $\theta$ )	SD SE( $\theta$ )	r	Mean SE( $\theta$ )	SD SE( $\theta$ )	r
Series of Figures	0.476	0.35	0.88	0.596	0.50	0.90
Matrices	0.404	0.20	0.73	0.453	0.26	0.75
Series of Numbers	0.434	0.25	0.76	0.521	0.48	0.85
Diagrams	0.598	0.34	0.80	0.668	0.43	0.81
G-factor	0.202	0.05	0.80	0.225	0.08	0.83

The results show that in general a more reliable theta value is estimated for participants from the MA sample, compared to the BA sample, whereas the mean SE( $\theta$ ) is evidently lower for the BA sample. The same items are provided to the participants of the BA and MA sample. The items are dispersed along difficulty level across the scale of theta. An item provides more information and in this way reduces more of the standard error, when the item has a difficulty level near the theta value of a participant. For participants of the MA sample, the items often may be more deviant from the theta level of a participant, which means that in particular relatively easy items do not contribute much to the estimation of theta. These items do not lower the standard error of the estimate for the MA sample as much as for the BA sample. Because the variance of the theta scale for the BA sample is lower compared to the MA sample, the differences between the two samples with respect to reliabilities are reduced, or even reversed.

It should be noted that for the G-factor the standard error of estimation is low for both samples. As the variance of this theta is relatively small (see also Table 3.10), the reliabilities are good, though not as high as might be expected from the results of the subtests.

An important consideration is that advice with respect to the suitability of a candidate for a given position or job using Connector Ability should always be based on the score of the candidate on the G-factor. Decisions should never be based on scores on individual subtests. In the report, each subtest score is given, as well as a bar around the score. The bar shows the margin around the score. In three-quarters of cases, the score will be within this margin is the candidate would take the test again. This is also especially stressed in the certification training for prospective users of Connector Ability.

### Selection sample

The data of the selection sample are obtained in a selection context where Connector Ability 1.0 was administered. Table 4.3 through Table 4.5 show the mean standard error as well as its standard deviation for the ME, BA and MA selection sample respectively. The reliabilities ( $r$ ) computed from the mean standard error are given as well.

It is seen that the reliability for all of the theta estimates are well above the mean reliability values computed for the construction sample, which was based on a fixed set of items. Of course the mean standard errors are lower, as could be expected as well. For the construction sample, each respondent responded to 14 items for each subtest. The selection sample that took the adaptive version of the test, had to respond to at least 10 and a maximum 15 items, depending on when the stop criterion was met. The stop criterion for the subtests was set at a standard error of 0.54. Again, the mean standard error of the G-factor is lower, as it combines the four subtests.

**Table 4.3 Mean standard error and reliability of four subtests as well as the G-factor based on ME selection sample data**

Subtest	Mean SE( $\theta$ )	SD SE( $\theta$ )	$r$
Series of Figures	0.263	0.08	0.72
Matrices	0.287	0.08	0.72
Series of Numbers	0.306	0.19	0.93
Diagrams	0.380	0.14	0.90
G-factor	0.139	0.02	0.83

N = 159

**Table 4.4 Mean standard error and reliability of four subtests as well as the G-factor based on BA selection sample data**

Subtest	Mean SE( $\theta$ )	SD SE( $\theta$ )	$r$
Series of Figures	0.298	0.14	0.88
Matrices	0.330	0.11	0.79
Series of Numbers	0.312	0.19	0.92
Diagrams	0.401	0.15	0.90
G-factor	0.149	0.03	0.87

N = 679



**Table 4.5 Mean standard error and reliability of four subtests as well as the G-factor based on MA selection sample data**

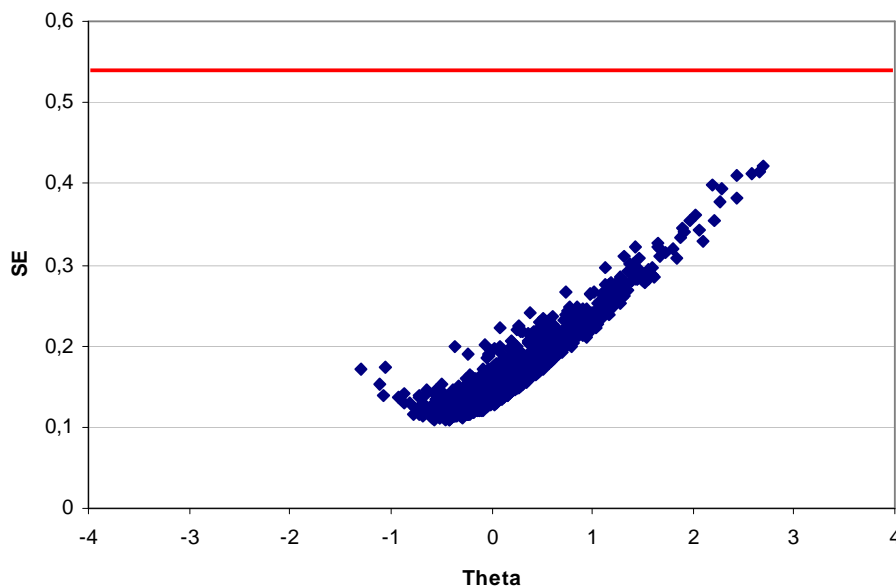
Subtest	Mean SE( $\theta$ )	SD SE( $\theta$ )	r
Series of Figures	0.372	0.20	0.92
Matrices	0.377	0.14	0.80
Series of Numbers	0.408	0.26	0.94
Diagrams	0.464	0.17	0.85
G-factor	0.173	0.05	0.88

N = 822

The tables also show that the mean standard errors have the lowest values for the ME sample, followed by the BA sample. This is not surprising, as the theta values of the ME sample are in a range where many items provide information. Many items have a difficulty parameter near their estimated theta values, and these items have relatively high discrimination parameters, which enhances the item information reducing SE.

Figure 4.1 shows for all candidates from the selection sample, their theta values and the corresponding standard error. This figure demonstrates that the standard error depends on theta, as well as on which items were administered to a candidate. For one set of items, the SE of a theta estimate depends on the value of theta, as the standard error is determined by the item parameters. Furthermore, Connector Ability is an adaptive test, which means that each candidate may respond to different items.

**Figure 4.1**  
Standard Error given Theta on G-factor



In Figure 4.1, a red line is drawn at an SE level of 0.54, which corresponds to a reliability value of 0.70. It is seen that the G-factor theta values are all measured reliably. Furthermore, in a range of theta from -1 through 0, theta is measured with the highest reliability. Higher values of theta show somewhat higher standard errors, though still well under an SE of 0.54.

One of the stop criteria is that a SE value of 0.54 or lower is obtained. When the estimate of theta has an SE of 0.54 or lower, the subtest will be ended, of course only when a minimum of 10 items has been administered. The second stop criterion is that a maximum of 15 items may be administered for one subtest.

In Table 4.6 the number and percentage of candidates is given for whom their theta value is estimated with an SE below 0.54 at the end of the (sub)test. The results are shown for the ME, BA and MA sample separately.

**Table 4.6** Number and percentage of candidates for whom theta is estimated with SE < 0.54 at the end of the (sub)tests for the ME, BA and MA sample

Subtest	ME		BA		MA	
	Number	Percentage	Number	Percentage	Number	Percentage
Series of Figures	157	99	648	95	732	89
Matrices	157	99	615	91	661	80
Series of Numbers	150	94	638	94	690	84
Diagrams	148	93	635	94	684	83
G-factor	159	100	679	100	822	100

It is seen that for most candidates their theta value is estimated with sufficient accuracy, in particular for the ME and BA sample. The percentages are somewhat lower for the MA sample, which can be explained by the relatively high theta values estimated for this sample. It is important to note that for selection decisions only the G-factor to be taken into account, which shows standard errors all well below 0.54. With candidates for whom the stop criterion of SE < 0.54 is not met, the theta values are all relatively high (above 1.5) or very low (below - 3.6) on a subtest level.

#### 4.1.2 Test-retest reliability

A sample of 124 respondents participated in the second pilot study to obtain test-retest data. The participants are all students from University Twente, with different majors. The participants were administered Connector Ability 1.0. The test was administered online and under supervision. Some characteristics of the sample are provided in Table 4.7.

**Table 4.7**      **Frequencies for test-retest sample**

Variable	Category	Frequencies	Percentage
Gender	Men	52	42
	Women	72	58
Age	< 30 years	119	96
	30 - 45 years	3	2
	> 45 years	2	2
Nationality	Dutch	100	81
	German	24	19

N = 124

The time interval between the two measurement moments was at least 2 weeks. Over 70 percent of the respondents took the second test within 3 weeks after the first administration. Table 4.8 shows specific information with respect to the time interval between the first and second administration of Connector Ability 1.0.

**Table 4.8**      **Frequencies of Time interval**

Time interval	Frequencies	Percentage
14 -20 days	74	60
21- 27 days	30	24
More than 27 days	20	16

T-scores below 30 are restricted to equal 30. The same is done for t-scores above 70, which are restricted to equal 70. A first reason is to cope with outliers. Second, a respondent receives a maximum t-score of 70 in their report. Above these values, specific estimation is not possible. T-scores are all based on the MA norm, as the participants are all students on MA educational level.

The test-retest correlation coefficient for the T-scores on the G-factor is .720\*\* for the total sample.

## 4.2 Validity

The validity of Connector Ability is investigated by means of several studies. First, the results with respect to construct validity are described, followed by two studies concerning criterion-related validity. Finally, discriminant validity is inspected as well.

### 4.2.1 Construct validity

A sample of 101 respondents participated in a study to obtain construct validity data. The participants are employees of a number of financial and insurance companies, a law office and a HRM consultancy firm and some students from University Utrecht with different majors.

The participants were administered the Connector Ability 1.0 and the PiCompany test Connector C 3.1. (PiCompany, 2005). Connector C 3.1 is a classic language dependent test for cognitive ability. Its total score has a reliability of above .90. Over 85 percent of the respondents took one of the tests in a selection context, the second test was administered voluntarily within this study. All of the tests were administered online and under supervision. Some characteristics of the sample are provided in Table 4.9.

**Table 4.9**      **Frequencies**

<b>Variable</b>	<b>Category</b>	<b>Frequencies</b>	<b>Percentage</b>
Gender	Men	50	50
	Women	50	50
Age	< 30 years	30	38
	30 - 45 years	32	40
	> 45 years	18	22
Ethnic Background	Autochthon	85	88
	Western minority	8	8
	Non-western minority	4	4

N = 101

The time interval between the two measurements varies. Over 70 percent of the respondents took the second test within 2 years after the first administration. Table 4.10 shows specific information with respect to the time interval between the first and second administration of the Connector Ability 1.0 or Connector C 3.1.

**Table 4.10**      **Frequencies of Time interval**

<b>Time interval</b>	<b>Frequencies</b>	<b>Percentage</b>
0 -7 days	15	15
0- 2 years	56	55
More than 2 years	30	30

In Connector Ability T-scores below 30 are restricted to equal 30. The same is done for t-scores above 70 that are restricted to equal 70. A first reason is to cope with outliers. Second, a respondent receives a maximum t-score of 70 in their report. Above these values, specific estimation is not possible.

The main difference between Connector Ability 1.0 and Connector C 3.1 is the by construction intended absence of possible differences between respondents in numerical and verbal competencies not directly related to G. In fact, this exactly was the reason to develop Connector Ability in the first place. Both tests should therefore not be regarded as strict equivalents, conceptually Connector Ability to be seen as a better approximation of what is meant by G.

The correlation coefficient for the T-scores on the G-factor is .552\*\* for the total sample. In view of the mentioned only partial equivalence between both tests as far as the measured construct is concerned, this is a value that lies within a range that might be expected.

#### **4.2.2 Criterion-related validity**

Two studies have been performed to estimate the criterion-related validity of Connector Ability.

##### **Association parental occupational level and intelligence**

Scientific research on the relationship between socio-economic status (SES) and intelligence leaves little doubt that people with higher scores on IQ tests are better educated, hold more prestigious occupations, and earn higher incomes than people with lower scores (Gottfredson, 1997, 2003; Jensen, 1980, 1998; Schmidt & Hunter, 2004).

SES comprises a number of indicators, including income, education and occupation of the parents or family. Measured properly at the level of the individual, SES reflects the occupations and thus the underlying levels of education and resulting incomes of the adult members of a household (Jeynes, 2002; White, 1982). A composite measure of SES is not available, though the parental occupational level is a background variable asked in Connector Ability to study its association with intelligence. It is expected that the occupational level of parents is to be positively related to scores on the G-factor.

##### *Subjects*

In a selection context, 2095 candidates were administered Connector Ability 1.0. The composition of the (selection) sample is described in Section 3.2.1.

##### *Measures*

Intelligence (G-factor) is measured by Connector Ability 1.0. The BA norms are used to compute the T-scores on the G-factor. T-scores below 30 are restricted to 30, and scores above 70 are restricted to 70.

Parental occupational level is asked through one question; "What describes the work of your parents best?". There are three response categories; low-level labour, administrative and other professional jobs; mid-level labour, administrative and other professional jobs; and high-level professional and scientific work.

### Results

The mean G-factor T-scores are computed for the three levels of occupation of the parents, see Table 4.11.

**Table 4.11 Mean, Standard deviation of G-factor T-scores as well as frequencies for three parental occupational levels**

Occupational level	Mean	SD	N
Low	51.45	8.9	361
Medium	53.33	9.3	877
High	55.51	8.8	694

A one-way ANOVA was used to test for differences in G-factor T-scores among the three parental occupational levels. T-scores differed significantly across the three occupational levels,  $F(3, 1949) = 17.32, p < .01$ . Tukey post-hoc comparisons of the three groups indicate that all groups differ in the mean G-factor T-scores.

The group with a lower parental occupational level scored lower ( $M = 51.54$ ) compared to both the medium level ( $M = 53.33$ ),  $p = .005$ , and higher level ( $M = 55.51$ ),  $p < .001$ . The medium and higher groups of parental occupation level differed significantly as well,  $p < .001$ .

### Associations of G-factor scores with educational effort

The educational system is based on differences in intelligence and expressed by the categorization of different educational levels. These educational levels are also a basis for composition of norms for different educational levels. People who obtain a degree on a certain educational level still vary with respect to their intelligence. Yet, all participants have obtained this degree. It can be argued that effort to obtain a degree on a given educational level depends, among others, on intelligence.

More intelligent individuals may have to put less effort to obtain their degree, compared to individuals who are less intelligent. The association of intelligence and the effort into education is studied from two perspectives. First, students who did and did not double a class during their secondary education were compared with respect to their intelligence. It is expected that these two groups differ in mean intelligence, where students who doubled a class are expected to show a lower mean intelligence compared to students who did not double a class. Students who doubled a class have put more effort, including time, into obtaining the same degree compared to students who did not double a class.

Second, it was examined whether time spend on homework is related to G-factor T-scores. The opinion of the student was asked, regarding the time spent on homework in comparison with their classmates. It is expected that participants with higher G-factor scores may have spent less time on their homework compared to participants with lower scores, as they have had to put less of an effort to complete their homework. Evidently, the kind or type of homework is an important variable. A minimum amount of time may be necessary to do

homework that requires writing or reading (we will refer below to a 'language' subject area). These skills are not directly related to cognitive skills. However, mathematics and physics (we will refer below to a 'math' subject area) for example require some skills that do rely for some part on cognitive skills. This results in the expectation that time spend homework for the 'language' subject area is less reliant on cognitive skills or intelligence compared to the 'math' subject area.

### *Subjects*

A sample of students from various disciplines has participated in the second pilot study to obtain test-retest data. More information about the sample is provided in Section 4.1.2.

### *Measures*

Intelligence (G-factor) is measured by Connector Ability 1.0. The MA norms are used to compute the T-scores on the G-factor, as all participants are students at a University to obtain a Masters degree. T-scores below 30 are restricted to 30, and scores above 70 are restricted to 70.

At the time of their first administration of Connector Ability, the participants were asked to complete a questionnaire concerning their educational background and related experiences. The participants were asked how long it took them to do their homework at secondary education, compared to their classmates. The response categories were 'Far less time', 'Less time', 'About an equal amount of time', 'More time', 'Much more time'. The question was asked for subject area 'language, history and society' (called 'language') as well as for 'mathematics, physics and technology' (called 'math'). Furthermore, it was asked whether the students had doubled a class during their secondary education.

### *Results*

There were 12 students that indicated to have doubled a class, and 75 students who did not double a class. A one-way ANOVA was used to test the hypothesis concerning doubling a class and G-factor T-scores. T-scores differed significantly for participant that have doubled a class, and those who did not,  $F(1, 91) = 4.66, p = .034$ . Participants who doubled a class showed a mean G-factor T-score of 53.3, whereas non-doubling participants showed a mean T-score of 58.6.

Next, Pearson's correlation coefficients were computed for the relation between time spent on homework and G-factor T-scores.

**Table 4.12 Correlations between time spent on homework for 'math' and 'language' subject areas and G-factor T-scores at measurement 1 and 2**

	T-score 1	T-score 2
Time 'math' homework	- 0.339 **	- 0.320 **
Time 'language' homework	- 0.191 *	- 0.129

The results in Table 4.12 show significant negative correlations for time spent on homework concerning 'math' subject areas, at both measurements. Participants indicated to have spent less time on their 'math' homework compared to their classmates as they obtain higher G-factor T-scores. The correlations for time spend on 'language' homework are lower. Intelligence seems mainly to be associated with effort with respect to time spend on 'math' homework.

### 4.5.3 Discriminant validity

Research with respect to personality structure seems to have agreed on the presence of five personality factors (Costa & McCrea, 1992). The five personality traits are Neuroticism, Extraversion, Openness to Experience, Agreeableness and Conscientiousness. Personality traits vary across the population, as does intelligence. However, these dimensions are unrelated (Gardner, 1983). For Connector Ability 1.1, this implies that the G-factor score should be unrelated to any big five personality trait. This hypothesis will be evaluated by correlating G-factor T-scores of Connector Ability with both factors and facets from a big five personality inventory.

#### *Subjects*

During an assessment, in general both Connector Ability 1.0 and Reflector Big Five Personality (PiCompany, 2007) are administered. A total of 356 persons were administered both tests during six months. The sample consists of 135 women (38 %) and 223 men (62 %). Most participants are autochthon (294), whereas 26 participants are part of a western minority group and 29 are non-western minority group members. A MA educational level is obtained by 134 participants (37 %), and a BA educational level by 150 participants (42 %). For the rest of the sample, other educational levels apply or the educational level is not registered (21 %). The mean age is 38 with a standard deviation of 10.

#### *Instruments*

Intelligence (G-factor) is measured by Connector Ability 1.0. T-scores on the G-factor are used to determine correlations with personality factors. To handle the presence of outliers, scores below 30 are restricted to 30, and scores above 70 are restricted to 70 (which is the range that is reported to a candidate).

The big five personality factors as well as the facets are measured by Reflector Big Five Personality (RBFP; PiCompany, 2007). The RBFP is a modern online personality questionnaire providing a comprehensive overview of how an employee scores on the five most important personality traits on which people differ, as well as a number of aspects underlying these five traits.



Reflector Big Five Personality covers the following five personality factors:

- Instability: the extent to which we respond emotionally to setbacks (corresponding to the factor often called Neuroticism);
- Extraversion: the extent to which we actively maintain contacts with others;
- Openness: the extent to which we look for new experiences and new ideas;
- Accommodation: the extent to which we place other people's interests above our own (corresponding to the factor often called Agreeableness);
- Conscientiousness: the extent to which we act in an organised and goal-oriented manner.

The questionnaire focuses on behaviours that people show in work situations. The report paints a portrait of the measured personality traits of an employee insofar as they match the competencies required for the work they do. Both factors and facets are reported in T-scores, like for Connector Ability.

#### Results

Table 4.13 shows the correlations between the T-scores of the G-factor and the five personality traits. The correlations are all very small, only Openness has a small significant correlation ( $p = 0.044$ ).

**Table 4.13 Pearson Correlations between T-score on G-factor and the five personality factors**

	G-factor
Instability	-0.008
Extraversion	0.047
Openness	0.107 *
Accommodation	-0.054
Conscientiousness	0.034

N = 356

These results support the assumption that personality and intelligence are unrelated. The correlations between intelligence and the facets underlying the personality factor have also been inspected. The correlations of the G-factor with the facets underlying the factors Instability, Extraversion, Accommodation and Conscientiousness are all small and not significant. Of the factor Openness, one facet is significantly correlated with the G-factor. The facet Complexity (the degree to which a person conceives matters as complex) shows a correlation with the G-factor of 0.195,  $p < 0.01$ . The correlation is significant though still quite small, which is in agreement with what is often found in literature (Ashton, Lee, Vernon, & Jang; 2000).

### 4.3 Adverse impact

A large organization providing tax advice, transaction advisory services, accountancy and legal advice has been using Connector C 3.1 (PiCompany, 2005) as a part of their personnel selection process. It noticed that cultural background had a relevant influence on the test results. To make a more fair comparison between candidates they have decided to adopt Connector Ability.

After having used Connector Ability 1.0 for a couple of months, the results on the two tests (Connector C 3.1 and Connector Ability 1.0) for candidates from different cultural backgrounds were compared. Candidates were classified in two groups; autochthon candidates and candidates from ethnic minority groups. The characteristics of the Connector C sample and the Connector Ability sample respectively are given below in Table 4.14 and Table 4.15.

**Table 4.14**      **Frequencies for Connector C sample**

Variable	Category	Frequencies		
		Autochthon	Minorities	Total
Gender	Men	134	120	254
	Women	66	79	145
	Unknown	0	1	1
Age	< 30 years	188	192	380
	30 - 50 years	9	6	15
	Unknown	3	2	5
Educational Level	BA	100	100	200
	MA	100	100	200

N = 400

**Table 4.15**      **Frequencies for Connector Ability sample**

Variable	Category	Frequencies		
		Autochthon	Minorities	Total
Gender	Men	386	176	562
	Women	204	110	314
Age	< 30 years	575	276	851
	30 - 50 years	15	8	23
	Unknown	0	2	2
Educational Level	BA	268	169	437
	MA	322	117	439

N = 876

Table 4.16 shows the mean G-factor T-scores for autochthon and ethnic minority members for both Connector C 3.1 and Connector Ability 1.0. The results are presented for the BA and MA sample separately.

**Table 4.16 Mean G-factor T-scores for different ethnic groups on Connector C and Connector Ability**

Education	Ethnic background	Connector C 3.1		Connector Ability 1.0	
		Mean	SD	Mean	SD
BA	Autochthon	48.93	9.07	53.89	7.85
	Minority	39.82	9.08	49.78	7.97
MA	Autochthon	52.36	8.79	54.25	9.22
	Minority	41.76	9.87	50.52	8.29

It is seen in Table 4.16 that candidates from ethnic minority groups have a 10 points lower score on Connector C compared to autochthon candidates. This deviation decreases to four points on Connector Ability.

Connector C consists of eight subtests. Four of them are, as regards substantive content, the relatively most comparable to those in Connector Ability. The mean G-factor T-scores based on these four subtests of Connector C are also compared to the scores of Connector Ability. The results are shown in Table 4.17 below. It is seen that candidates from ethnic minority groups have an eight points lower score on the four subtests Connector C compared to autochthon candidates.

**Table 4.17 Mean G-factor T-scores for different ethnic groups on Connector C four subtests and Connector Ability**

Education	Ethnic background	Connector C four subtests		Connector Ability	
		Mean	SD	Mean	SD
BA	Autochthon	53.29	8.48	53.89	7.85
	Minority	44.86	8.94	49.78	7.97
MA	Autochthon	55.11	8.54	54.25	9.22
	Minority	46.79	10.26	50.52	8.29

Generally, criteria were specified that candidates had to meet in order to be selected for a position. Here, the selection criterion is a minimum T-score that has to be obtained on a cognitive ability test, i.e. T-score > 45 or > 50. To illustrate the consequences for personnel selection, Tables 4.18 and 4.19 present the rate of autochthon and minority candidates that have been selected for both tests.

**Table 4.18 BA selection rate for different ethnic groups by Connector C or Connector Ability, applying two selection criteria**

Criterion	Ethnic background	Selection rate	
		Connector C	Connector Ability
T > 45	Autochthon	65%	89%
	Minority	27%	68%
T > 50	Autochthon	44%	66%
	Minority	13%	47%

**Table 4.19 MA selection rate for different ethnic groups by Connector C or Connector Ability, applying two selection criteria**

Criterion	Ethnic background	Selection rate	
		Connector C	Connector Ability
T > 45	Autochthon	80%	84%
	Minority	56%	71%
T > 50	Autochthon	65%	65%
	Minority	17%	50%

It is seen that the differences in selection rates between autochthon and ethnic minority group members substantially decrease when Connector Ability is used, compared to the use of Connector C, thus reducing adverse impact to a large extent as should be expected from the characteristics of both tests.

# References

- Angoff, W.H. (1993). Perspectives on differential item functioning methodology. In P.W. Holland, & H. Wainer (Eds.), *Differential item functioning* (p. 3-24). Hillsdale, NJ: Lawrence Erlbaum.
- Ashton, M.C., Lee, K., Vernon, P.A., & Jang, L. (2000). Fluid intelligence, crystallized intelligence and the Openness/Intellect factor. *Journal of Research in Personality, 34*, 198-207.
- Bochhah, N., Kort, W. & Seddik, H. (red.) (2005). *Richtlijnen gebruik diagnostische instrumenten bij etnische minderheden*. [Guidelines for the use of diagnostic instruments among ethnic minorities] Rotterdam, LBR/NIP.
- Bochhah, N., Kort, W. & Seddik, H. (red.) (2005). *Toepasbaarheid van enkele psychologische tests bij personeelsbeoordelingen bij etnische minderheden*. [Suitability of a few psychological tests when performing personnel appraisals on ethnic minorities] Rotterdam, LBR/NIP.
- Chuah, S.C., Drasgow, F., & Leucht, R. (2006). How big is big enough? Sample size requirements for CAST item parameter estimation. *Applies Measurement in Education, 19*, 241-255.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Costa, P.T., & McCrae, R.R. (1992). *NEO PI-R: Professional Manual*. Odessa, FL: Psychological Assessment Resources.
- Dodd, B.G., Koch, W.,R., & De Ayala, R.J. (1989). Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement, 13*, 129-144.
- Embretson, S.E., & Reise, S.P., (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- Gardner, H. (1983) *Frames of Mind: The theory of multiple intelligences*. New York: Basic Books.
- Glas, C.A.W. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica, 8*, 647-667.

- Glas, C.A.W. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika*, 64, 273-294.
- Green, B.F., Bock, R.D., Humphreys, L.G., Linn, R.L. and Reckase, M.D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, 21, 347-360.
- Gottfredson, L. S. (2002). Where and why *g* matters: Not a mystery. *Human Performance*, 15, 25-46.
- Herrnstein, R. & Murray, C (1994). *The Bell Curve: Intelligence and Class Structure in American Life*. New York: The free press.
- Jensen, A. R. (1980). *Bias in mental testing*. New York: The Free Press.
- Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.
- Jeynes, W. H. (2002). The challenge of controlling for SES in social science and education research. *Educational Psychology Review*, 14, 205–221.
- Kline, P. (1992). *Intelligence: The psychometric view*. London: Routledge
- Kowall, M. A., Watson, G. M. W., & Madak, P. R. (1990). Concurrent validity of the test of nonverbal intelligence with referred suburban and Canadian native children. *Journal of Clinical Psychology*, 46, 632-636.
- Mackintosh, N. J. (1998). *IQ and human intelligence*. Oxford: Oxford University Press.
- Mislevy, R.J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56, 177-196.
- Naglieri, J. A., & Ronning, M. E. (2000). Comparison of White, African American, Hispanic and Asian children on the Naglieri Nonverbal Ability Test. *Psychological Assessment*, 12, 328-334.
- PiCompany (2005), *Connector C 3.1: Achtergrond en ontwikkeling van een computer ondersteunde intelligentietest*. [Connector C: Background and construction of a computer supported intelligence test]. Utrecht, The Netherlands: PiCompany
- PiCompany (2007), *Workplace Big Five (Reflector Big Five Personality) 2.0: Professional Manual*. Utrecht, The Netherlands: PiCompany.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124, 262-274.

Schmidt, F. L., & Hunter, J. E. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology*, 86, 162-173.

Thissen, D., Chen, W.H., & Bock, R.D. (2003). *Multilog 7.0 [Computer program]* Chicago: Scientific Software.

Van der Linden, W. J., & Glas, C. A. W. (Eds.) (2000). *Computerized adaptive testing: Theory and practice*. Norwell, MA: Kluwer.

Wainer, H. (Eds.) (2000). *Computerized Adaptive Testing: A Primer*. Mahwah, NJ: Lawrence Erlbaum.

White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin*, 91, 461–481.





## **Appendix A**

### **Candidate Brochure**



**pi**Company

 **Connector** *ABILITY*

*Candidate brochure*

P E O P L E I M P R O V E  
P E R F O R M A N C E

# Content

1	Why this brochure?	3
2	Why tests?	3
3	What is Connector Ability: what is being tested?	4
4	How does computer testing work?	4
4.1	Taking the test	4
4.2	Which steps to take when taking the test?	4
4.2.1	Personal data and background questions	4
4.2.2	General instruction	5
4.2.3	Instruction per subtest	5
4.2.4	The actual test	5
4.3	Protection of personal data	6
5	How to prepare?	6
5.1	The basis for reliable test conditions	6
5.2	Instruction and sample questions	6
5.2.1	Series of Figures	7
5.2.2	Matrices	14
5.2.3	Series of Numbers	22
5.2.4	Diagrams	27

## 1 Why this brochure?

The purpose of this brochure is to give you insight into the upcoming testing procedure. It is important for you to start the testing procedure well-informed and to be relieved from any uncertainties about the ins and outs of the tests in the procedure.

Specifically, you will receive an instruction and sample questions for the Connector Ability test.

The brochure wishes to answer the following questions:

- Why tests?
- What is Connector Ability: what is being tested?
- How does computer testing work?

It is very important that you prepare yourself for the test. Then you know what to expect and, besides that, it is important that everyone who takes this test knows what the test is about. Therefore, this brochure thoroughly enters into:

- How to prepare?
- Instruction per subtest and sample questions.

## 2 Why tests?

Tests are used to gain an image of the candidate that is as objective as possible. Each candidate is given a large number of questions to answer and problems to solve. The test situation is the same for each candidate. The candidate's results are compared with the results of a large group of people who have taken the same test and who have had a similar education. The results of these people have been processed into a table of comparison, also referred to as a norm table.

Tests generally give a reliable image of a person's intellectual abilities and personal characteristics. Characteristics which are not relevant, such as race, sex or appearance, have no bearing on the result.

Finally, tests are used because they have a relatively high predictive value. The connection between the test results and (later) behavior in the position is examined. If, for example, it appears that a lot of people with the same test results do well in a particular profession, then we can expect that someone with a similar test result will also perform well in that profession.

### **3 What is Connector Ability: what is being tested?**

Connector Ability is an intelligence test, a test that measures problem-solving ability.

Connector Ability consists of four subtests:

- Series of Figures. This subtest measures the ease with which someone can complete logical reasoning;
- Matrices. This subtest measures the ease with which someone can analyse and continue complicated relationships;
- Series of Numbers. This subtest measures the ease with which someone can analyse and continue the relationship between numbers;
- Diagrams. This subtest measures the ease with which someone can make connections between concepts.

### **4 How does computer testing work?**

#### **4.1 Taking the test**

The Connector Ability will be administered with the help of a computer. Even if you have little experience with computers you will notice that computer testing is very easy.

The questions appear on screen, after which you choose from a number of possible answers. This is done with the computer mouse. Pen and paper will be ready for use. These are also the only aids that you may use during the test and which you are to hand in at the end. Cellular phones, calculators and this brochure are not allowed in the testing room.

#### **4.2 Which steps to take when taking the test?**

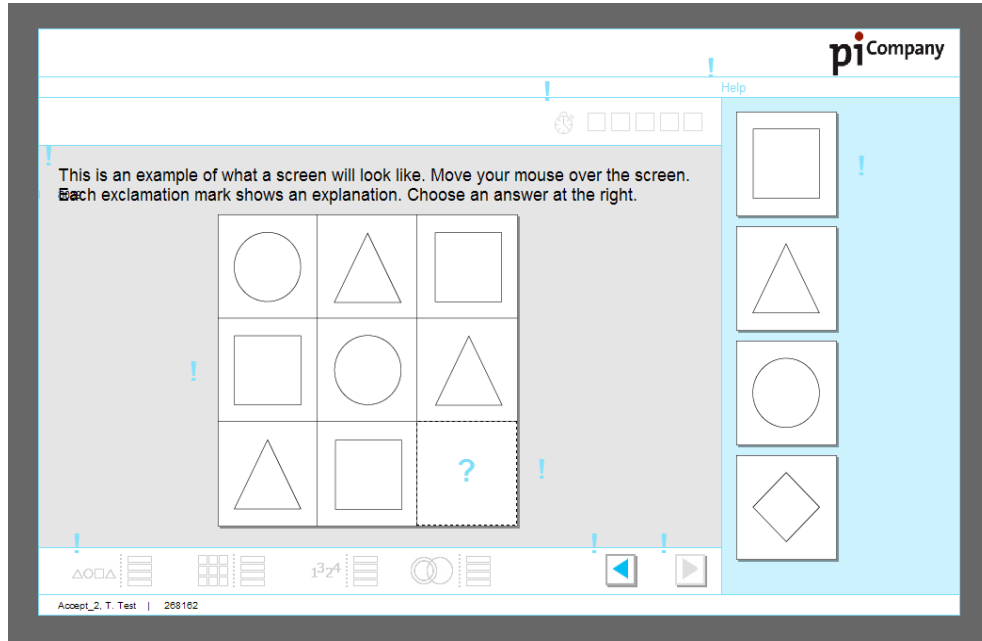
##### **4.2.1 Personal data and background questions**

Before you start with the test, you are asked to indicate whether your personal data are correct. These data refer to your name and birth date.

Next, we ask you to provide some more data on your personal background, for example on your education, work experience and on your and your parents' country of birth. This information is used for research purposes only. Data are processed anonymously and are not in any way used in reporting or in interpreting your test results.

#### 4.2.2 General instruction

Based on the screen below we will explain how the test works.



By moving the mouse across the screen the various parts of the screen are explained.

The **problem** can always be seen in the middle of the screen.

The **answering buttons** are on the right of the screen. By clicking on these buttons you choose an answer. This answer can always be changed (by clicking on a different answer).

At the top of the screen you see a **time bar**.

#### 4.2.3 Instruction per subtest

After exiting the test screen, the instruction of the first subtest will start. This instruction is identical to the instruction presented in this brochure. Also in case of the actual test, you can go through this instruction and sample questions at your own pace. Once you have understood the examples you can start the actual test. The time will not start to run until then.

#### 4.2.4 The actual test

After the instruction, the actual subtest will start. For example: Series of Figures.

The number of questions you receive in each subtest, depends on the answers you give. The computer program offers questions until it has been able to estimate your problem-solving ability based on your answers. For each question you have to give an answer within a limited number of minutes, however, for most people this time is sufficient to be able to answer the question.

Your answers will then be checked and stored by the computer. In the report, your scores will be compared to the scores of people with a comparable level of education.

### **4.3 Protection of personal data**

You may wonder about the protection of the information entered into the computer. Measures have been taken to prevent that your personal and test information, which are stored in the computer, can be accessed by anyone without authorization. Your personal data are stored in a file in such a way that those who are not meant to have this information have no access to it. This way your personal details are protected.

## **5. How to prepare?**

### **5.1. The basis for reliable test conditions**

What is most important is that you are fit and relaxed. Should you not feel well on the day of the test, please tell the test assistant this beforehand (by telephone, if necessary). As an alternative (if possible) you can be tested on another day. If you do the test, the results will be valid.

If you are dyslexic or if you expect that other language problems may influence the test results, please tell the test assistant this beforehand (by telephone, if necessary).

When the test is administered you will be given an explanation about how you are supposed to handle the different parts. We strongly advise you to read through this brochure beforehand, so that you know what you will be dealing with and you can concentrate fully on the questions.

### **5.2 Instruction and sample questions**

For each subtest, you receive an instruction below. You can practice with sample questions, so that you will know what the particular subtest is all about before you start taking the actual subtest.

The sample questions will give you a good impression of what the particular subtest is all about; however, questions in the actual test and the sample questions can differ in difficulty.

Note: You can also practice on PiCompany's website: [www.picompany.nl](http://www.picompany.nl). Click on 'English', on the link 'Connector Ability' and 'Practice the Connector Ability'. Note: the test report gives an indication of your general intelligence. The purpose however of this test is to practice, by answering questions that are comparable to the actual test questions.

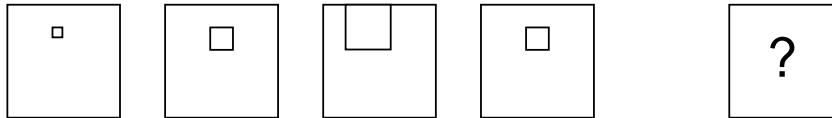


### 5.2.1 Series of Figures

You are shown four boxes. Each box contains a figure.

The question is: **What should the fifth box contain?**

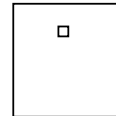
*An example:*



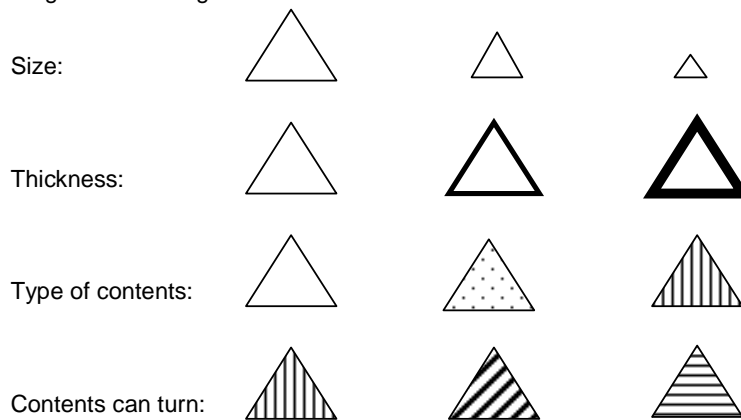
Something changes each time. From one box to the next. Look closely at what changes. And continue that change.

The square first increases in size. And then it becomes smaller. The square in the last box should therefore be smaller again.

This is the figure that should replace the question mark:



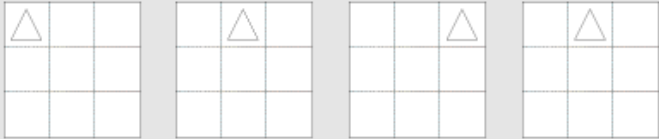
A figure can change:



Each box actually consists of nine (invisible) boxes. Or: nine places for a figure.

A figure can go to a different place in the box, as follows:

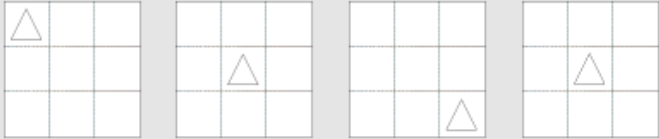
- From left to right  
(or from right to left);



- From top to bottom  
(or from bottom to top);



- Diagonally.

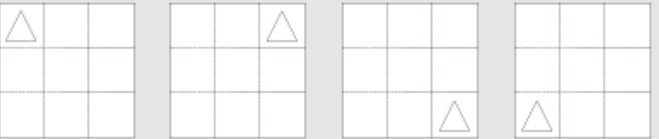


A figure can also turn like the hands of a clock (clockwise or anti-clockwise):

- Moving 1 place each time;



- Moving 2 places each time;



- Moving 3 places each time.



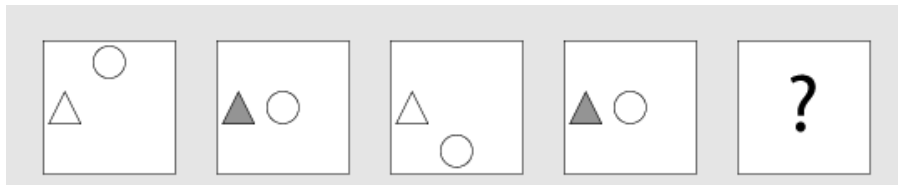
The row or column the figure is in can also determine how the figure changes.. *For example:*

- All the figures in a particular column (left, middle, right) change in the same way (become dark or light);
  
- All the figures in a particular row (top, middle, bottom) change in the same way (become dark or light).



Several things can change in one figure (for example: contents and place). And there are often several figures in one box. Each figure then goes through its own change.

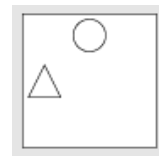
*An example:*



The triangle stays where it is and changes its contents.  
 The triangle in the last figure should therefore be empty again.

The circle moves up and down in the middle.  
 The circle in the last figure should therefore go up further.

The question mark should therefore be replaced by this figure:

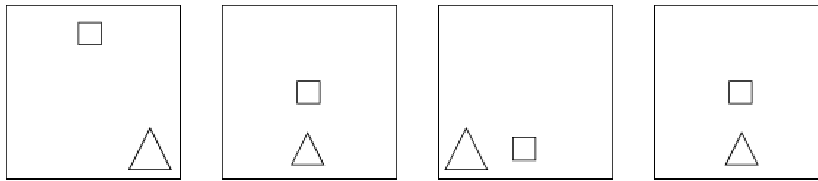


### Sample questions

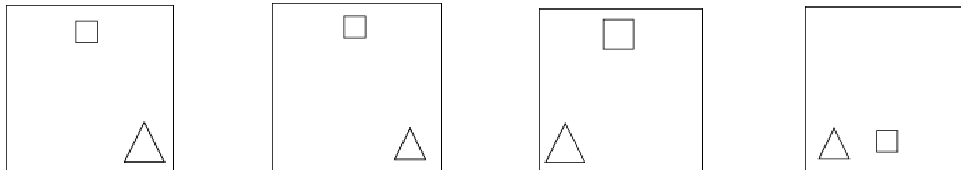
You will now be presented with three sample questions. The answer on each sample question is presented on the next page.

#### Sample question 1:

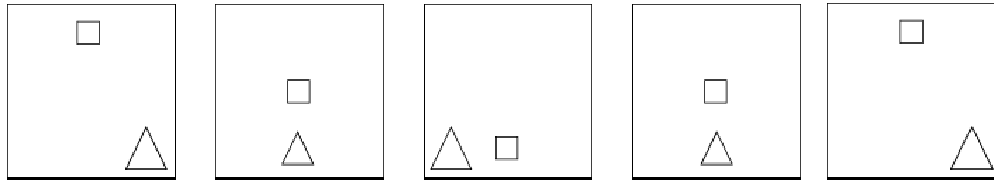
Which figure most logically continues this series?



Choose one of the answers below:



The right answer is:



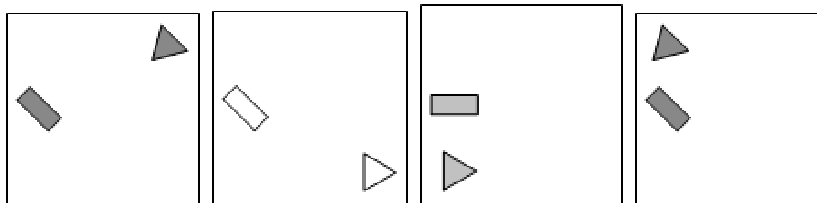
Imagine that each square is divided into nine boxes.

Look from left to right. Form one square to the next. This is what changes:

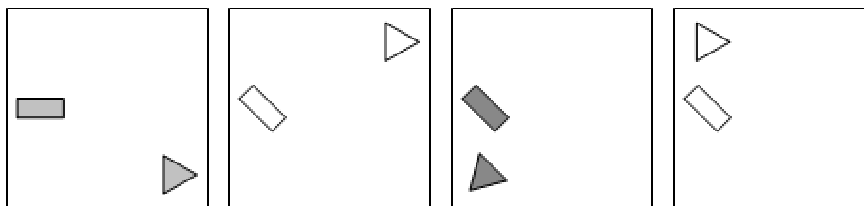
- The square first goes down and then back up again (in the middle column).
- The triangle moves from the right to the left, and then to the right again (bottom row).
- And the size of the triangle changes (first becomes smaller and then larger again).
  
- So the last square in this series of figures moves up one step further and is now in the top row.
- The triangle moves one more step to the right and is now in the bottom right corner. And the triangle becomes larger again.

Sample question 2:

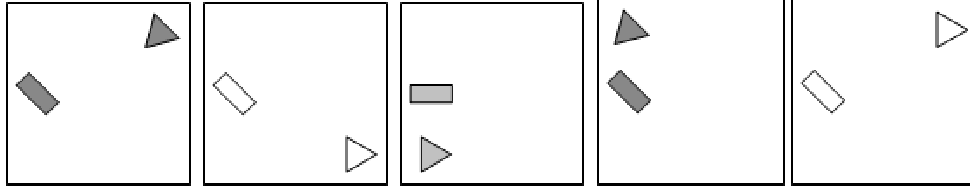
Which figure most logically continues this series?



Choose one of the answers below:



The right answer is:



Imagine that every box is divided into nine sections. This is what changes:

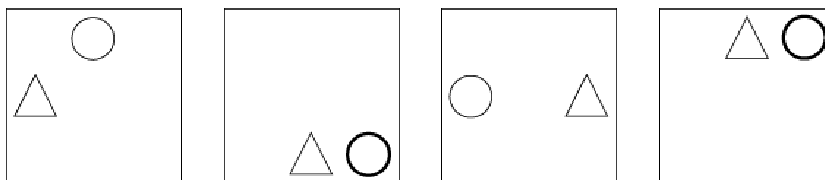
The triangle keeps moving two places, clockwise. In the fourth box the triangle is in the top left corner and it will move to the top right corner.

Two times the rectangle takes a diagonal position, followed by a horizontal position. In the fourth box, the rectangle again takes a diagonal position and now it stays like that.

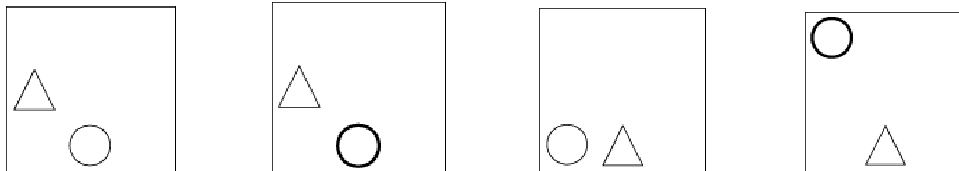
The contents of both the triangle and the rectangle change: very dark-empty-a little bit darker-very dark. So after becoming darker, these shapes turn empty again.

Sample question 3:

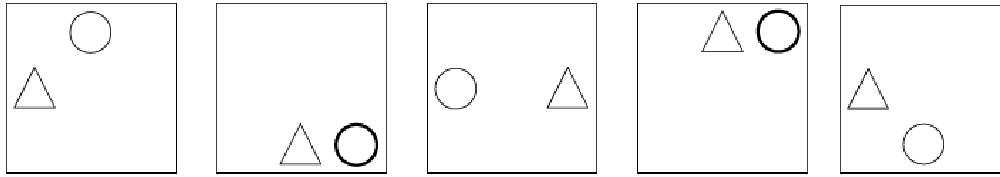
Which figure most logically continues this series?



Choose one of the answers below:



The right answer is:



Again imagine that each square is divided into nine boxes.

Now look from left to right. From one square to the next. This is what changes:

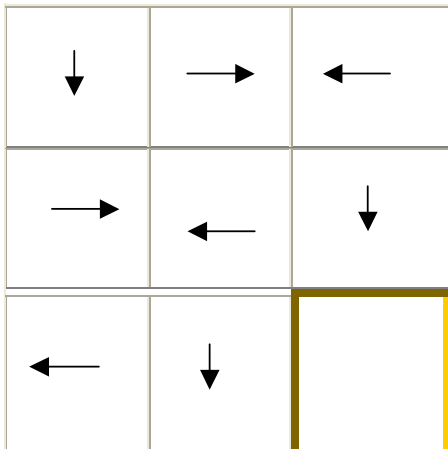
- The triangle moves two places each time, anti-clockwise.
- The circle moves three places each time, clockwise.
- And the thickness of the circle changes (first it becomes thicker and then thinner again).
- So the last triangle moves two places further anti-clockwise, and is now in the left column, middle row.
- The circle moves three steps further clockwise, and is now in the middle column, bottom row.
- And the (thick) circle now becomes thin again.

## 5.2.2 Matrices

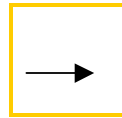
You are shown a large square. The square is divided into nine boxes. There is a figure in each box. Only the bottom right-hand box is empty.

The question is: **What should be in the bottom right-hand box?**

An example:



The answer (bottom right) is:



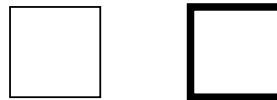
The top and middle rows contain:

- an arrow pointing down
- an arrow pointing to the right
- an arrow pointing to the left

**The bottom row does not yet contain an arrow that points to the right.**

The figure can also change. Changes can take place in:

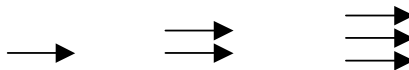
Thickness:



Type of contents:



Number:



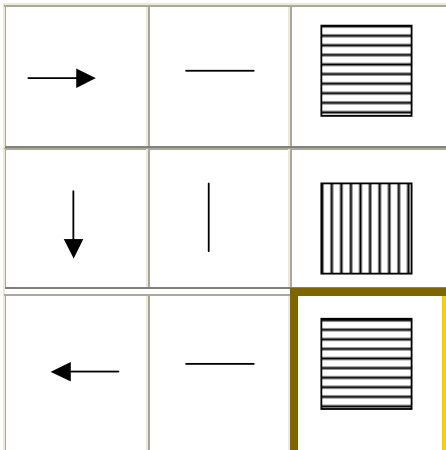


The figure can also turn.

Turning with the same angle each time.

*An example*

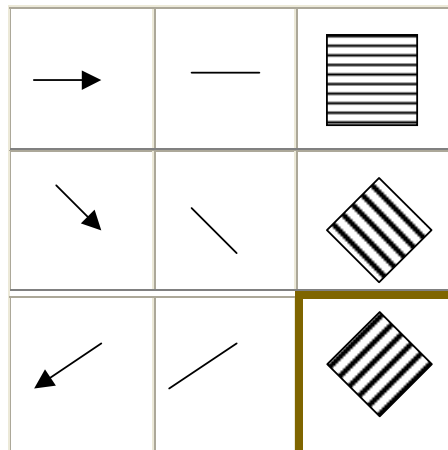
*(the figure keeps turning + 90 degrees, from row 1 to 2 and from row 2 to 3):*



Turning with a larger or smaller angle.

*An example*

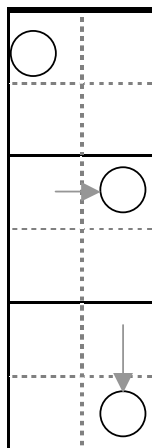
*(the figure turns + 45 degrees from row 1 to 2, the figure turns + 90 degrees from row 2 to 3):*



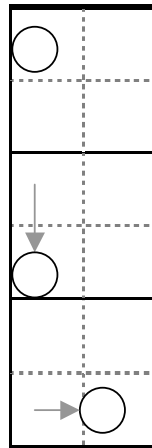
Each box actually consists of four (invisible) boxes. Or: four places for a figure.

A figure can move to a different place in the box, as follows:

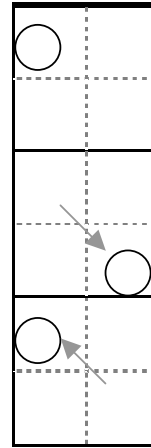
By moving 1 place  
(clockwise):



By moving -1 place,  
(anti-clockwise):

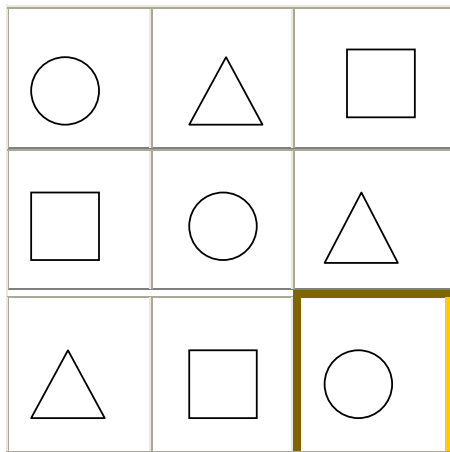


By moving 2 places:

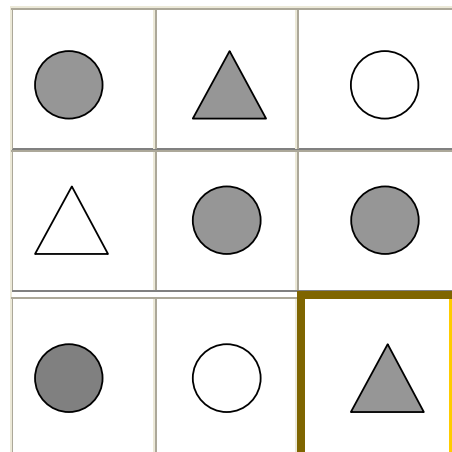


A figure and/or a change can happen once or more times in each row.

*An example of once in each row  
(each row contains 1 circle, 1 triangle and 1 square):*



*An example of twice in one row  
(each row contains 2 circles and  
each row contains 2 dark figures):*



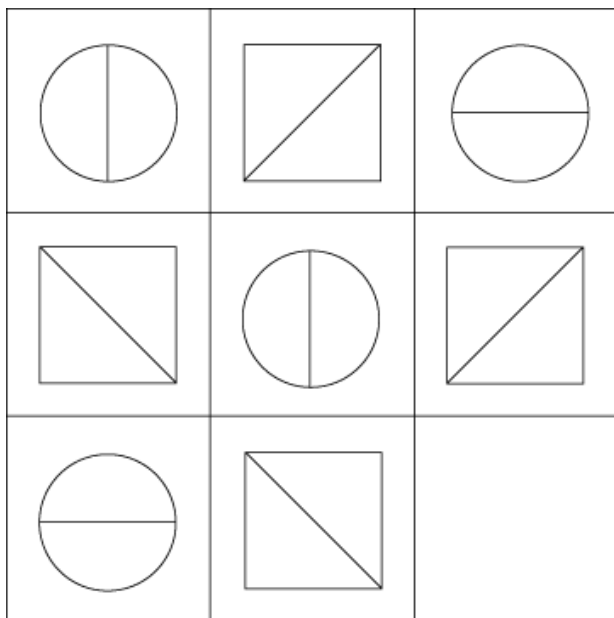
Several things can change in one figure (for example: contents and place). And there are often several figures in one box. Each figure then goes through its own change.

### Sample questions

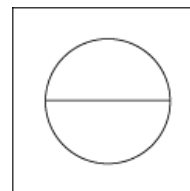
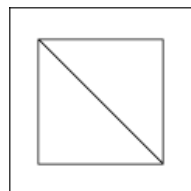
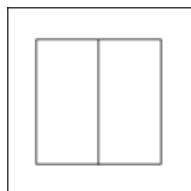
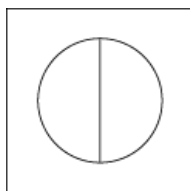
You will now be presented with three sample questions. The answer on each sample question is presented on the next page.

#### Sample question 1:

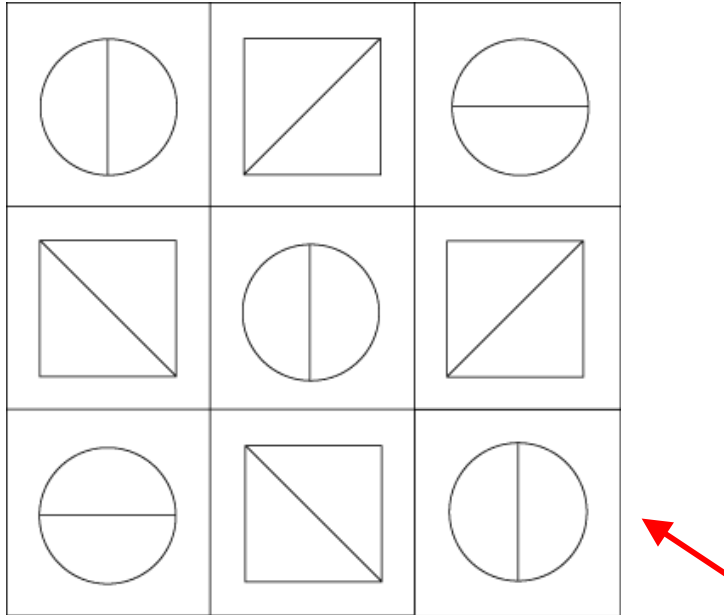
Which figure most logically continues this matrix (bottom right)?



Choose one of the answers below:



**The right answer is:**











Look from left to right. And from top to bottom:

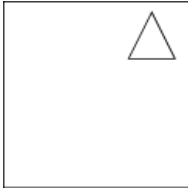
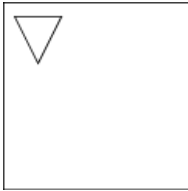
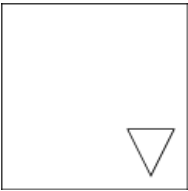
- In each row and in each column one figure occurs only once. And the other figure occurs twice.
- The bottom row contains the same figures as the top row (a circle or a square).
- Only the line inside the top circle or the top square has turned 90 degrees, going from top to bottom.
- At the top right is a circle. So at the bottom right is a circle too.
- The line inside the top circle turns 90 degrees. The line was horizontal and now becomes vertical in the bottom circle.

Sample question 2:

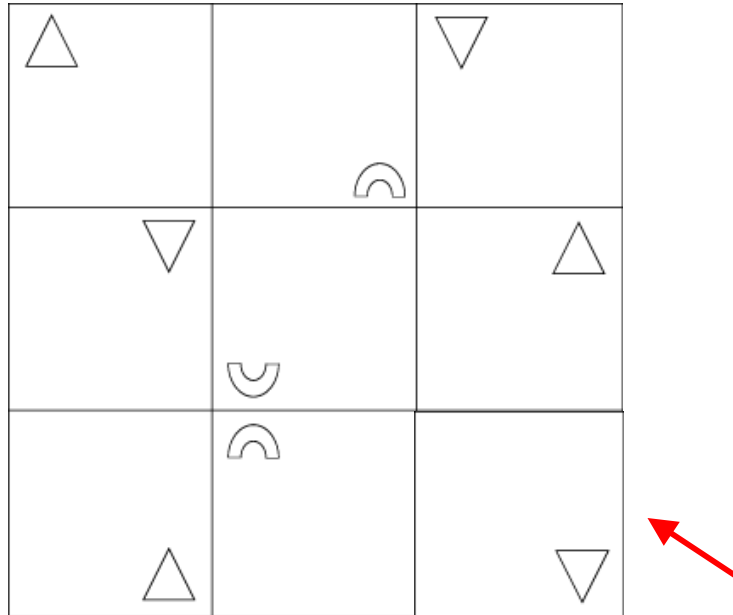
Which figure most logically continues this matrix (bottom right)?

Choose one of the answers below:



The right answer is:



Imagine that each box is divided into four sections. Now look at the left column first, from top to bottom. This is what changes:

- The triangle moves one place to the right each time and then to the left again, clockwise.
- And the triangle turns 180 degrees.









The same thing happens to the figure in the middle column. Again, look from top to bottom.

And also in the right column, the same occurs:

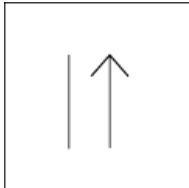
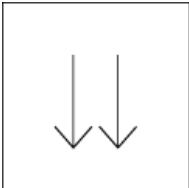
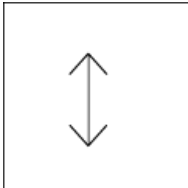
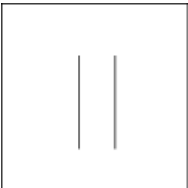
- So the last triangle moves one place further to the right (and therefore goes down). It is now at the bottom right.
- And this triangle, that first pointed upwards, now turns 180 degrees. Now it points down.

Sample question 3:

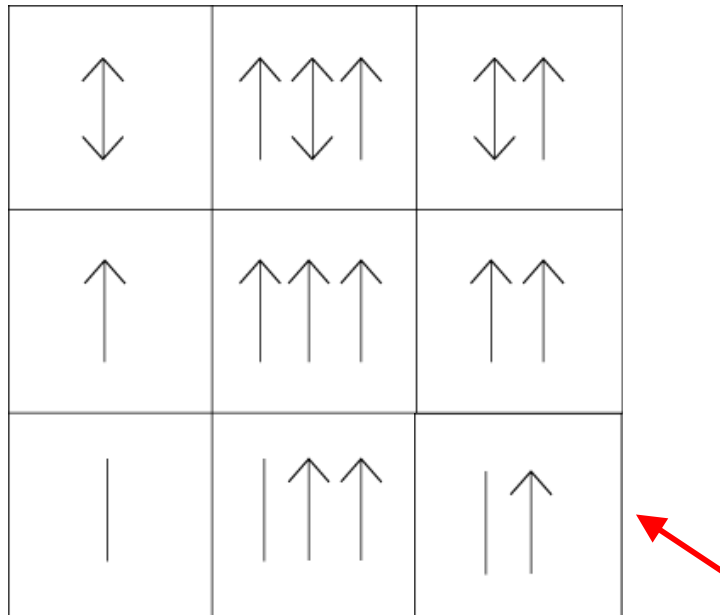
Which figure most logically continues this matrix (bottom right)?

Choose one of the answers below:



The right answer is:



Look from left to right. And from top to bottom:

- Every column (left, right, middle) contains the same number of lines, with or without arrow head.
- And in every row, one arrow head disappears. For this, look from top to bottom.
- The last two lines in the right columns, each have one arrow head. So one of these arrow heads now disappears.
- Therefore, at the bottom right one line with arrow head and one line without arrow head remain.

### 5.2.3 Series of Numbers

You are shown a row of numbers. Something changes each time, in a logical manner. From left to right. We want you to look closely at what changes. And continue that change. **What is the next number?**

An example:            2        4        6        8

In this example two is added to produce the next number.

The next number will be:  $8 + 2 = 10$





One change may take place in a series of numbers. From one number to the next.

An example:            1        4        7        10  
                                  +3        +3        +3

The next number will be:  $10 + 3 = 13$

Sometimes *two* changes take place in a series of numbers.  
Something changes from the first to the third number. And from the second to the fourth number. So two changes have taken place.

An example:            2        200        5        300  
                                  +3        +100        +3

The next number will be:  $5 + 3 = 8$

### Sample questions

You will now be presented with three sample questions. The answer on each sample question is presented on the next page.

#### Sample question 1:

**Which number most logically continues this series?**

4        8        16        32

**Choose one of the answers below:**

60        62        64        96

The right answer is 64.

4      8      16      32      64

Each time the numbers are multiplied by 2 ( $\times 2$ ).

The last number (32) should therefore be multiplied by 2.  $32 \times 2 = 64$ .

4      8      16      32      64  
↘ ↗ ↘ ↗ ↘ ↗ ↘ ↗  
x2 x2 x2 x2

Sample question 2:

Which number most logically continues this series?

10      12      18      28

Choose one of the answers below:

32      38      42      56

The right answer is 42.

10    12    18    28    42

First, 2 is added to the first number.

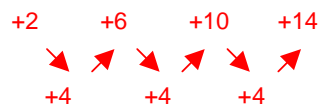
Then 6 is added to the second number.

And after that, 10 is added to the third number.

So the number being added is always increased by 4.

So now  $10 + 4 = 14$  has to be added to the last number (28). And that makes:  $28 + 14 = 42$ .

10    12    18    28    42



Sample question 3:

Which number most logically continues this series?

55    49    43    37

Choose one of the answers below:

29    31    32    33

The right answer is 31.

55    49    43    37    **31**

6 is subtracted from a number each time (-6).

So from the last number (37) again 6 has to be subtracted.  $37 - 6 = 31$ .

55    49    43    37    **31**  
↘ ↗ ↘ ↗ ↘ ↗ ↘ ↗  
-6    -6    -6    -6

#### 5.2.4 Diagrams

You are shown three words. Imagine that each word is in the form of a circle. The question is:

**Which three circles best represent the relationship between the words?**

*An example:*                    **animal, bird, parrot**

A bird is a type of animal. So bird belongs in the circle that represents animal. A parrot is a bird (and also an animal). So parrot belongs inside the two other circles.

These circles best represent the relationship between the three words:



The outer circle represents: animal.

The second circle represents: bird.

The inner circle represents: parrot.

The space between the circles means something too:

*Example:*                    **animal, bird, parrot**

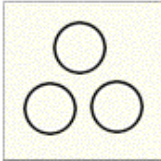
The outer circle represents: animal. The middle circle represents: bird. The inner circle represents: parrot.



But there are also other species (not just birds) that are animals. The space between the outer and middle circles represents: other species. For example: fish. And there are also other birds (apart from parrots) that are birds. The space between the middle and the inner circles represents: other birds. For example: swan.

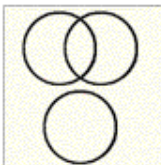
Possible relationships are:

*For example:* **tree, fish, jacket**



Tree, fish and jacket are completely different things. The circles are completely separate.

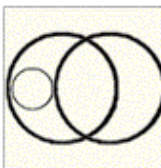
*For example:* **dog, black, apple**



Some (but not all) dogs are black. There are other black things apart from this dog. These two circles overlap. An apple is something quite different. An apple is never a dog and an apple is never black. The circle for apple is separate from the other two circles.

Other relationships are possible too. Here are two more examples:

*For example:* **party dress, party clothes, trousers**



A party dress is party clothes. The circle for party dress is inside the circle for party clothes. And some (but not all) trousers can be party clothes. That is why the circle for trousers partly overlaps the circle for party clothes. The circle for trousers does not overlap the circle for party dress (which is in the circle for party clothes). Because: trousers are never a party dress.

*For example:* **trousers, green, wool**



Some (but not all) trousers are green. And some (but not all) trousers are made of wool (the middle part). It doesn't matter which word comes first or last. Neither does it matter how big or small the circles are.

It doesn't matter which word comes first or last. Neither does it matter how big or small the circles are.

### Sample questions

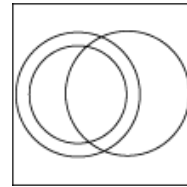
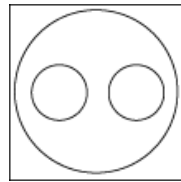
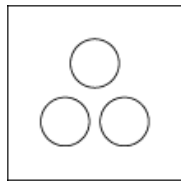
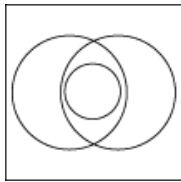
You will now be presented with three sample questions. The answer on each sample question is presented on the next page.

#### Sample question 1:

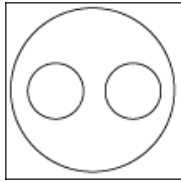
**Which figure best describes the relationship between these three concepts?**

**animal    fish    bird**

**Choose one of the answers below:**



The right answer is:



**animal    fish    bird**

A fish and a bird are both animals.

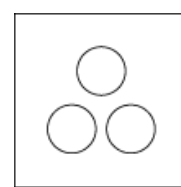
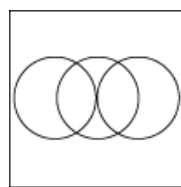
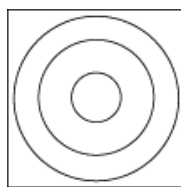
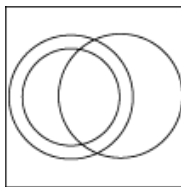
Fish and bird therefore are both within the big circle which stands for animals.

Sample question 2:

Which figure best describes the relationship between these three concepts?

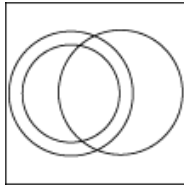
**clothes    jacket    cotton**

Choose one of the answers below:





The right answer is:



**clothes      jacket      cotton**

A jacket is a type of clothes. That is why the circle representing jacket is inside the left-hand circle that represents clothes.

Some (but not all) clothes are made of cotton. And some (but not all) jackets are made of cotton.

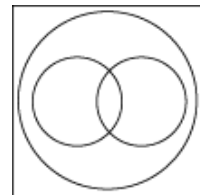
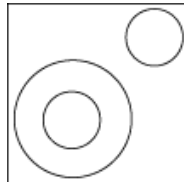
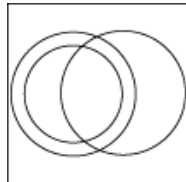
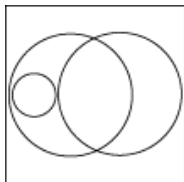
That is why the right-hand circle (that represents cotton) partly overlaps the circles that represent clothes and jacket.

Sample question 3:

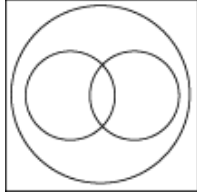
Which figure best describes the relationship between these three concepts?

**transport      freight transport      boat**

Choose one of the answers below:



**The right answer is:**



**transport    freight transport    boat**

Freight transport and boat are both transport. That is why freight transport and boat both belong in the circle that represents transport.

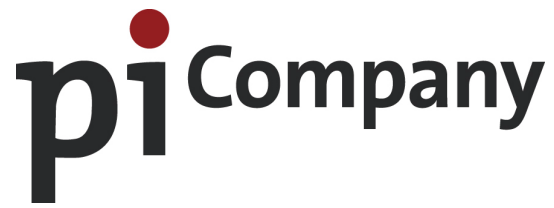
And sometimes a boat is also freight transport (for example: a freight boat). And there are also other sorts of transport that do not include transport by boat (for example: transport by car).

That is why the circle that represents freight transport partly overlaps the circle that represents boat.

## **Appendix B**

### **Best Practice Guidelines**





*Best Practice Guidelines*

P E O P L E I M P R O V E  
P E R F O R M A N C E

# Table of contents

Best Practice for applying Connector Ability 1.1.....3  
Connector Ability and dyslexia.....5  
Use of T scores.....9

# Best Practice for applying Connector Ability 1.1

- Does a person have the required intellectual ability for the job or potential job?
- Is a person capable of taking a course at a particular level?

When you are faced with such questions, Connector Ability can offer you an insight into a person's cognitive abilities, or intelligence, and thus an insight into how easily that person will be able to solve problems and become familiar with new knowledge.

The aim of this document is to assist you in applying this instrument appropriately and professionally. To this end, this document offers answers to the following questions:

- What can I expect from the application of Connector Ability?
- When should I use Connector Ability?
- What should I take into account when applying Connector Ability?

## 1. What can I expect from the application of Connector Ability?

Applying Connector Ability will give you:

- An insight into candidates' general level of intelligence.
- An insight into whether they have strong and less strong abilities in the sub-areas and where they lie.
- A measurement of their cognitive abilities free of cultural bias and independent of language skills or academic knowledge.
- An efficient, focused measurement of their cognitive abilities, since candidates are given very specifically determined questions that depend on their previous answers ('adaptive' questioning).

## 2. When should I use Connector Ability?

### *a. For selecting staff:*

- When an insight into cognitive abilities, or intellectual ability is important for a person to be able to perform well. Research has shown that insight into cognitive abilities is the best predictor of how a person will perform in the work situation.
- When an estimate of potential comes into play, an estimate of the extent to which someone will develop in a position, it is especially important to gain an insight into how easily a person will become familiar with new knowledge.

*b. In combination with other selection instruments from the Connector package:*

When an estimate of a broader collection of competencies is required besides a measurement of cognitive abilities. Connector Big Five Personality sheds light on the ability to develop competencies from a personality perspective; a STAR interview is essential for checking the extent to which competencies are present in practice.

An example of such a selection process:

Connector Ability + Connector Big Five Personality + STAR interview.

*c. In development or career orientation processes*

When it is important to be able to estimate:

- How easily a person will be able to master new knowledge in the work situation
- How easily a person will be able to complete a particular course
- Whether a person has the required intellectual level for a course.

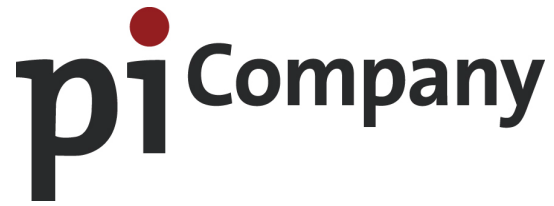
### **3. What should I take account of when applying Connector Ability?**

The following aspects are important for the appropriate use of tests in general, but for a cognitive abilities test in particular, in which it is extremely important that people are able to work with as much concentration as possible in the time allotted:

- Good preparation. Candidates should be sent the brochure in advance to enable them to go through it and see what sort of questions they can expect.
- The conditions under which candidates fill in their answers during the test must be optimum: the candidate should feel fit and well and not be disturbed.
- The test assistant should be well prepared to answer any questions and should monitor the test conditions.  
See 'Manual Test assistant' and 'Candidate brochure'.
- It is also important to take account of diversity in the selection procedure. PiCompany can advise your organisation on how to set up a selection process in which gender, age and cultural background have no effect on the outcome of the process.
- Dyslexia will affect the test results of Connector Ability to a lesser extent than is the case with many traditional, language-based intelligence tests. Nonetheless, it is still a good idea to know in advance whether candidates are actually dyslexic and to take account of this when administering the test.

See Best Practice 'Connector Ability and dyslexia'.





# *Connector Ability and dyslexia*

*Best Practice guidelines*

P E O P L E I M P R O V E  
P E R F O R M A N C E

## **Appropriate test usage for dyslectics**

### **Non-language-based tests reduce the effect of dyslexia**

Dyslexia only has a limited effect on the results of Connector Ability. This is due to the following features of the test:

(1) The Connector Ability test has been developed free of reliance on language: instead of measuring language skills, the test measures general intelligence. Questions in the 'Diagrams' section do use verbal concepts to ask for an indication of the relationships between these concepts. On the other hand, only generally known, simple, non-compound words are used. Compared with tests in which whole passages of text have to be read and analysed, this test hardly calls for any reading skill and there is no question of complicated words, either in terms of meaning or spelling.

(2) Besides this, Connector Ability is a test that measures *whether* you are able to answer an item within a standard length of time and not exactly how much time you needed to do so. The time people are given in which to answer a question is the time the vast majority of people need in which to answer that sort of question. How quickly someone answers the question within that time does not play a significant part in the final score. From this point of view, the subtest "Diagrams" is particularly relevant for people with "dyslexia". But although these people may need a little more time to read the three words, the available time for each item in this section is more than enough.

For these reasons, dyslexia affects the Connector Ability test results to a lesser extent than is the case in many traditional, language-based intelligence tests. Nonetheless, it is a good idea to know in advance whether candidates are actually dyslectic and to take this into account when administering the test. A few 'best practice' guidelines are set out below.

### **What should you do if a candidate tells you he/she is dyslectic prior to taking Connector Ability?**

#### **1. How severe is the dyslexia?**

If possible, try to find out, before administering the test, whether the person is dyslectic. If so, check whether an official diagnosis has ever been made by a recognised body (see appendix).

***If the diagnosis of dyslexia is not official:***

If the diagnosis has been made informally, 'dyslexia' can just mean being less good at verbal tasks. If, for example, the diagnosis was made on the basis of poor reading performance and/or word recognition, dyslexia may only be another word for "relatively low verbal abilities". And that is exactly what you want to test: Connector Ability includes this in its prediction of intelligence. The application of Connector Ability is therefore relevant.

***If there is an official diagnosis of dyslexia:***

Even when the diagnosis of dyslexia is official, we still advise administering the Connector Ability test, but to take special care when giving instructions and when these are being interpreted (see the next sections). Connector Ability cannot specifically correct for dyslexia, but because general intelligence, the G factor, is measured from several sub-areas and therefore in several different ways, its effect is probably minor.

## **2. Administering Connector Ability**

We do not consider it desirable to correct for dyslexia when administering the tests for the following reasons:

- Any slower speed of reading comprehension will also apply in practice. That is why a procedure in which this 'handicap' is taken into account will lead to the actual performance in practice not being correctly predicted.
- Giving a person more time to complete an intelligence test actually means that intelligence is no longer being measured. After all, intelligence is defined as 'cognitive performance in standard tasks within a *standard length of time*'.
- To be able to estimate someone's level of intelligence, it is essential to compare them with other people (norm groups). This becomes impossible if the time allotted is not standard.

## **3. Instructions to the participant**

It is important to put the person taking the test at their ease and one way is to give them a brief explanation of how dyslexia is dealt with. Emphasise the following aspects:

- Only the 'Diagrams' section of Connector Ability is based on language, while other sections are based on numbers or figures. This means that the ultimate general intelligence score is based on several tests, on several measurements of intelligence. That is why it is still able to produce a picture of a person's cognitive abilities.
- All the tests, including those that measure language, must be done in a limited time-frame. This is necessary to be able to compare the results with the results other people attain in the same tests.
- Very important: emphasise that the participant should take as much time as he/she wants to read the instructions so that he/she knows exactly what is expected of him/her before the test starts.

#### **4. Interpreting the results**

The results of the verbal subtest will reveal how much any dyslexia affects that person's verbal abilities. That, too, is relevant input for selection and development processes.

After that, it is important to find out what effect dyslexia will have on the way that person functions in practice. Ask specific, preferably STAR/criterion-based questions, to try to get an idea of how dyslexia affects the person's work and learning, how they deal with this in practice and what the result of their efforts is. For example: ask for recent, concrete examples of work situations in which the person was confronted by their dyslexia, how they dealt with it and what the result was. But also: what are that person's educational results (marks, length of time spent in forms of education), how did they manage to attain these results/diplomas?

### **Appendix: Information about dyslexia and referral**

More Information about what dyslexia is:

**British Dyslexia Association:**

[www.bdadyslexia.org.uk](http://www.bdadyslexia.org.uk)

**The International Dyslexia Association:**

[www.interdys.org](http://www.interdys.org)

# *Use of T scores*

*Best Practice guidelines*

## Why T scores?

Various data can be used to interpret test results. The most commonly used data are:

- Percentile or decile scores.
- T scores.

### **Advantages and disadvantages of using percentile scores**

Percentile scores indicate how great a percentage of the norm group has a lower score for the test than the candidate in question. For example: a percentile score of 80, which is the same as a decile score of 8, means that 80% of the norm group scored lower and 20% of the norm group scored higher.

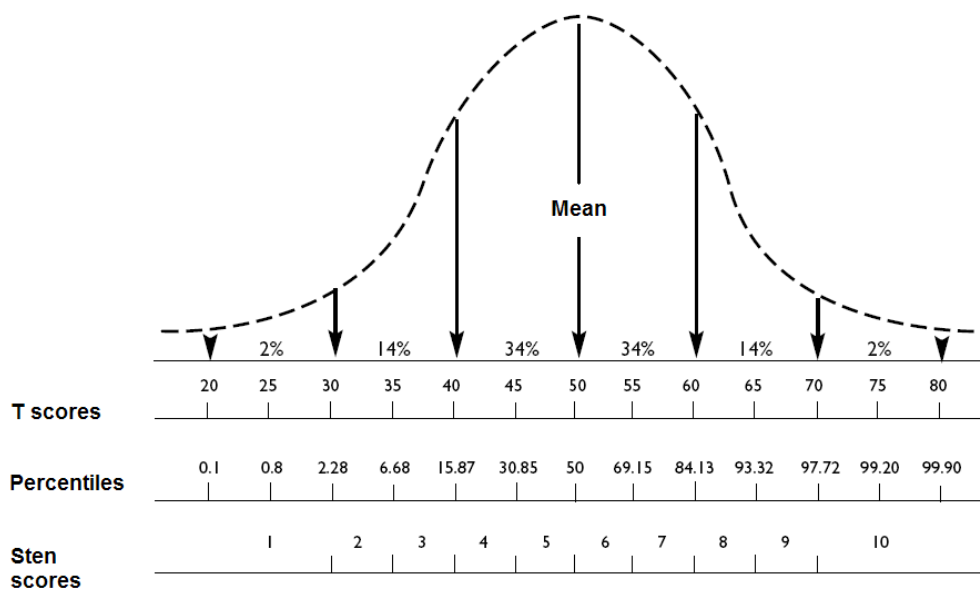
The advantage of the use of percentile scores is that users find them transparent and easy to understand. From a psychometric point of view, however, there are a number of disadvantages. Percentiles only indicate an order of merit: how many people score higher, how many people score lower. This level of measuring is called ordinal. At the same time, going by a percentile score means that the differences in scores that lie close to the average are magnified. For example: someone who gets 13 questions right (a 'rough score' of 13) and, when compared to the specific norm group, comes out at the 27<sup>th</sup> percentile, could come out at the 50<sup>th</sup> percentile with a rough score of 17. Conversely, the differences in scores that are far away from the average are diminished, of all things. For example: a rough score of 32 lies on the 97<sup>th</sup> percentile, while a score of 36 lies on the 99<sup>th</sup> percentile. This produces a distorted picture.

### **Most important reason not to base selection decisions on percentile scores:**

People who perform above average are too quickly thrown onto a heap (too little attention is paid to differences), whereas there is too much of a distinction made – unjustifiably - between people who perform about average (differences are made too much of).

### **Why use standard scores like the T score?**

Standard scores show whether a candidate has scored above or below the norm group average and, at the same time, the extent to which the candidate's score deviates from the average norm group score. A standard score therefore provides more information about the candidate's results compared to the norm group and, furthermore, the distortion seen with percentile scores does not occur (interval scale). Because of this, T scores can be used immediately to compare people's scores with those of others. For example: person A's score lies twice as far from the average as that of person B.



T scores are standardised standard scores chosen such that the average score lies at 50 and a standard deviation from the average is 10. That makes calculating easy, partly because the scale can easily be divided up into 10 equal blocks (known as 'Sten scores'). This is depicted in the figure above, which also shows how percentile scores are concentrated around the average.

For example: when an organisation puts the 'bar' at a Sten score of 5 (a T score of 50), this means that only people who score at least average compared with the norm group of, for instance, academics, may go on to the next phase. In percentiles this would mean that the organisation wants to select the best 50%. But the bar can of course also be set lower, when desirable in view of the job (high risk of failure) and when the supply of applicants allows for this.

**To summarise:** a T score gives a clearer picture of the level at which a person performs and does not distort, as does happen when percentile or decile scores are used. You will not reject people undeservedly or take the wrong people on because you will be estimating more precisely and more correctly how a person will perform compared to a norm group. T scores are moreover easier to convert to a 10-point scale and T scores can be readily compared to each other.

That is why the T score has a central role in the reports on the most recent Reflector Big Five Personality test (2.0), Connector C (3.1.), and Connector Ability. In addition, Connector C also displays percentile scores showing which percentage of the norm group scored higher or lower than the person in question.





## **Appendix C**

### **FAQ Top 10**





*Frequently Asked Questions Top 10*

P E O P L E I M P R O V E P E R F O R M A N C E

## Connector Ability

### Frequently Asked Questions Top 10

1. The average score on the subtests differs from the G-factor score. Is the calculation correct? .....	3
2a. With some tests it is possible to distinguish between someone's numeric, abstract or verbal reasoning ability. Is this also possible with this test? .....	3
2b. How can PiCompany assess someone's language ability with this adaptive test? .....	3
3. Is it allowed to practise before taking the test and, if so, how to practice? .....	3
4. How long does the test take? .....	4
5. How to deal with dyslexia? .....	4
6. How do I determine what a candidate should score? .....	4
7. When a candidate does not know the answer to a question, is it better for a candidate to gamble or to give no answer at all? What is the best strategy for answering? .....	5
8. What if a candidate already started the test but mentions to feel unwell. Is it allowed to let the candidate take the test again? .....	5
9. What is the difference between G-factor and IQ? .....	5
10. Does the test meet the (international) quality guidelines for tests? .....	5

**1. The average score on the subtests differs from the G-factor score. Is the calculation correct?**

The calculation is correct, because the G-score is not calculated based on averaging the subtest scores. The G-score is computed based on a special calculation model.

The G-score can be higher as well as lower than the mean of the subtests. The reason is that the calculation is based on all answers on all subtests at the same time, taking into account the specific contribution of each separate question on the G-factor.

**2a. With some tests it is possible to distinguish between someone's numeric, abstract or verbal reasoning ability. Is this also possible with this test?**

It is important to distinguish between the talent or natural ability to learn certain things (for example, to learn to calculate) versus the skills to, for example, calculate. In predicting success in whatever job, the general talent to quickly learn new cognitive tasks is of primary importance.

There are ability tests available that measure specific verbal or numeric reasoning ability. These tests do not directly measure the talent to learn, but a combination of actual skills and talent. Based on this type of test, it is not possible to assess which part is talent and which part is skill.

**2b. How can PiCompany assess someone's language ability with this adaptive test?**

It is not possible to assess language ability with this test and this test is also not meant to assess this. Based on someone's general intelligence (talent), one can estimate the level of language ability a person can reach when adequate education/training is available.

**3. Is it allowed to practise before taking the test and, if so, how to practise?**

To make sure that before taking the test candidates understand what is expected, it is even essential that candidates practise before taking the test. This is why the test includes practise questions. Also when guiding a candidate, it is important that – regardless of cultural background- a candidate knows exactly what is expected from him/her. Before the actual test starts, this is also checked explicitly: when candidates repeatedly do not succeed in answering the practise questions correct, they can repeat the practise questions.

We advise to sent candidates a candidate brochure before they take the Connector Ability test and/or let candidates take the practise test at the PiCompany website ([www.picompany.nl](http://www.picompany.nl), English, Connector Ability, Practise the Connector Ability). This practise test is available in English and in Dutch for all educational levels (lower vocational level– VMBO; mid-level vocational education – MBO; Bachelors (BA); Masters (MA).

#### **4. How long does the test take?**

In the subtests Series of Figures, Series of Numbers, and Matrices, a candidate has 90 seconds (so 1,5 minutes) to answer a question. In the subtest Diagrams, the time per question is 45 seconds. Our research showed that about  $\frac{3}{4}$  of the norm group succeeds in answering the question within that timeframe. Most candidates taken about 45 to 60 minutes to take the test (including the instruction).

#### **5. How to deal with dyslexia?**

The extent to which dyslexia can affect the results on the Connector Ability is relatively small because Connector Ability does not measure language skills but general intelligence. In the subtest 'Diagrams', only generally well-known, simple words that are not composed, are used. Compared to tests in which a lot of text needs to be read and analysed, this subtest hardly relies on reading skills, and words are not complex, neither in meaning nor in spelling.

Besides, the Connector Ability is a test that measures per item *if* someone can give the correct answer within the standard time, and not the amount of time necessary to do so. The time available to answer a question, is the mean time the gross majority of people need to answer a similar question. Therefore, how fast someone answers the question within this timeframe, does not have a significant impact on the final score. For someone with dyslexia, particularly the subtest Diagrams is relevant for taking into account. But even in the case that someone with dyslexia would need a little more time to read these three words carefully, still, in general, the available time per item will be more than enough.

Nevertheless, it is still important to know before starting the test, whether candidates are really dyslectic and to take this into account when using the test. It is important to pay attention to: To what extent is dyslexia confirmed? Is this really a case of dyslexia, then give candidates ample time for the instruction (take this into account when planning the test). With respect to the test results/interpretation: it is not very likely that dyslexia will affect the results of Connector Ability (as explained above). Also, the G-factor is being estimated based on 4 subtests.

Even though it is important to know beforehand if a candidate is really dyslectic, the possibly lower speed in reading comprehension will also affect actual performance. Therefore, measuring this remains relevant, but in the interpretation of the general intelligence factor this possible limitation should be taken into account. In case of dyslexia, the remaining subtest of the Connector Ability – Series of Figures, Series of Numbers, and Matrices- will be more pure indicators of the G-factor.

#### **6. How do I determine what a candidate should score?**

This is elaborated in the certification training. This depends on the level needed for successful performing in the particular job and on what decision errors you are prepared to accept. The higher the score, the lesser the number of people that will be unjustly accepted. Therefore,

determining the so-called 'cut-off score' means weighing the risks of unjustly accepting versus unjustly rejecting.

**7. When a candidate does not know the answer to a question, is it better for a candidate to gamble or to give no answer at all? What is the best strategy for answering?**

When someone does not fill in an answer, this automatically counts as a wrong answer. When someone gambles, the answer could be correct. Someone will not get 'punished' for giving the wrong answer. Because of the adaptive nature of the test, the total pattern of answers determines whether another question will be offered and if so, which question. Should someone 'unjustly' (or by accident) give a wrong or a correct answer, then the program will, as it were, correct this. So the only result is that someone will be offered more questions until someone's G-factor can be estimated reliably. This does not apply for a traditional test.

In short: When someone really does not know the answer, it is no use to wait until the time for answering the question has passed. A person can just as well gamble.

**8. What if a candidate already started the test but mentions to feel unwell. Is it allowed to let the candidate take the test again?**

If the candidate really thinks that he/she can not continue with the test, then it is better to take the test another time. Better not to take the test on the same day, because it is important that the candidate feels well and is able to concentrate well.

**9. What is the difference between G-factor and IQ?**

There is no real difference. IQ is just a specific measurement scale that refers to the total world population (mean 100, range 15). In general, organisations that wish to use intelligence tests, like Connector Ability, want to compare a person to a more specific norm group, most of the time a group of people with a specific educational level (for example Masters). Then it is not necessary to estimate the IQ-score in addition to the t-score, the score used to compare the score of a person with the norm group scores.

**10. Does the test meet the (international) quality guidelines for tests?**

The test has been developed in line with the guidelines of the European Federation of Psychological Associations (EFPA) and the International Test Commission (ITC). The test will soon be sent to an independent, external association (4TP) which will judge the test based on these guidelines. The Netherlands Institute for Psychologists (NIP) complies with the same guidelines.





## **Appendix D**

### **Example of Test Report**





*Personal report for  
B. Smit*

PEOPLE IMPROVE PERFORMANCE

## ***Personal details of participant:***

<b>Name</b>	B. Smit
<b>Date of birth</b>	12 June 1972
<b>Gender</b>	Man

## ***Test data:***

<b>Test date</b>	17 August 2007
<b>Test number</b>	54321.123456
<b>Norm group</b>	MA
<b>Test language</b>	English

## **Disclaimer**

When interpreting this report, account should be taken of the attributes of the specific instrument. This report and the instrument it refers to may only be used by people whom PiCompany deems to have the appropriate expertise to do so. PiCompany is not liable for the consequences of improper use of this report; this liability lies entirely with the organisation that makes use of the instrument in question.

This report has been generated automatically.

# Connector Ability

---

This is the Connector Ability report. The Connector Ability is a test that measures a person's cognitive abilities.

Cognitive ability refers to how easily and quickly a person is able to solve various types of cognitive problems. This problem-solving ability gives an indication of the ease and the speed with which the candidate will tackle problems in the position being applied for.

## Explanation of the page 'Connector Ability Test Report'

### – G factor

The G factor reflects the participant's general ability to solve problems ('G' stands for 'General Ability'). The G score is based on the scores attained in the four subtests: Series of Figures, Matrices, Series of Numbers and Diagrams.

### – Norm group

The G factor that the participant attains in the test is compared to the scores attained by a norm group. A norm group is a group of people who are comparable to the participant in a certain respect. You will see which norm group the participant's G factor has been compared to under 'norm group'.

### – Subtests

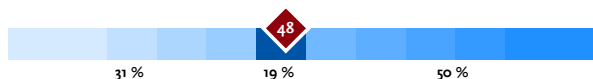
Finally, you will see the scores the participant has attained for each of the four subtests: Series of Figures, Matrices, Series of Numbers and Diagrams.

## Meaning of the scores in 'Connector Ability Test Report'

The G factor and the scores for each of the subtests are represented by so called t-scores. These t-scores are shown above the bar. A t-score of 50 reflects the norm group average. A score that deviates strongly from the average occurs relatively less often. For example, approximately only two percent of the norm group scores above 70. And similarly, about two percent of the norm group scores below 30, while about nineteen percent has a score between 45 and 50.

The bars under the t-score are divided into sections, each representing five t-scores. Percentages for the G factor and each of the four subtests indicate the percentage of people from the norm group who score lower and the percentage of people from the norm group who score higher than the participant. These percentages appears below the bar.

An example

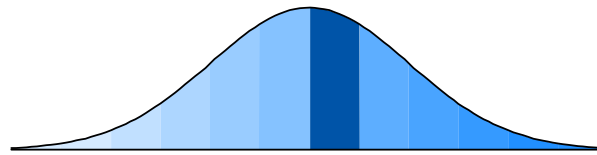


***This score is between 45 and 50. This means that 31 percent of the norm group score lower and 50 percent score higher than the participant. 19 percent have about the same score as the participant.***

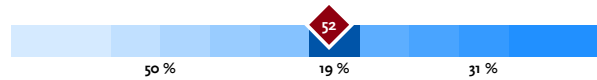
## Confidentiality

Test data is handled with the greatest possible confidentiality. In so doing, PiCompany conforms to the guidelines of the Netherlands Institute of Psychologists (NIP), see: [www.psynip.nl](http://www.psynip.nl), and those of 4TP ([www.4tp.nl](http://www.4tp.nl)).

# Connector Ability Test Report



**G factor - MA**



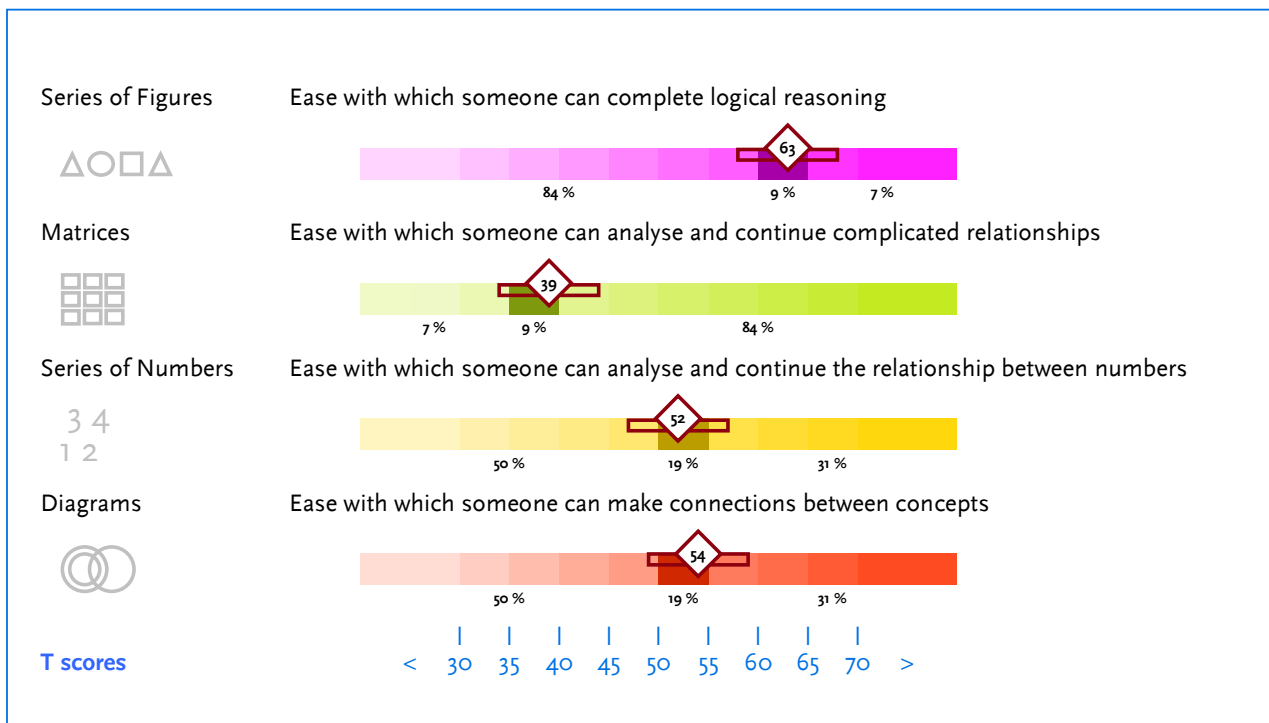
**T scores**



The candidate has scored 52. This score is between 50 and 55. This means that 50 % percent of the people in the norm group with **MA**-education had a lower score and 31 % percent had a higher score than the candidate. 19 % percent scored about the same as the candidate.

## Subtests

The G factor is calculated based on the following four subtests.



The bar shows the margin around the score. In three-quarters of cases, the score will be within this margin if the participant would take the test again.

## **Appendix E**

### **Online Testing Process; an illustrative example**







*Online Testing Process;  
an illustrative example*

P E O P L E I M P R O V E P E R F O R M A N C E

# Online testing process

## An illustrative example

A test case will be described, providing an example of how the adaptive procedure is independently checked for a number of test cases. The results are compared to the output obtained from the system that is used for the adaptive procedure.

A candidate is given the first item randomly drawn from a set of items with a difficulty parameter near zero. The candidate gives an answer, which is correct (score 1) or incorrect (score 0). As no estimation of theta is possible when there is just one item response, the theta value is updated by adding or subtracting one 'step-size'. The next item is selected that is the most informative given the estimate of theta at that point. The procedure continues as described in Section 2.3.4. The process is terminated when one of the stop criteria is met, see also Section 2.3.5.

## Test case example

A table for each subtest is provided, based on Connector Ability 1.1, showing the procedure of adaptive testing. An item count is given, the ID of the item and its respective discrimination ( $\alpha$ ) and difficulty ( $\beta$ ) parameters. The response by the person is given, either 1 (correct) or 0 (incorrect). Next, the theta value is estimated after the given response. The item information (I) can be obtained from Equation 1, using the discrimination parameter and the probability of a correct and incorrect answer given the estimated value of theta before the response is given. According to Equation 3, the corresponding SE can be computed, based on all information available up to that point after having responded to the item.

## Series of Figures

The first item selected is Figitem11 with a mean difficulty parameter ( $\beta = -0.33$ ) and high discriminative power ( $\alpha = 3.10$ ). The item information value, given the initial theta value of 0, forms the basis for item selection. The person gives an incorrect response (0). The theta value now decreases with one step-size, to a value of -0.7. Given this value of theta and the item parameters the standard error is computed.

The next item is selected, of which the item information is largest given the value of theta estimated at this point (see column I). This item, Figitem183, contributes most to the estimation of theta. It is seen that the discrimination parameter of the item is very high (3.01) and the difficulty parameter (-0.83) is situated near the estimated value of theta (-0.7).

**Table A Adaptive procedure for Test case 1 for Series of Figures**

# item	Item-ID	$\alpha$	$\beta$	Response	theta	SE	I
1	Figitem11	3.10	-0.33	0	-0.7	0.75446	1.87384
2	Figitem183	3.01	-0.83	0	-1.4	0.81692	2.17654
3	Figitem657	3.87	-1.28	1	-1.08578	0.40880	3.54071
4	Figitem12	2,98	-1.18	1	-0.92188	0.36082	2.17615
5	Figitem84	2.81	-1.10	1	-0.80922	0.33605	1.86419
6	Figitem132	2.91	-0.66	1	-0.63594	0.32014	2.02934
7	Figitem9	2.57	-0.78	1	-0.53859	0.30865	1.59453
8	Figitem7	2.40	-0.33	1	-0.40969	0.30665	1.35182
9	Figitem265	2.31	-0.47	0	-0.50859	0.27825	1.32578
10	Figitem8	2.32	0.56	1	-0.42656	0.27269	1.34179

Again, an incorrect response is given (0), thus the theta estimate decreases with another step-size to -1.4. The corresponding item information is computed, given the theta estimate at this point. The standard error is based on all item responses given up to this point. The SE now increases, though one would expect it to decrease with the administration of more items. However, the first item (with mean difficulty level) does not contribute much information, as the theta estimate is far from the difficulty parameter.

A third item is chosen with a difficulty parameter close to -1.4 and high discriminative power ( $\alpha = 3.87$ ). Now, a correct response is given. As there are both correct and incorrect answers, a theta value can be estimated according to MLE (see Section 2.3.4). The standard error of estimation has decreased drastically. The selection of new items continues, according to the maximum information given the theta value estimated at that point.

The standard error of estimation decreases further. Stop criteria are when a maximum number of 15 items is administered, or when an SE value below 0.54 is reached. This latter criterion is met after the administration of just 3 items. However, the minimum number of items to be administered was set at ten. Once ten items are administered, this subtest is terminated. The theta value of -0.426 is estimated with a reliability of 0.93 (see Section 4.1.1).

This procedure is followed in order to check all computations in the program underlying the adaptive test.

### Matrices

The same is done for the subtest Matrices. As for Series of Figures, the first Matrices item is also answered incorrectly. The second item, however, is answered correctly, after which MLE theta estimation can be started.

**Table B** Adaptive procedure for a test case for Matrices

# item	Item-ID	$\alpha$	$\beta$	Response	theta	SE	I
1	ravitem299	2.03	-0.31	0	-0.7	1.06445	0.92872
2	ravitem060	2.86	-0.73	1	-0.445	0.60246	2.04086
3	ravitem302	3.04	-0.37	1	-0.0722	0.51277	2.27549
4	ravitem614	2.63	-0.03	1	0.21516	0.50236	1.72179
5	ravitem121	2.73	0.29	0	0.02	0.38678	1.83781
6	ravitem423	2.67	-0.26	1	0.13273	0.36312	1.55900
7	ravitem830	2.19	0.23	0	0.02	0.32658	1.17990
8	ravitem009	2.18	-0.11	0	-0.0928	0.30178	1.16177
9	ravitem030	2.10	-0.06	1	0	0.29202	1.09619
10	ravitem102	2.11	-0.32	0	-0.1033	0.27565	0.99642

The SE value quickly arrives at values below the stop criterion. After ten items are administrated, the subtest is terminated, as the SE value is 0.276, i.e. reliability of 0.92. Half of the items were answered correctly, and half of the items incorrectly. The theta value is close to zero, -0.103.

### Series of Numbers

In Table C, the procedure for the subtest Series of Numbers is shown.

**Table C** Adaptive procedure for a test case for Series of Numbers

# item	Item-ID	$\alpha$	$\beta$	Response	theta	SE	I
1	cijf80	2.60	-0.12	0	-0.7	1.00276	1.64594
2	cijf546	3.97	-0.70	0	-1.4	0.95575	3.94418
3	cijf535	3.83	-1.44	0	-2.1	0.95189	3.63633
4	cijf67	4.00	-2.06	1	-1.7675	0.41703	3.96859
5	cijf51	3.42	-1.85	1	-1.59297	0.34842	2.87072
6	cijf540	2.62	-1.37	0	-1.68609	0.31724	1.56977
7	cijf59	3.53	-2.17	0	-1.93969	0.28709	1.62580
8	cijf30	2.70	-2.23	1	-1.87891	0.27107	1.56081
9	cijf11	2.17	-1.83	0	-1.94484	0.25943	1.17850
10	cijf502	3.11	-2.52	1	-1.91969	0.25023	1.18316

Only after four items have been administered, theta can be estimated by means of MLE. As for Matrices, half the items have been answered correctly and incorrectly. However, the theta estimate for Series of Numbers is far lower compared to Matrices. The items that were answered correctly are relatively easy. The SE value quickly arrives at values below 0.54, resulting in an estimated reliability of 0.94.

### Diagrams

The last subtest is Diagrams. The first four items are answered correctly. Only after the fifth item is administered, a MLE estimate of theta is possible.

**Table D Adaptive procedure for a test case for Diagrams**

# item	Item-ID	$\alpha$	$\beta$	Response	theta	SE	I
1	V3023	2.03	-0.15	1	0.7	1.37521	1.00671
2	V3020	2.2	0	1	1.4	1.65245	0.70359
3	DT38	1.53	0.31	1	2.1	2.11835	0.31304
4	DT80	1.37	0.37	1	2.8	2.73671	0.14739
5	DT58	0.77	1.42	0	1.74	1.14711	0.11220
6	V3021	1.37	0.37	1	1.949844	1.13769	0.21608
7	V2057	1.09	0.31	1	2.131641	1.14897	0.14593
8	DT44	0.97	0.46	1	2.332656	1.17384	0.12935
9	DT47	0.87	0.65	0	1.717422	0.81867	0.11530
10	DT64	1.62	0.13	1	1.783906	0.80046	0.17305
11	D19	1.75	-1.03	0	1.234297	0.58725	0.15736
12	DT84	1.88	0.13	1	1.306953	0.57450	0.35116
13	V3012	1.83	0.14	1	1.360469	0.56191	0.31654
14	V3010	1.26	0.12	1	1.431094	0.56287	0.22737
15	V3047	1.12	0.16	1	1.494297	0.56410	0.19622

It is seen that the item information values are relatively low, compared to the previous subtests. The discrimination parameters are relatively low, which has large effects on the item information values. Also, the difficulty parameters are not always very close to the estimated theta value at that point. There were at that point not many items with a difficulty parameter close to one or higher, though these items would contribute in the estimation of theta, as this is estimated to be larger than one.

After ten items, the SE value has not yet reached the stop criterion of a value below 0.54. Though the standard error decreases, it does not reach this point. The subtest is terminated after the 15<sup>th</sup> item was administered, with a value of 0.564, which corresponds to a reliability of 0.68.

