Explanatory Item Response Modelling of an Abstract Reasoning Assessment:

A case for modern test design

by

Fredrik Helland

Thesis

for the degree of

Master of Philosophy in Education

Department of Education
Faculty of Educational Sciences
University of Oslo

June 2016

Explanatory Item Response Modelling of an Abstract Reasoning Assessment:

A case for modern test design

Title:	Explanatory item response modelling of an abstract reasoning assessment: A case for modern test design
Author:	Fredrik Helland
Academic advisor:	Associate professor Dr. J. Braeken
Associated department:	Centre for Educational Measurement (CEMO)
Exam:	PED4391 - Masteroppgave - Allmenn studieretning
Semester:	Spring 2016
Keywords:	abstract reasoning, intelligence, psychometrics, assessment, cognitive lab, artificial intelligence, modern test design, item model, item response theory, explanatory item response modelling, automatic item generation,
Print:	Reprosentralen, UiO; `repro@uio.no`

Acknowledgements

Johan. Thank you for coming up with the project, supplying the know-how, for constantly pushing me to learn new skills and for the patience to explain things the first, the second, third and even the fourth time. I am very grateful for the opportunity to work with you, and without your supervision, this would not have gone as it did. It should go without saying that I have enjoyed this last year tremendously.

Thanks to Ronny for not scaring me away when I initially showed up looking for a project, and to the rest of the CEMO people for tolerating my continued presence while also being awfully nice. Thanks to Melaku for sharing your office with me, even after you had no more use for me for your project, making my time much more enjoyable.

Thanks to Olav for help finding participants, and to Annette in the iPED administration for providing much valued assistance during data collection. The cognitive lab would have gone far less smoothly had it not been for your helpfulness.

I must thank all my friends for bearing over with me while I was burying myself in the thesis. Special thanks to Stian, Kristian and Espen for proof-reading my text and giving valuable feedback. Also thanks to Espen for getting me started with LaTeX, and to Lily for making his template work.

Lastly I have to give thanks to Siri, for always being there, for pushing and loving me. You are my caregiver and my role model, and I love you.

Fredrik Helland
June 14, 2016

Abstract

Assessment is an integral part of society and education, and for this reason it is important to know what you measure. This thesis is about explanatory item response modelling of an abstract reasoning assessment, with the objective to create a modern test design framework for automatic generation of valid and precalibrated items of abstract reasoning. Modern test design aims to strengthen the connections between the different components of a test, with a stress on strong theory, systematic item design, and advanced technological and statistical tools. Such an approach seeks to improve upon the traditionally weak measures found in education and social sciences in general.

The thesis is structured in two parts. Part one presents the theoretical basis of the dissertation, and part two presents the empirical analysis and results of the assessment. The first chapter establishes an understanding of the general field of which this study has been conducted. The second chapter delves into the particular content domain relevant for the assessment. The third chapter presents the actual assessment that is the object of investigation. The fourth chapter presents a comprehensive report on a cognitive lab. The fifth chapter presents the factors on which the actual explanatory item response modelling of the assessment is founded on. The last chapter present a general discussion and conclusion of the study.

Contents

List of Tables

List of Figures

List of Abbreviations

1PL	one parameter logistic model
AI	artificial intelligence
AIC	Akaike information criterion
BIC	Bayesian information criterion
CAT	computer adaptive test
HR	human resources
IRT	item response theory
NASA TLX	NASA task load index
NSD	Norwegian Social science Data services
PISA	Programme for International Student Assessment
RPM	Ravens Progressive Matrices
TIMSS	Trends in International Mathematics and Science Study

List of Symbols

α	Significance level
β	Regression weights of the explanatory predictors
η	Predictor component
σ	Random variance
θ	Estimated ability parameter
ε	residual term

b	Difficulty parameter
i	Item
k	Item property
p	Person
SD	Sample standard deviation
X	Item predictor values
Y	Item response

Part I

Introduction to the Methodological Field, the Content Domain, and the Assessment

The following part presents the theoretical basis of the dissertation. The first chapter establishes an understanding of the general methodological field under which this study has been conducted. The second chapter delves into the particular content domain relevant for the assessment. The third chapter presents the actual assessment that is the object of investigation.

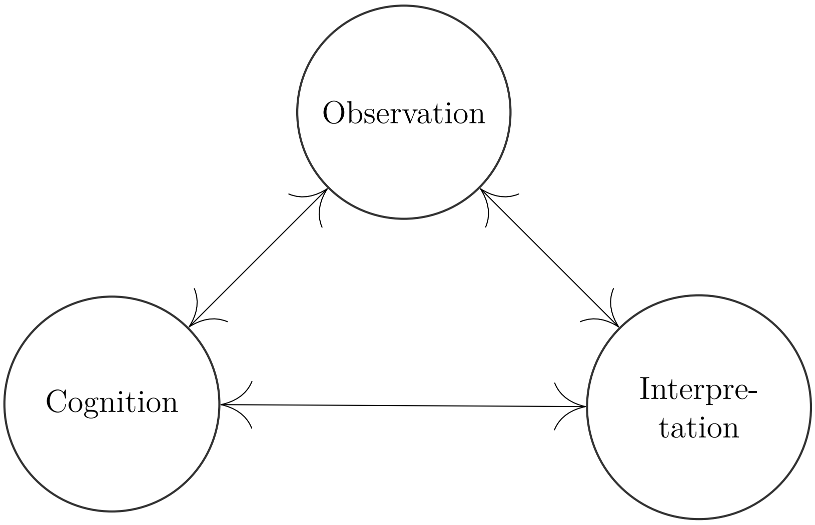
Chapter 1

Modern Test Design
– *the Methodological Field* –

Assessment is an integral part of education. Teachers make test quizzes for assessing their students learning, governments set up national exams to monitor and safeguard educational quality, and even internationally there is an interest in comparing educational outcomes by means of large scale educational assessments such as Programme for International Student Assessment (PISA) (OECD, 2016) or Trends in International Mathematics and Science Study (TIMSS) (Mullis & Martin, 2013). Tests have also widely used for a long time to screen for candidates by for instance human resources (HR) companies (Raven, 2000a) or universities (Sternberg, 1985). Pellegrino (2003) nicely outlines three core component that should be part of every and each of such assessments:

> Any assessment must meld three key components: cognition, which is a model of how students represent knowledge & develop competence in the domain; observations, which are tasks or situations that allow one to observe students performance; and interpretation, which is a method for making sense of the data relative to our cognitive model. Much of what weve been doing in assessment has been based on impoverished models of cognition, which has led us to highly limited modes of observation that can only yield extremely limited interpretations of what students know.(Pellegrino, 2003, page 49)

Figure 1.1: Components of an assessment



Hence, when interested in assessing how proficient a person is in content domains like for instance mathematics, English grammar or abstract reasoning we need to design a measurement instrument or test that can differentiate between high and low performers. For that you first need tasks or situations that give opportunity to *observe* mathematics performance - this role can be filled by test items in a traditional mathematics exam for instance. Secondly you would need *theory* to help define which and in what way these items are relevant to the skills and learning of the mathematics domain. Thirdly, models and other tools are needed to make inferences and decisions based on performance on these items in light of the theoretical foundations. Finally all three components need of course to interact with each other to create an integrated assessment. This is graphically represented in the assessment triangle in Figure 1.1. The triangle would generally apply to assessments made for most content domains.

What you intend to assess is considered a latent construct as it only manifests itself in observable behaviour through the item tasks. Hence, the test items define the bounds of the measure that can be obtained and determine the concrete operationalisation of the construct that is to be assessed. In the social sciences, the items thought to represent the latent constructs have traditionally been poorly defined. The design and specification of a test and its items has often been a rather artisanal undertaking. The items have mostly been individually hand-crafted by content experts (Drasgow, Luecht, & Bennett, 2006), leaving it be a more of a creative inspiration-driven exercise where the main theoretical foundation is the assumed common understanding of the item writers. This has left item generation rather disconnected from scientific theory on cognition and learning (Hunt, Frost, & Lunneborg, 1973). The combination of an unsystematic approach to item design and the weak theoretical foundation for the entire measurement instrument (Drasgow et al., 2006) are part of the reasons why education and social sciences in general, are still haunted by weakly defined measures.

Modern test design approaches that go under different labels such as Evidence Centered Design (R. Mislevy & Riconscente, 2006; R. Mislevy, Steinberg, Almond, Haertel, & Penuel, 2001; R. J. Mislevy, Almond, & Lukas, 2003, see e.g.), in Assessment Engineering (Luecht, 2003, 2013), and in Cognitive Systems Design (Embretson, 1994, 1995, 1998) hope to counter this tradition of weakly defined measures. All the modern test design approaches have in common that they aim to strengthen the connections between the different components with a stress on strong theory, systematic item design, and advanced technological and statistical tools.

1.1 Validity and calibration of items

Traditionally, calibration and validation of the test would mostly happens a posteriori, with psychometric calibration and theoretical validation only coming into play after the test has been created and administered (Lai, Gierl, & Breithaupt, 2012). The validation has traditionally been performed on the scale level, by mapping the nomothetic span of the construct (Embretson, 1998). This is done by correlating test scores with other scores on selected measures of constructs that are expected to converge or diverge with the construct underlying your own test (Campbell & Fiske, 1959; Cronbach & Meehl, 1955).

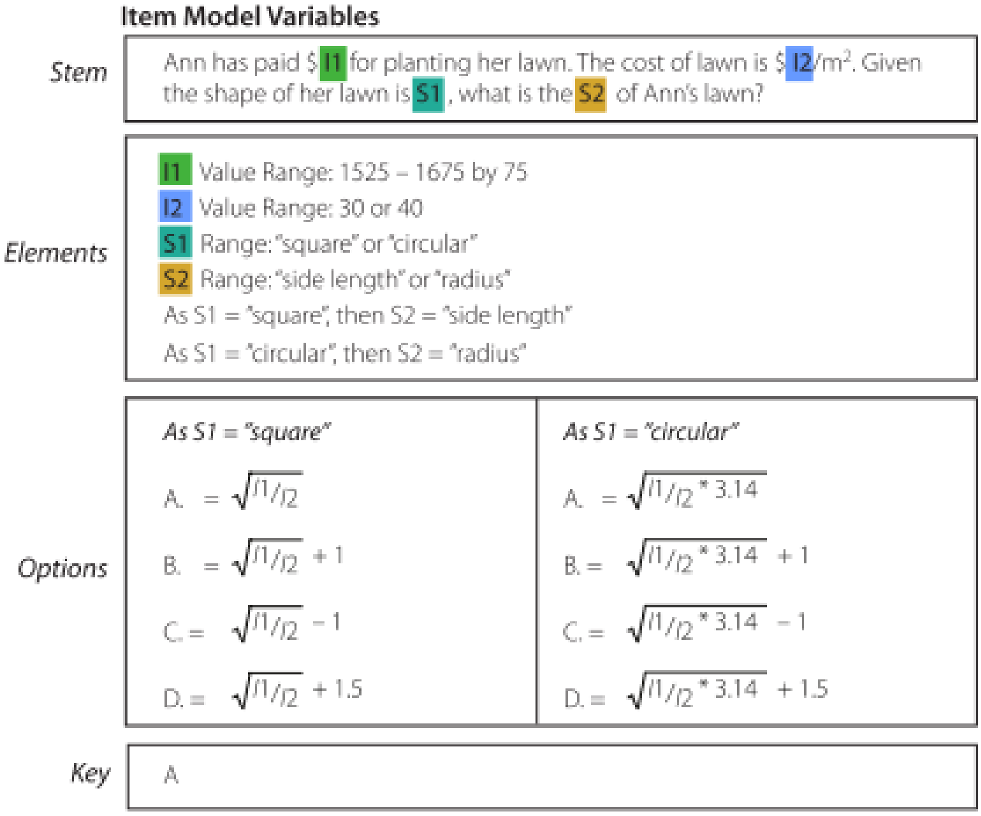
Yet, validity is more that just knowing the network of relations between constructs - it is also about knowing the theoretical mechanisms underlying the behaviour of your items (Whitely, 1983). More specifically, you want to identify the specific properties or facets of the items that might influence how participants respond, also known as radicals, as well as facets which are merely cosmetic, which are known as incidentals (Irvine &

Kyllonen, 2002). In essence, radical and incidental elements form the core of what is often called the item model (Drasgow et al., 2006). Your item model would be the operationalisation of a cognitive theory on your assessment (Embretson, 1998). It is the explicit representation of all the variables in your item (Gierl, Zhou, & Alves, 2008) and thus defines the measure.

Under a weak theoretical design approach, validation does not really occur at the item level. Instead focus is on statistical detection of specific items with ill-functioning psychometric characteristics such that detected "misfitting" items can be physically eliminated from the item pool (Bond & Fox, 2001). It is established that these items do not work out as intended, but we have no clue or interest in the reasons why they are inadequate to measure the construct. Generating a large pool of items to expand or renew an existing test or item bank either relies on a lot of inspiration of many item writers or a reuse of old items by a process called item cloning. The former approach does require some informal convergence of item writers on what type of items are eligible, comparable to the existing one, and in line with guidelines and blueprints from the weak theoretical framework. The latter cloning approach (Glas & van der Linden, 2003) copies an existing item but merely makes some cosmetic changes to create the new cloned item. This means that clones should be radically the same, but incidentally different. Yet, if the only knowledge you have on the item level is a descriptive account of item difficulty, you are left without anything but hypothetical knowledge of which item facets are radicals and incidentals under a weak theory approach. The decision of what item facets to change is largely based on the intuition of the designer, and this substantially limits how much you can do without unintentionally altering the way the item behave.

What is needed first and foremost for making the most of modern test design approach is a viable item model. Unfortunately, generating valid item models is not always realistic in all situations where assessments are being made. Especially in poorly defined content domains, there are few established principles on which to generate a model. In contrast, if your construct is sufficiently well-defined, strong theory on item radicals and incidentals is available. These radicals and incidentals can be employed as factors in an experimental design for the items of an item bank. Creating templates that define what your items are supposed to look like Gierl et al. (2008) (e.g., stem, item structure, response alternatives etc. For an example template, see Figure 1.2) then even allow you to automatically generate tons of theoretically motivated items and easily fill up an item bank. If the item model is also transformed to a statistical model that can predict the impact of item radicals on psychometrical item behaviour, then in principle we can also simultaneously precalibrate newly generated items without having to field trial them. One could in essence be able to create individually tailored items on demand, based on nothing but a list of item facets and a mathematical model to combine them. Alas this type of item model would be very hard to attain with absolute certainty, and might or might not be an utopia.

Figure 1.2: Example of an item template used for generating items in mathematics



Item Model Variables

Stem

Ann has paid $ I1 for planting her lawn. The cost of lawn is $ I2/m^2. Given the shape of her lawn is S1 , what is the S2 of Ann's lawn?

Elements

I1 Value Range: 1525 – 1675 by 75
I2 Value Range: 30 or 40
S1 Range: "square" or "circular"
S2 Range: "side length" or "radius"
As S1 = "square", then S2 = "side length"
As S1 = "circular", then S2 = "radius"

Options

As S1 = "square"	*As S1 = "circular"*
A. $= \sqrt{I1/I2}$	A. $= \sqrt{I1/I2 * 3.14}$
B. $= \sqrt{I1/I2} + 1$	B. $= \sqrt{I1/I2 * 3.14} + 1$
C. $= \sqrt{I1/I2} - 1$	C. $= \sqrt{I1/I2 * 3.14} - 1$
D. $= \sqrt{I1/I2} + 1.5$	D. $= \sqrt{I1/I2 * 3.14} + 1.5$

Key

A

Note. The template has some fixed features, and some variables. The two variables first on the element list, I1 and I2, changes numbers only, hence in this case are merely incidentals. The other two variables act as radicals, and they switch between geometric concepts of varying degree of difficulty. If there was no strong theory on difficulty of geometry items, only the two incidentals could be altered safely. In that case any generated items from this template would be a clone. This illustration was borrowed from the report of Gierl et al. (2008) on assessment engineering.

1.2 Item response theory

It is however possible to strive for a representationally valid test, by developing a strong theory and using a method called item response theory (IRT). IRT is not really a theory, but a family of statistical measurement models used for item response data (Baker & Kim, 2004). It is the current method of choice in assessment practice. The core of IRT are the common measurement scale(s) locating both persons as well as items in the same metric system. Furthermore, any item response model formulates the the response probability of a person on a given item of the test as a function of characteristics of both the person as well the item. This leads to a whole IRT family of models that all have three main assumptions in common:

1. **Dimensionality:** A small set of dominant factors explain individual differences and response variation between persons and response interdependence within a person;
2. **Conditional Independence:** Once you know someone's position on the latent dimensions, a response on one item cannot provide any extra information anymore with respect to the person's response on another item;
3. **Monotonicity:** The item characteristic curves specifying the relation between the latent dimensions and the observed item response are always non-decreasing (i.e., more proficient persons will never have a lower chance of answering an item correctly, than low proficient persons).

The reason why IRT is so popular, is exactly because it lets you calibrate both the item and person parameters jointly on the same (usually logit) scale. Persons can be compared to persons, items to items, and even persons to items all on the same scale using the same measurement units. This is more useful than the other alternative scaling method, classical test theory, as the latter leads to person-focused norm-referenced measurement scales. Here there is no scale link between the person and item parameters. As soon as the selected set of items changes, new norms need to be computed and a new scale and score units arise. In contrast, if the IRT model fits for item bank, it implies that we can meaningfully compare persons to persons regardless of the set of items that were actually administered to each person (i.e.,technically this is sometimes called the IRT invariance property). This is because both persons and items are positioned on the same scale with scores in the same metric, and this opens up for all kinds of exciting opportunities for test assembly.

Furthermore, having established a common scale, it becomes possible to extend the IRT model with explanatory predictors based on the item model (De Boeck & Wilson, 2004). This implies that item parameters can now be modeled not on an unique item-wise individual basis, but on a more general theoretical basis as a function of the radicals defined in the item model and not determined by the incidental cosmetic elements. This explanatory IRT approach opens up for generalising the scale from the observed sample of calibrated items to unobserved instances in the theoretical item population, essentially giving just a little glimpse of utopia still (Section 1.1).

1.3 Test assembly

Whereas traditionally the assembly of a new test from an existing item bank would involve selecting a fixed set of items that comply to some vaguely defined blue-print criteria, modern test design and use of explanatory IRT opens up effective and efficient pathways for the assembly of more tailored tests, because you can generate valid and pre-calibrated items on demand.

Tests can be tailored with respect to a specific purpose. Through strong theory we know how to generate valid items that will behave as intended. For diagnostic screening purposes of low performers we need a test targeted at the lower side of the scale. Strong theory tells us what items to make as we know what to manipulate and how.

Similarly for selection purposes, you would want to assembly several equivalent test forms targeted at the upper-end of the scale. Here as well, strong theory can guide us which items to generate and include in the item bank for the tests, with IRT allowing us to create the equivalent test forms as all generated items are scaled in the same metric.

Yet tests can also be tailored to each specific individual by setting up a so-called computer adaptive test (CAT) (Van der Linden & Glas, 2009; Wainer, 2000). Computational and statistical techniques from item response theory (IRT) and decision theory are combined to implement a test that can behave interactively during the test process and adapts towards the level of the person being tested. The implementation of such a CAT relies on an iterative sequential algorithm that searches the pool of available items (a so-called item bank) for the optimal item to administer based on the current estimate of the persons level (and optional external constraints). The subsequent response on this item provides new information to update the persons proficiency estimate. This selection-responding-updating process continues until the person ability level is pinpointed with sufficient certainty or you run out of available items. The advantage of this is twofold. Firstly, the number of items needed to get a reliable estimate of the person diminishes. Secondly, as there would be no need for the persons to answer items that are way too easy or too difficult for them, tests would be gentler on the persons (i.e., test burden decreases). For a CAT to work well you want a large IRT-calibrated item bank, which in a modern test design is easily accommodate with theory-driven automatic item generation and psychometric item models.

1.4 Bridge to content domain

The intention of this overall project is to create a modern test design framework for an existing abstract reasoning assessment, with the long term goal of automatic item generation and the development of a CAT. The role of this particular thesis study is mainly to serve as a proof of concept for the feasibility of the long-term goal, but also to lay the groundwork for an item model.

The rest of the chapters in Part I deals with the theoretical foundation of the study. Relevant cognitive theory from the content domain will be presented in Chapter 2, before describing the current abstract reasoning assessment in Chapter 3. Part II contains the empirical work laid down in the study. Chapter 4 reports the comprehensive results from a cognitive lab, consisting of a think-aloud, an interview, as well as a small performance questionnaire. Chapter 5 presents the results of a logical-rational task analysis based on the creation an artificial intelligence to solve the test, synthesised with the already defined radicals from the general content domain and findings from the cognitive lab to create a tentative item model - formulating the concrete hypotheses of the study. These hypotheses are then tested using explanatory item response modelling of the existing abstract reasoning assessment. Chapter 6 is a general discussion of the implications of the findings for further development of the assessment, with concluding remarks.

Chapter 2

Abstract Reasoning

– *the Content Domain* –

2.1 Abstract Reasoning, Fluid Intelligence and Complex Problem Solving

Reasoning is a very old scientific discipline, going back to the ancient Greek study of logic and rhetoric. The ability to string together logical arguments was an important mark of intelligence, and this was naturally an important requirement for participating in the politics of the time, as it is today. This ability to argue effectively and coherently is still a valued skill in much of society, and hence reasoning tests are a popular criterion for mapping, screening or ranking for selection, and many popular intelligence tests usually include both verbal and figural reasoning tasks or subtests (Hunt, 2011). Reasoning in itself can be described as an explicit sequential application of a set of rules to a formal problem (Gilhooly, 2004). Common operationalisation principles behind the assessment of reasoning are to present cases or sets of premises and either make the person generate a conclusion themselves, judge the validity of a presented conclusion or rule, or make an inference from a problem (Leighton, 2004).

Reasoning tasks with a rich basis in meaningful cultural content might have a tendency to confound the measurement of intelligence with other factors such as social and cultural background (see Cattell, 1940; Flynn, 2007). To amend such culturally rich tasks, researchers introduced abstract figural tasks, as for instance used in the Ravens Progressive Matrices (RPM) (Raven, 2000b) and the Cattell culture fair test (Cattell, 1940), that are supposed to tap into the intelligence construct without taxing language ability or prior knowledge (Raven, 2000a). It is debatable whether this actually worked (Wüstenberg, Greiff, & Funke, 2012), but the general idea is that these tests should be more fair for a diverse group of people. Abstract reasoning tests generally differ from other complex problem solving measures in that it is supposed to be more context- and content-independent (Leighton & Sternberg, 2003), contrasted to incorporating the context as part of the assessment (Raven, 2000a). Abstract reasoning is also characterised as being fairly static, where you are given all the necessary information from the very start, as opposed to for instance assessments of more complex problem solving measures, where persons have to strategically interact with some sort of system in order to gather the relevant information to solve the problem (Wüstenberg et al., 2012).

The generality of abstract reasoning tests has made these tests popular for a long time. After all, abstract geometric problems are considered the classic way of assessing individual difference in fluid intelligence (Marshalek, Lohman, & Snow, 1983). There are many factors that might limit (or bound) rational problem solving or reasoning ability, both human-, and task-wise (Simon, 1972). The rational way of solving problems, based on deep processing with cognitive scripted procedures, is more effective than superficial heuristic strategies (Sweller, 1988), but require sufficient processing power in accordance with the complexity of the task. People are quite limited in terms of information-processing capability (Miller, 1956), and when making assessments of individual differences in reasoning or other intellectual abilities, we essentially want to take advantage of this fact.

To differentiate between the persons taking the test, it is important to include items of varying difficulty. A general theory on what facets constitutes the problem difficulty, the so-called "radicals", in an abstract reasoning test is somewhat challenging to determine, as most tests differ in some way, both with the item structure as well as the type of elements and rules the test contain (Jacobs & Vandeventer, 1972). According to Stanovich, Sá, and West (2004), reasoning errors usually have two sources: a lack of mental capacity as well as the challenge of understanding the premises of the rules. Identifying potential theoretically-motivated radicals that tap into these two sources is the primary goal of this section.

2.2 Potentially relevant Cognitive Processes and Resources

One of the most famous and extensively studied reasoning tests in this domain are the RPM (Raven, 2000a; Wüstenberg et al., 2012), which consist exclusively of what is called matrix and completion problems (Carpenter, Just, & Shell, 1990; Embretson, 1998, 2004). An example of an easy matrix and completion problem can be found in Figure 2.1, and a more difficult one in Figure 2.2. These are mostly based around pattern recognition across the two dimensions of a figural matrix, and are perhaps best described by the task analysis of Carpenter et al. (1990):

> *Each problem consists of a 3 x 3 matrix, in which the bottom right entry is missing and must be selected from among eight response alternatives arranged below the matrix. [sic] (Note that the word entry refers to each of the nine cells of the matrix). Each entry typically contains one to five figural elements, such as geometric figures, lines, or background textures. The test instructions tell the test-taker to look across the rows and then look down the columns to determine the rules and then to use the rules to determine the missing entry.* (Carpenter et al., 1990, page 4)

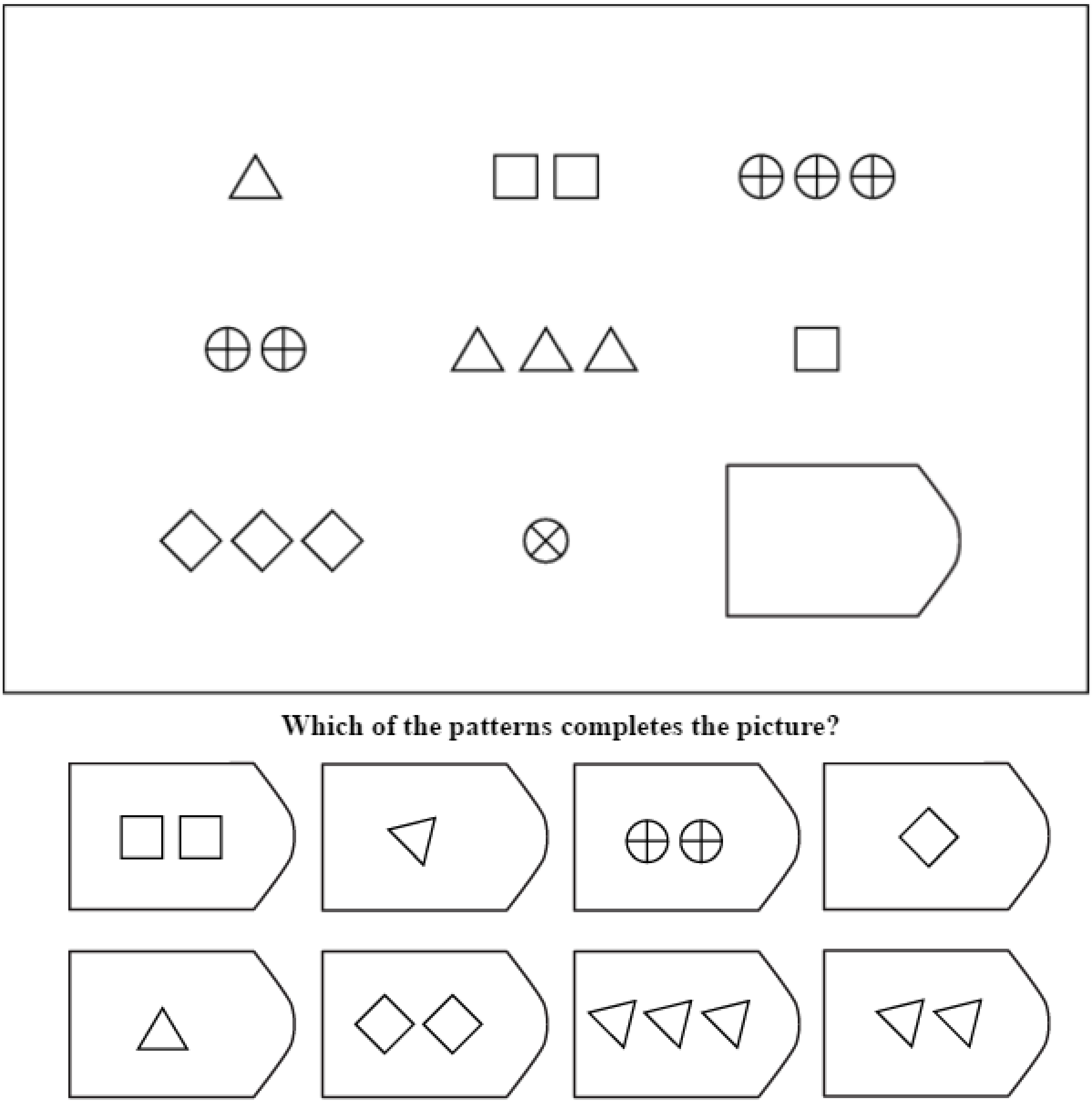
Quite a decent amount of work has been done on identifying the radicals of matrix and completion problems, and in particular the RPM. Item facets putting greater demand on the working memory of the persons should be considered as a starting point for working out the radicals in most cases. In his review, Primi (2001) claims that radicals of a matrix test can be divided into three types: A) the amount of information, B) the type of rules, and C) the perceptual organisation of the item. Although originally targeting matrix completion tasks, this general synthesis framework should have some merit for understanding other abstract reasoning tasks such as the assessment that is the focus in Chapter 3 and should prove useful in identifying potential radicals that can be later used to construct more specific item models.

2.2.1 Amount of Information

Amount of informations refers to the quantity of elements and rules in the problem, which has generally been associated with working memory load (Arendasy & Sommer, 2005; Hosenfeld, 1997; Mulholland, Pellegrino, & Glaser, 1980).

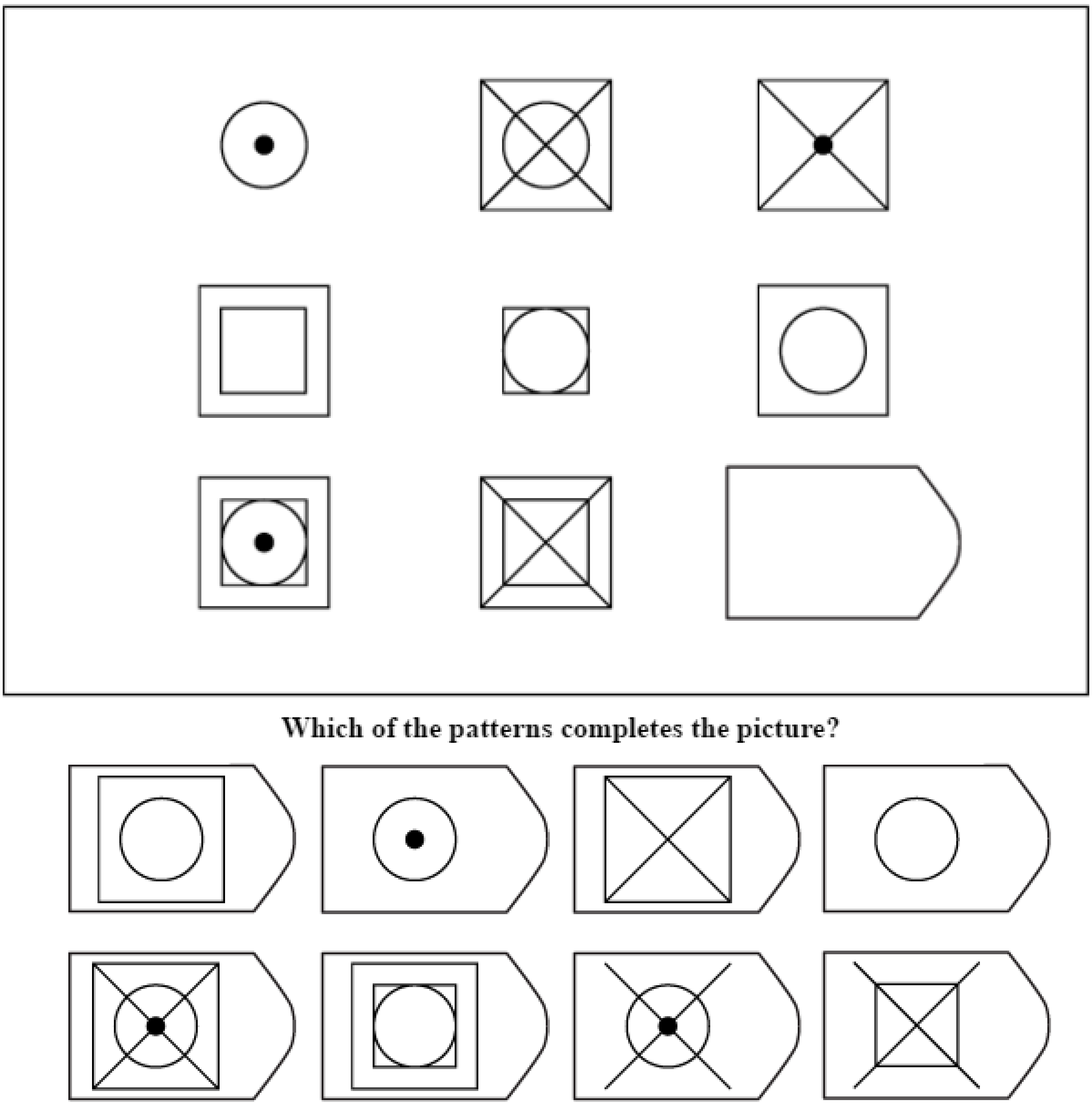
Gilhooly (2004) defines two main approaches to working memory: a single pool of resources approach, and a multi-components approach.The single pool working memory varies both between as well as within persons. It has a storage component and a processing component. The resource functions as a bottleneck for information processing, where the amount of information one is able to recall, defines ones working memory capacity. Items with a high information density would put a higher load on the working memory capacity of the person, thus leading to individual differences in this respect. The multiple components approach is based on the Baddeley and Hitch (1974) three-pronged working memory model. It consist of two working memory systems, the phonological loop and visuospatial sketchpad, which are regulated by a higher-order central executive. Verbal and spatial information is thus processed in separate independent systems. The executive functions constitute a persons goal management ability among other things (Miyake et al., 2000). This is thought to be a central factor determining abstract reasoning ability, as the ability to partition up the reasoning task into tangible chunks and sub-goals is crucial for effectively solving them (Carpenter et al., 1990).

Figure 2.1: Easy example item from a matrix and completion test



Note. This item was borrowed from a demonstration test at the Concerto platform developed by the Psychometric Centre, University of Cambridge (The Psychometrics Centre, n.d.). Here the horizontal rule cycles features and the vertical rule numbers within a set. It can also be referred to as distribution of three-values.

Figure 2.2: Slightly more difficult example item from a matrix and completion test



Note. This item was borrowed from a demonstration test at the Concerto platform developed by the Psychometric Centre, University of Cambridge (The Psychometrics Centre, n.d.). The rule in question here is a subtraction rule, also referred to as distribution of two-values. More features are involved here than in Figure 2.1.

Persons construct mental models of the reasoning problems by modelling the relations between the premises of the problem (Johnson-Laird, 2004). If there are enough premises and relations, the persons will not be able to reason properly because that load is too large (Stanovich et al., 2004). As working memory capacity most likely determines the person's ability to maintain complex representations, or mental models, of the relations between the premises in the problem (Stanovich et al., 2004), a problem which complexity exceeds the capacity of the person, should be an important cause of reasoning errors.

2.2.2 Type of Rules

Next to the more quantitative aspect of amount of information, also the qualitative aspect of the rules themselves is important within the reasoning process. The quality aspect refers to the different changes that can happen to the elements in a problem. Jacobs and Vandeventer (1972) categorised rule types from 1335 matrix items from 22 different tests, and ordered them into a universe of twelve types of rules: Identity, shape, shading and size change, movement in a plane, flip-over of a shape, reversal of order of elements, adding elements, addition of features, unique addition (duplicates get negated), number series addition, and alternation between elements in a set. For the RPM, Carpenter et al. (1990) synthesised the rule categories of A) constants in a row, B) quantitative pairwise progression across the rows, C) figure addition or subtraction (facets are arithmetically operated to produce a product), D) distribution-of-three-values (alternations within a set of elements in a row) and E) distribution of two-values (identical features negate each other, leaving only the unique). From this theory, it was possible to successfully generate new matrix and completion items (Embretson, 1998). Of course all these rule types are based on problems that has a 3x3 matrix structure, so all of them would not be directly applicable to other cases.

Primi (2001) made a three level rule complexity taxonomy. Simple rules like shape or size changes should be the most tangible at level one. Level two rules has to do with spatial changes, like rotation of a shape, which would be less tangible than the simple changes. The most abstract rules at level three are classified as complex or conceptual rules, where groupings of the item features are mainly on a structural or conceptual level that makes simple pattern recognition not necessarily straightforward, like for instance adding features or feature attributes, or altering features with respect to conceptually abstract categories instead of visually salient progressions.

Having a diversity of rule types in the test could contribute to individual differences in problem solving success by challenging the persons ability to connect changes happening to the item features in a meaningful way. There is some grounds to claim that persons apply a fixed repertoire of rules when reasoning (Stanovich et al., 2004). Hence if there is a lack of conception of, or misconstruation of a rule present in a given problem, the reasoning would be erroneous right from the start.

2.2.3 Perceptual Organisation

Information processing is a key cognitive mechanism. According to Atkinson and Shiffrin (1968), information is processed by the sensory memory, before it enters the short term memory (hereafter referred to as working memory). The information that is rehearsed may eventually be stored in the long term memory, where it can be retrieved back to the working memory under certain conditions. This process is an integral part of successful reasoning, as it contributes to controlling, regulating and actively maintaining task-relevant information (Miyake & Shah, 1999, in Gilhooly, 2004).

Not only the amount of information or the rule type are a cause of errors in the reasoning process. Perceptual factors not directly stemming from rule complexity can be the cause of failure to understand the problem. According to Primi (2001), there are perceptual features independent of the other categories that contribute to item difficulty on their own by conforming to or violating gestalt principles of perceptual harmony. More specifically, it relates to “ambiguity, contradiction among perceptual and conceptual groupings, and the number of misleading cues” (Primi, 2001, page 50). In other words, items with features that incidentally come across as perceptually ambiguous, challenge the problem solving process of the persons by blurring the systematic patterns in the problem.

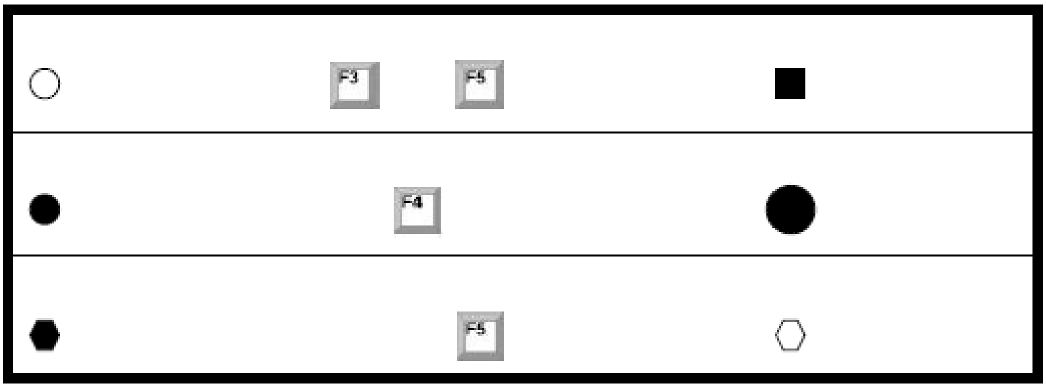
Chapter 3

Abstract Reasoning Test – *the Assessment* –

The particular assessment under investigation in this study is an older version of a high stakes job recruitment test. It is a linear power test, computer-based, and the persons are under test-wise time constraint. The assessment is characterized by being an unknown system where the participants have to decipher some problem rules in order to solve a problem. It is modelled on other abstract reasoning tests, like the RPM.

3.1 Basic description of the test

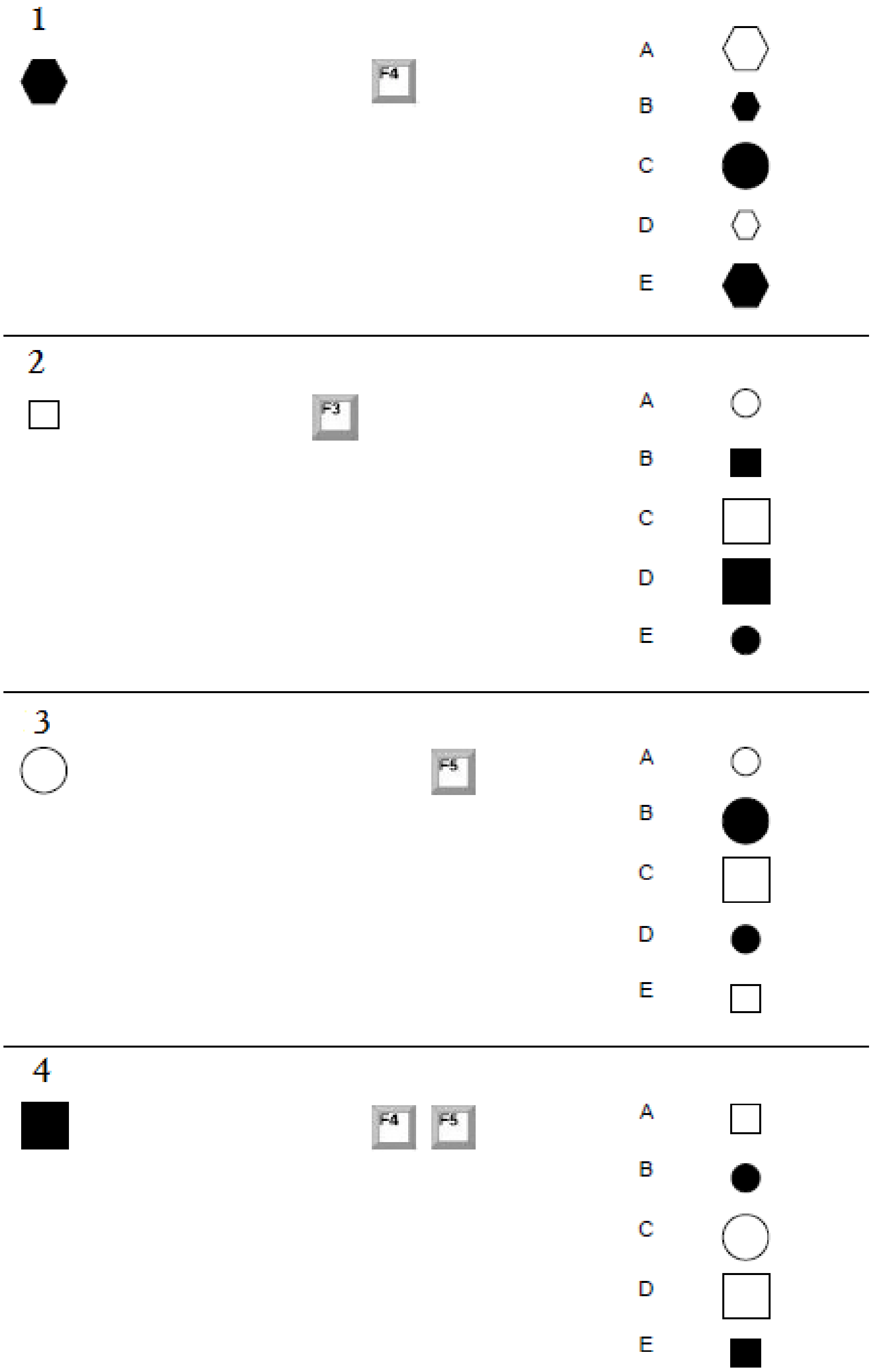
Figure 3.1: Example of an easy problem from the abstract reasoning test.



The test consist of ten testlets. A testlet is basically a set of items attached to a common stimulus (Wainer & Kiely, 1987), and in this case the stimulus is a problem (Figure 3.1), with four individual items connected to it (Figure 3.2). Each problem consists of two sets of figures, respectively the start figures and end figures. Between the two figure sets are a number of buttons. The buttons determine which operation is performed on one or more basic figures. The changes happening between the start and end figures then have to be matched with the corresponding buttons in each row to find out what rule they operate. The person must then apply the relevant rules to the start figures in four related items, choosing, from five alternatives, which end figure he or she thinks is the right one given the buttons (in essence, rules) present in that item. Only one of the alternatives is correct.

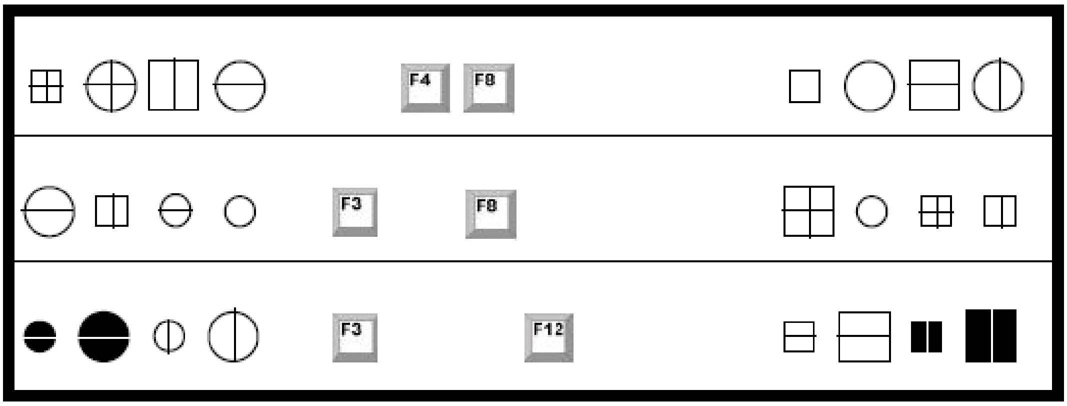
In Figure 3.1, we see three rows of figures, each with one figure in the start and end set. There are also three different buttons in the problem, arranged across the rows. We can see that the buttons F4 and F5 are isolated in their row, and it is thus an easy task to identify the rules associated with them, which is respectively enlarging the figure and turning it white. The two buttons in row one are dependent on each other, so you need to decipher other rows to solve them, but since it is know from row three that F5 changes the colour, F3 must then turn the circle into a square. Although this is the logic by design, it is not explicitly stated whether the operation is a one-way or a two-way operation (the size change in Figure 3.1 an example). Persons have to figure this out for themselves, either when working on another row where the changes are reversed, or when

Figure 3.2: The items connected to the example problem in Figure 3.1.



Note. The correct answers are respectively B, A, B and A trying to apply the derived rules to the items (for example item one in Figure 3.2).

Figure 3.3: Example of a slightly scarier problem from the abstract reasoning test.



A person solving the test is presented with testlets of cumulatively escalating difficulty (see for example Figure 3.3), and has to solve as many as possible until the time runs out.

3.2 Data and data processing

The test data were acquired from the international HR company that owns the test, and consist of a sample of Belgian nationals ($n = 6689$) that took the test as part of a job recruitment process.

Table 3.1: Table of descriptives for the test data.

	Original n	Adjusted n
Total	6689	6519
Male	4255	4144
Female	2063	2016
Unknown	371	359
Flemish speaking	4552	4438
French speaking	2137	2081

It is natural that some persons took the test without making serious attempts at solving it, either because they just skimmed through or because of some test administration error or failure. Thus any person with a total time spent being in the bottom 2.5% of the sample were removed from further analysis (see Table 3.1). Remaining was a total of 6519 persons, where 4255 were male, 2016 female and the remaining 359 were classified as unknown. The data also included information on their native language, of which 4438 reported speaking Flemish, compared to 2081 French.

The persons were under test-wise time pressure, which means that many persons were unable to answer all the items before the time ran out. In situations where you are interested in information on the persons, not reaching some items could be interpreted as an indicator of ability, and thus a non-response should be scored as a wrong answer. When you are interested in the items however, running out of time gives you no additional information on the difficulty of any of the items the person did not reach (R. J. Mislevy & Wu, 1996). Since the focus in this thesis is to investigate item properties, responses with zero recorded reaction time were presumed not reached, and considered missing.

In the study, the term *participant* exclusively refers to the particular sample of people participation in the cognitive lab. When referring to people in general, the term *person* is used, to keep with the terminology of IRT.

3.3 Feasibility of Modern Test Design

The abstract figural item contents of this reasoning test makes it a prime candidate for a redesign according to modern test design principles. In order to modernise the test design, it will be necessary to further establish a sound theoretical basis for the creation of an item model. In the first line, this could be achieved by reverse-engineering the existing assessment to identify concrete item facets - radicals and/or incidentals - based on both cognitive theory as well generated input from initial analyses of the existing item pool. Alternative methods, be it qualitative or quantitative, can be used to assist in making causal explanations by letting you explore phenomena and triangulate your findings (Shadish, Cook, & Campbell, 2002). Exploring how the test is structured logically as well as how people solve it in practice are natural steps in generating theory on how it works. In order to investigate the human factors when solving the test, a cognitive lab study was set up. The assessment was also studied from a more rational perspective by performing a task analysis and creating an artificial intelligence algorithm (artificial intelligence (AI)) to solve the items. Input from theory, cognitive lab, and the task analysis were synthesised in order to create a preliminary item model. In order to provide an initial validation of this item model, factors from the item model will be operationalised as explanatory predictors which are put to the test by performing explanatory IRT modelling on the actual test data.

There are some constraints that come along with building an item model based on the existing test design. Because the test consist of testlets, the effective item pool size is smaller than the actual item pool size. Every item does not present a unique problem, but every set of four (in essence, testlet) does. Furthermore, the item pool of the existing test does not have the design-wise rigour of a true experiment. There is no experimental design or randomization across any of the item facets that we might end up investigating, and thus it would be difficult to determine or disentangle all causal mechanisms underlying the responses on the item pool. Combining the lack of experimental design and the limited item pool size, we see that any conclusions stemming from reverse-engineering the existing assessment cannot be guaranteed to be robust. Hence, this is only the initial phase of a larger project, and the establishment of any radicals or incidentals would have to be corroborated later in properly designed experiments.

3.3.1 Software and graphics

The statistical analyses were coded in the software environment R (R Core Team, 2015) using the package `lme4` (Bates, Mächler, Bolker, & Walker, 2015) to estimate the item response models. All statistical graphics were made using the R package `ggplot2` (version 2.0, Wickham, 2009). Other figures were constructed using `Tikz` (Tantau, 2015), a graphics package for LaTeX. Initial analysis of the audio-data from the cognitive lab were done using NVivo 10 (QSR International Pty Ltd., 2012), before being processed with R.

Part II

Generation and Empirical Validation of an Initial Theoretical Item Model

The following part presents the empirical analyses and results to start and inform the initial reverse-engineering of the assessment. The fourth chapter puts forward a comprehensive report on the cognitive lab. The fifth chapter covers the explanatory item response modelling of the assessment. The sixth and last chapter presents a general discussion and conclusion of the study.

Chapter 4

Theory generation – *Cognitive laboratory* –

The method chosen for the exploratory procedure is the cognitive laboratory. This is a group of methods often used to investigate cognitive processes during person-artefact interactions (Ericsson & Simon, 1993; Nielsen, Clemmensen, & Yssing, 2002), and has its historical origins in the introspection procedures of early German experimental psychology, having since become an important research tool for both psychologists and educationalists (Leighton, 2005). The most basic elements in the lab toolbox (Katz, Moon, & King, 2015) are the concurrent think-aloud procedure and the retrospective interview, in addition to the usability study. There are numerous variations of the method, but according to Katz et al. (2015), they differ in general in that the concurrent think-aloud is a procedure where the participant talks aloud whilst performing the task, with the researcher being broadly non-interfering; the participant in the retrospective interview first solves the task, then gets to explain his thinking; the researcher continuously prompts the participant in the usability study to interact with the task in a certain way. Of course these can be, and often are, used in combination.

Variations of the method have been used in psychometrics and educational measurement studies on multiple occasions. Katz, Bennett, and Berger (2000) used the concurrent think-aloud procedure to generate categories of strategies when solving mathematics problems, and investigated the effect of item format on use of said strategies; Carpenter et al. (1990) used the think-aloud method in conjunction with eye-trackers to generate theory on how people solve matrix problems in an intelligence test; Winter, Kopriva, Chen, and Emick (2006) administered a mathematics test with a post-hoc interview, in order to obtain information on item interpretation, cognitive processes involved when solving each item, as well as the saliency of specific features of each item. The method is also used in a wide range of other research fields, from development of diagnostics tools (see e.g. Paap, Lange, van der Palen, & Bode, 2015) to human-computer interaction research (Bastien, 2010; Nielsen et al., 2002).

Originally used to specifically gather information on what goes on in a person's short-term memory (Ericsson & Simon, 1993), a cognitive laboratory approach was deemed to be well suited for enriching the information on the abstract reasoning test in question. Finding information on what item features functions as radicals was one key objective when doing the cognitive lab, as these features constitutes differences in item difficulty (Gierl & Haladyna, 2013). I was also interested in anything the participants might express that could shed light on consequences of having a testlet based test structure. To investigate this, the time participants spent deciphering the problem rules and applying them to each item, was coded to see if there were any interesting patterns. The expectation was that participants with a strong focus on rule learning would spend a greater proportion of time deciphering, and would be more efficient when solving the items.

4.1 Methods and materials

This part of the study utilises a cognitive laboratory procedure where participants were asked to solve a sample of test items while thinking aloud, fill out a questionnaire and participate in a post-hoc interview. The think-aloud are designed to uncover what is going on in a participant's short-term memory before they internalize the experiences and in that way taint the information with their own rationalisations. The questionnaire used is the NASA task load index (NASA TLX), a tool widely used to study workload in person-artefact interactions (Hart, 2006). The questionnaire should be well suited to shed light on the workload involved when solving the test. The interview makes participants give their own retrospective interpretation of what happened during the think-aloud, in order to give an account of their perspective and to clear up any issues.

The set-up roughly correspond to a breakdown of the cognitive lab procedures that is common in the field today, the concurrent and retrospective accounts of what goes on in the cognition of the participants, as well as the usability aspect of taking the test.

4.1.1 Materials

Before beginning the data gathering, the method was piloted, to see whether it was possible to get good, usable data from the procedure, and to check the quality of the data gathering protocol. This led to a number of changes to the protocol that supported the feasibility of the method. The number of items the participants were supposed to solve was reduced, as exhaustion poses a danger to the quality of the obtainable data (Nielsen et al., 2002). Starting from a universe of 40 items distributed across ten testlets, the original intention was to use every testlet and ease the strain on the participants by converting seven of the ten into single item problems. However, the piloting suggested that this conversion did not reduce the strain on the participant noticeably, so a sample of six testlets was selected purposefully to represent both the easier and more difficult parts of the test. Every second testlet in the sample was designated to be a single item problem, and thus got three of their four items removed at random. This leaves a sample of three full testlets and three single item problems. The sampled problems was ordered from easiest to most difficult according to the test set design, and gathered into a booklet. A protocol for the think-aloud was made, so as to ensure a uniform procedure for every participant. The interview was set up to be a semi-structured interview, which is a set-up where an interview guide is created with basic topics and questions, but with the intention of deviating from it as new information surfaces and the theory gets saturated (Creswell, 2012). The actual protocol for the think-aloud can be found in Appendix A, and the interview guide in Appendix C. The think-aloud and interviews were all conducted in Norwegian, and consequently most of the materials were in the same language. The instructions for the test had been translated from Dutch to Norwegian (see Appendix B for the translated instructions). The test itself was not translated, as the instructions were deemed sufficient for the participants to know what to do. The NASA TLX was kept in its original English language. All participants were administered the same booklet of test items. The participants were supplied with pens for filling out the test, with no specific instructions on how to use them aside from what was in the test instructions.

4.1.2 Sample

When looking for participants, internal comparability was emphasised over generalisability. Given that information redundancy increases quite rapidly in cognitive lab studies (Nielsen, 1994), six participants were deemed to be sufficient. A convenience sample of ethnic Norwegian, young adult, social science and humanities students was chosen, of which half were in their first year of bachelor studies, with the rest being in their late bachelor and master studies. The gender balance was equally male and female. Most of the participants reported having some experience with similar tasks or games, but none stood out in this regard. Each participant was awarded for participating with a lottery ticket.

4.1.3 Procedure

As recommended by Fonteyn, Kuipers, and Grobe (1993), participants were scheduled for individual sessions in a quiet and undisturbed environment at the university. Based on the pilot, it was estimated that the think-aloud would take roughly 20 minutes and the interview 10 minutes, with 30 minutes set aside as a margin in case these estimates were wrong. Time was set aside between the sessions, for the observer to write out the notes. After giving informed consent, the participant was given brief instructions on the think-aloud procedure. The participant was then instructed to read the instructions for the actual test itself (Appendix B), with minimal guidance from the observer, before again being reminded of the think-aloud instructions. These was to think aloud while solving the problems, and that the participant was not supposed to converse with the observer. If the participant for some reason stopped talking aloud, the observer would prompt him or her to continue talking. Because of experiences from the piloting regarding comprehension of the audio recordings, the participant was asked to be explicit about what he or she referred to when talking. Immediately after finishing the think-aloud, the participant went on to quickly fill out the questionnaire, before commencing with the interview.

The participants were asked to try to solve the test efficiently, preferably within the twenty minutes specified by the instructions. Sessions that exceeded this time constraint were not aborted, however, as the participant's ability to cope with time pressure was deemed less important than learning how the problems were being solved. Hence, avoiding incomplete data was deemed more important than comparability. The full session was audiotaped, for subsequent analysis.

4.1.4 Reflections on research credibility

Verbalising internal speech always entails some degree of interpretation and restructuring by the person (Vygotsky, Hanfmann, & Vakar, 2012), and this risks affecting both the think-aloud and the interviews. The interviews in particular risks participants rationalizing their thoughts, instead of giving pure and objective access to their cognitive processing. There is simply no way to guarantee that participants are able to put words to their own thoughts, or even understand exactly what is going on in the first place (Nisbett & Wilson, 1977). All the interviews were conducted just after the participants had finished the think-aloud and the NASA TLX, which might influence the trustworthiness of the information, taking into account the primacy and recency effect on human memory (Ebbinghaus, 1913), the richness of the accounts made by the participants might be somewhat biased against the middle testlets, with the effect being especially strong with the participants who spent the longest time on the later testlets, which also was the impression of the interviewer after undertaking the sessions.

The content validity of the cognitive lab could be better, as the sample of items used can neither be said to represent the unrestricted universe of possible abstract reasoning items in this format (there are many possible item features and rules that have yet to be tried out), nor the restricted universe of this test (Kane, 1982).

The results from the think-aloud can be said to have challenges regarding its ecological validity (Cole, 1996) relative to a real test situation. In a real test situation, the participants would be under a much stricter time constraint. The test would likely be administered on a computer, and the participants would of course not be thinking aloud. They would likely be more externally motivated in the job recruitment setting than the participants in the present cognitive lab. Time data from the think-aloud procedure will not perfectly match time data from the real test situation. The patterns uncovered should, however, still be able to shed some light on what is going on when participants interact with the items, providing a basis for operationalising factors to use on real test data.

4.1.5 Ethical considerations

The participants all gave oral informed consent, in accordance with the specifications set by the Norwegian Social science Data services (NSD) (see Appendix E). Files and documents were stored securely in accordance with Norwegian data security laws. The data were anonymised before the analysis, and explicit identifying information was not kept with the data. Some of the participants were aware of each other's participation, so to avoid compromising anonymity I have elected to not use numbered labels when presenting the data. Using randomised numbering of the participants would imply ordering, and since this is not there, they were given random aliases taken from an existing list of names (Meteorologisk Institutt, n.d.). Although the literature shows that there are some gender differences in intelligence on aggregate (see e.g. Halpern, 1997), no major patterns were observed in this small sample. Gendered names were therefore given at random, as to insure maximum anonymity.

4.2 Analysis

The data were analysed by categorising themes and concepts that arose when the participants solved problems. The analysis evolved with emerging theoretical saturation. Although there were some initial assumptions about what to expect, extracting the categories was an iterative sense-making process, where the data sources had to be continuously reassessed during the analysis. Original expectations were reshaped during the cognitive lab situation, and theory was later revised during analysis of the recorded materials.

4.2.1 Think-aloud

Getting thoroughly transcribed and reliable verbal reports (as prescribed by Ericsson & Simon, 1993; Ohlsson, 2012) from the think-aloud procedure was deemed to be infeasible at present, taking time constraints on the project into account, and also impractical given the exploratory . Based on impressions from the initial lab situation, it was decided that the ideal way to analyse the think-aloud data, was to count the duration each participant spent on the problem and items in the test, in addition to any other behaviour at the given time, essentially turning the think-aloud into a poor man's eye-tracker.

4.2.2 Interview

Meaningful information from the interview were analysed, and coded into categories (as well as some information from the think-aloud), with representative or interesting information translated and written up. Because of the time constraints on this project, the data were not transcribed verbatim, but time frames were electronically coded into categories on the audio files, with select parts being written up in English and presented. The quotations used in this document were translated by the author. Square brackets, [ ], have been utilized to fill in information about context, to skip digressions, or to condensate and clarify utterances that would have been impractical to write up directly. All quotations have been written down with varying degrees of interpretation, in order to insure good readability, and are thus not verbatim transcripts of the audio data.

4.3 Results

This section contains time counts and observational data from the think-aloud, excerpted parts from the interviews, and presentation of results from the NASA TLX as well as from the test itself. The order of presentation corresponds to the chronology of the actual procedure.

4.3.1 Observations

In the think-aloud, it became apparent that the number of figures in each problem set challenged the decoding ability of the participants. When turning the page to a new problem, several participants seemed to be taken aback when faced with an escalating number of figures. They seemed quite overwhelmed by the sheer number of figures, even before they had started analysing the problem in detail. The number of figures had not originally been considered a noticeable theoretical feature in itself, but now had to be considered further as a radical at the problem level.

Some of the participants did manage to differentiate between addition or subtraction of a line and rotation of the whole figure, but most had trouble with pinpointing exactly what was going on when the figures contained no obvious clues to pinpoint rotation, like imperfect shapes or similar. In these cases, the participants mostly either interpreted the rule as a rotation or a double rule. Some instances with multiple rules altering similar figure features, like lines pointing in different directions, seemed to cause substantial frustration, with the perceptual saliency of the characteristic being especially important. Participants generally struggled to solve the whole problem in these instances, with several exclaiming “I really don’t understand what this button does, but I’ll give it a shot anyway” (Synne), or the like.

With the more salient figure features, like shapes and colours, rule finding strategy seemed to be a quite straight-forward procedure of matching the changes to buttons. As the participants encountered figure features that stood less out, however, they seemed less able to perceive the changes right away, thus having to resort to a rule finding strategy of cycling through any rule that might seem relevant to the problem. In addition, as they continue to solve the test, more rules were introduced, increasing the problem space. As more rules were introduced, the participants had to evaluate an increasing number of possible solutions.

Two of the six participants chose to take notes roughly from the start, and the rest either waited to do this until at least half-way into the test, or they never did at all. Ole and Petra started taking notes at the kick-off (spending more time and getting the most correct), while Tor and Nina never wrote anything down (spending the least amount of time, but getting a lot of incorrect answers). Roar and Synne started taking notes along the way, the former at problem F, and the latter at problem C.

The participants turned out to be very dutiful to their instructions of avoiding gambling when unsure. When they had to guess, it was after spending a lot of time and effort trying to solve the problem, and as a consequence, most guesses were partially informed, that is, the participants had solved most of the testlet problem, but were missing one or two rules to completely get it. This was especially obvious in problems with rules that were hard to differentiate, where participants often were unsure about their choices.

4.3.2 Results from the think-aloud

Coding tree

There was no prior established criterion for distinguishing between which observed instances should be labelled as *item oriented* or *problem oriented* (Figure 4.1). This distinction was made based largely on the degree of pragmatism the participants showed (that is, more interested in just the item-relevant buttons), what elements they seemed to focus their thinking around, whilst also taking into account notes from the observer on where their attention seemed to be.

Table 4.1 shows the number of codes that were assigned to each of the nodes in the model, as well as the number of sources (that is, participants) that had codes associated with each node. As we can see, all nodes from the think-aloud had at least one code assigned to every participant, except for the item-orientation node, in which one participant, Petra, was exceptionally diligent and stuck to her starting strategy of deciphering each problem completely, paying no regard the attached items until satisfied with the rule derivation. The codes from the child nodes are aggregated into their parent node. It became obvious quite early that the participants not necessarily solved the items of each testlet in order, so some codes were associated specifically with the testlet node, usually when there was no clear indication what item the participant was focusing on. This is shown in Table 4.1, where the number of codes for some of the testlets surpass the sum of codes for the items belonging to them. The consequences of this is further illustrated in Figure 4.5 and might affect how some of the results, most notably Figure 4.4 should be interpreted.

Figure 4.1: Code tree from the think-aloud

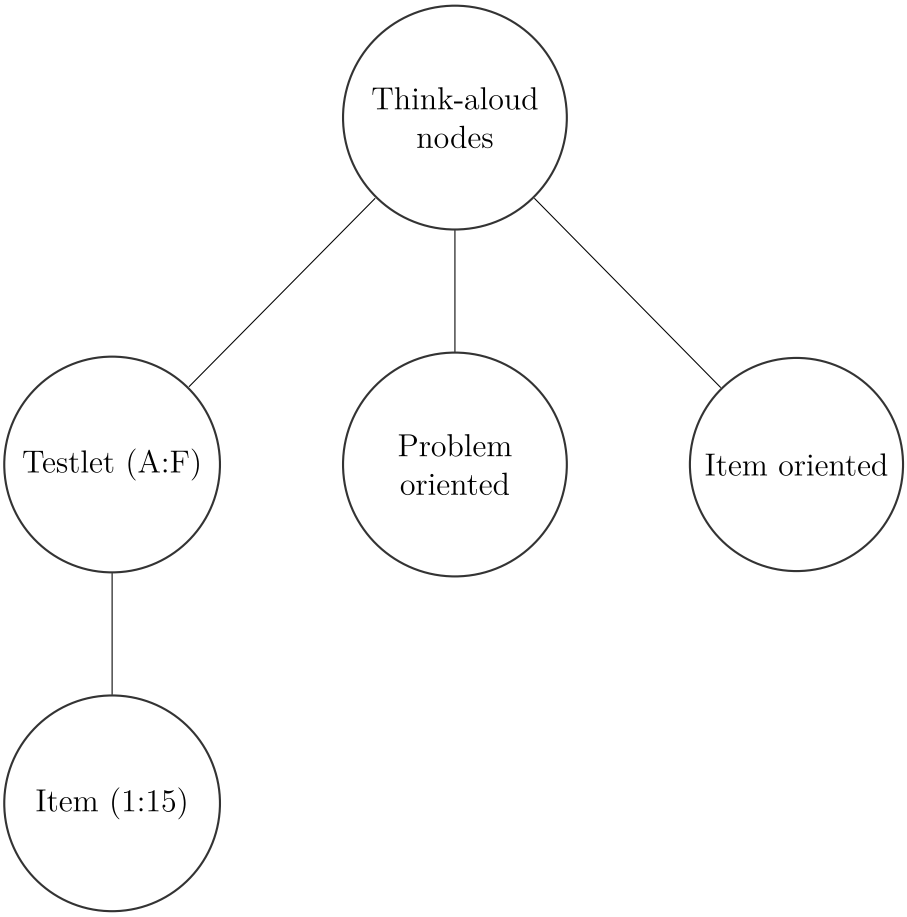


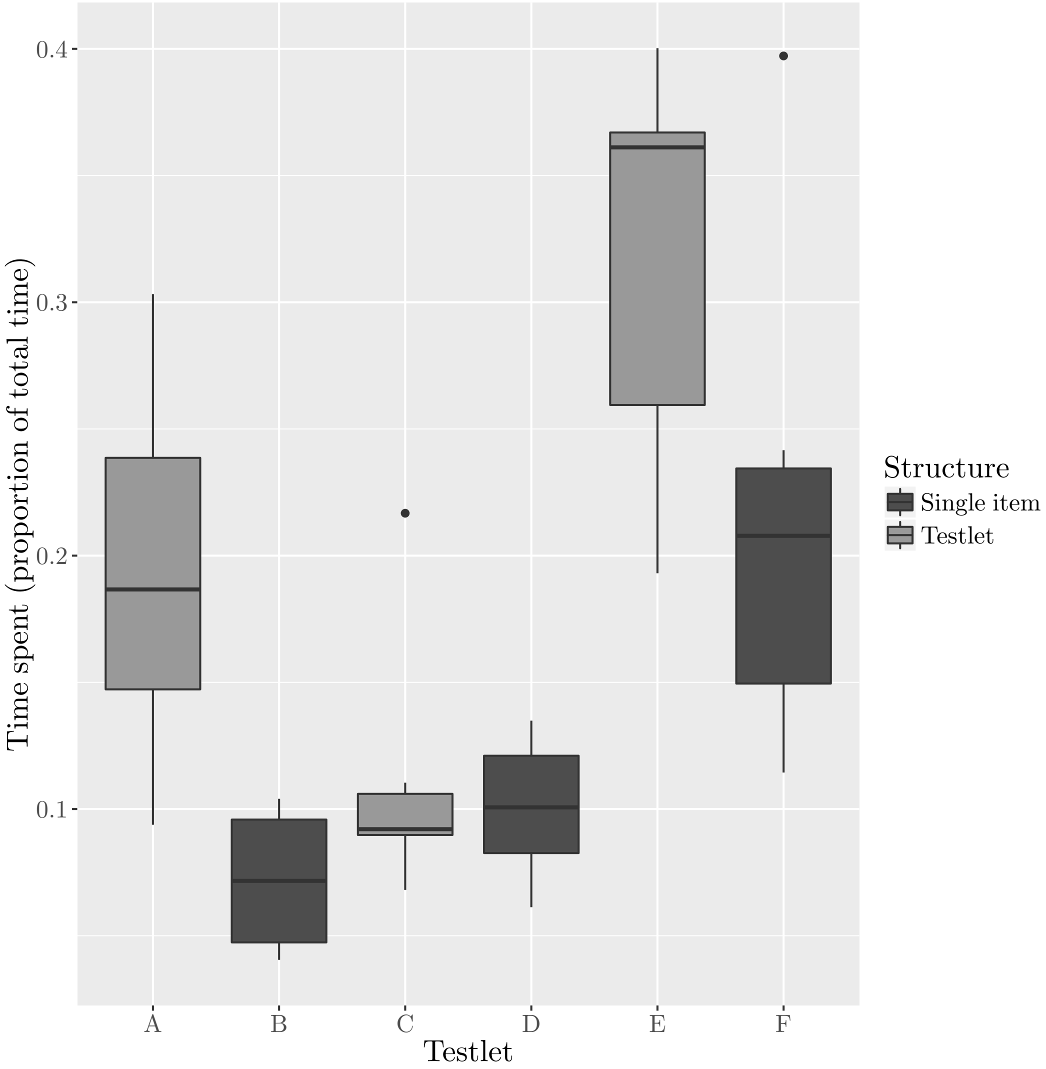
Table 4.1: Frequency table of the number of codes used during the think-aloud procedure.

	Grandparent node	Parent node	Node	Codes	Sources
1	Think-aloud nodes	A	1	6	6
2	Think-aloud nodes	A	2	8	6
3	Think-aloud nodes	A	3	6	6
4	Think-aloud nodes	A	4	6	6
5	Think-aloud nodes	B	5	6	6
6	Think-aloud nodes	C	6	6	6
7	Think-aloud nodes	C	7	6	6
8	Think-aloud nodes	C	8	6	6
9	Think-aloud nodes	C	9	7	6
10	Think-aloud nodes	D	10	6	6
11	Think-aloud nodes	E	11	7	6
12	Think-aloud nodes	E	12	6	6
13	Think-aloud nodes	E	13	6	6
14	Think-aloud nodes	E	14	7	6
15	Think-aloud nodes	F	15	6	6
16	Think-aloud nodes	Think-aloud nodes	A	31	6
17	Think-aloud nodes	Think-aloud nodes	B	8	6
18	Think-aloud nodes	Think-aloud nodes	C	28	6
19	Think-aloud nodes	Think-aloud nodes	D	9	6
20	Think-aloud nodes	Think-aloud nodes	E	31	6
21	Think-aloud nodes	Think-aloud nodes	F	10	6
22	Think-aloud nodes	Think-aloud nodes	Item oriented	28	5
23	Think-aloud nodes	Think-aloud nodes	Problem oriented	31	6
24	Think-aloud nodes	Think-aloud nodes	Think-aloud nodes	176	6

Note. Table of descriptives of the number of codes used during the think-aloud procedure. The text columns correspond to the model in Figure 4.1. The code and source numbers are aggregated into their parent node. Some sequences were coded directly into the parent nodes, thus making the total bigger than its parts. The maximum number of sources is six

Time spent

Figure 4.2: Boxplot of time spent on each testlet problem.



.

Note. The colour fill indicates whether the problem in question was a single item or a full testlet

The participants generally spent a lot of time on the first, as well as the last two problems in the test, while the middle problems performed roughly equivalent of each other. Problem E in particular stands out as taking a lot more time than the other problems (Figure 4.2). The observer noted a warmup effect, in that the participants were somewhat inefficient when they started out, as they had to get accustomed to the format of the test, and this can also be seen from the figures.

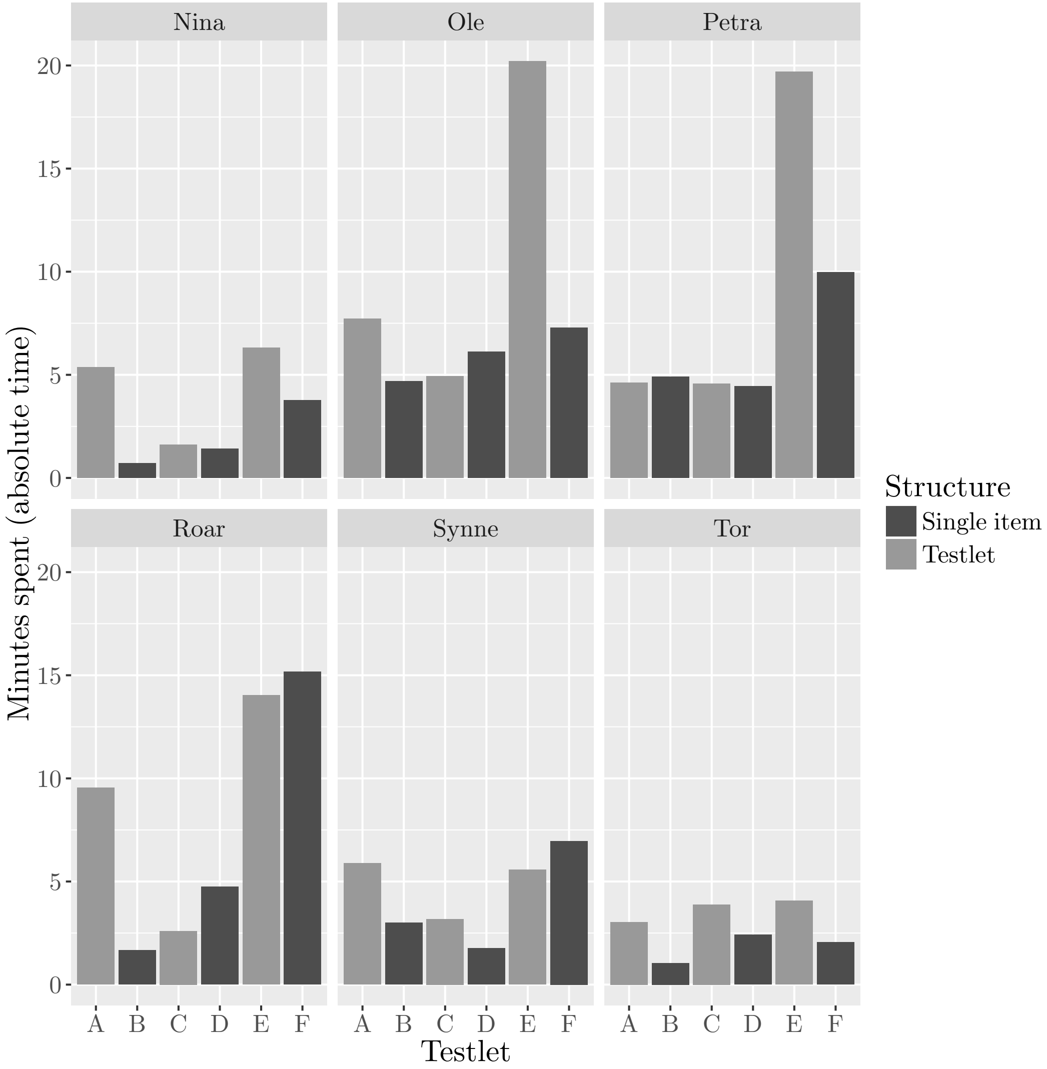
Figure 4.2 shows a substantial amount of variation in how much time the participants spent on solving the test. All of the participants but Tor and Petra spent substantially more time on the first testlet problem (A) than on the second (C), despite the latter being expected to be more difficult (Figure 4.3).

Figure 4.4 shows time spent on each item, grouped by which testlet problem they belong to. The first two items within problem A have more time coded to them than the last two. Problem D, E and F has huge variation, and generally more time spent on each item. This is not surprising, given that these problems should be considered the most complex. In other words, the participants either cracked the code, spent outrageous amounts of time solving it, or they guessed and moved on. The single items problem (B, D and F) showed a bit higher time spent than each of the items in the testlet problems, but the last two of these varied greatly.

However, this figure is not telling the whole story. We can see that the duration on problem D and F in Figure 4.4 looks quite similar, but there is a substantial difference between problem D and F in Figure 4.2. This discrepancy is caused by the way the variables were created, and if we look at Figure 4.5, we se that there is a lot of time spent on each testlet where the participants were not coded as spending time on any item, which means that they were preoccupied with solving the problem as a whole. Figure 4.2 and Figure 4.5 are also interestingly similar, showing that the testlets the participants spent the longest time on in general, were also the ones that they spent the most time not focusing on the items. Corroborating evidence can be found in Figure 4.6, where the degree of pragmatism showed by the participants was more directly coded for. We see that testlet F elicits more problem oriented behaviour than testlet D. This might actually be because the participants perceived problem F as more intimidating than problem D at first glance, as it has more rules and figures.

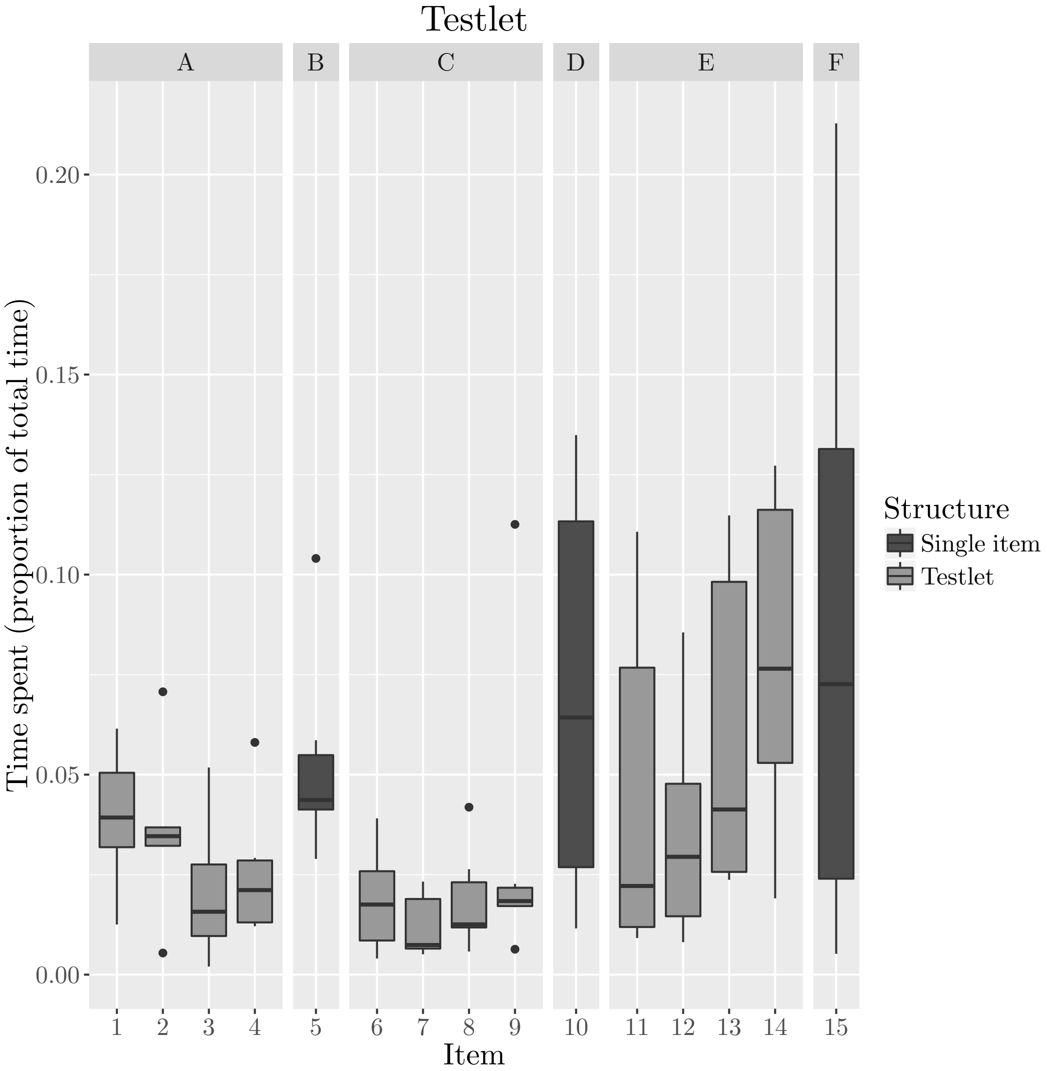
As shown in Figure 4.6 most time was spent focusing on either solving the general problem or figuring out each item. On the earlier items, most of the time was spent solving the problem and then applying the rules to the items, whereas when the test got harder, more time was spent focusing on the items.

Figure 4.3: Barplot of the amount of time each individual participant spent on each testlet.



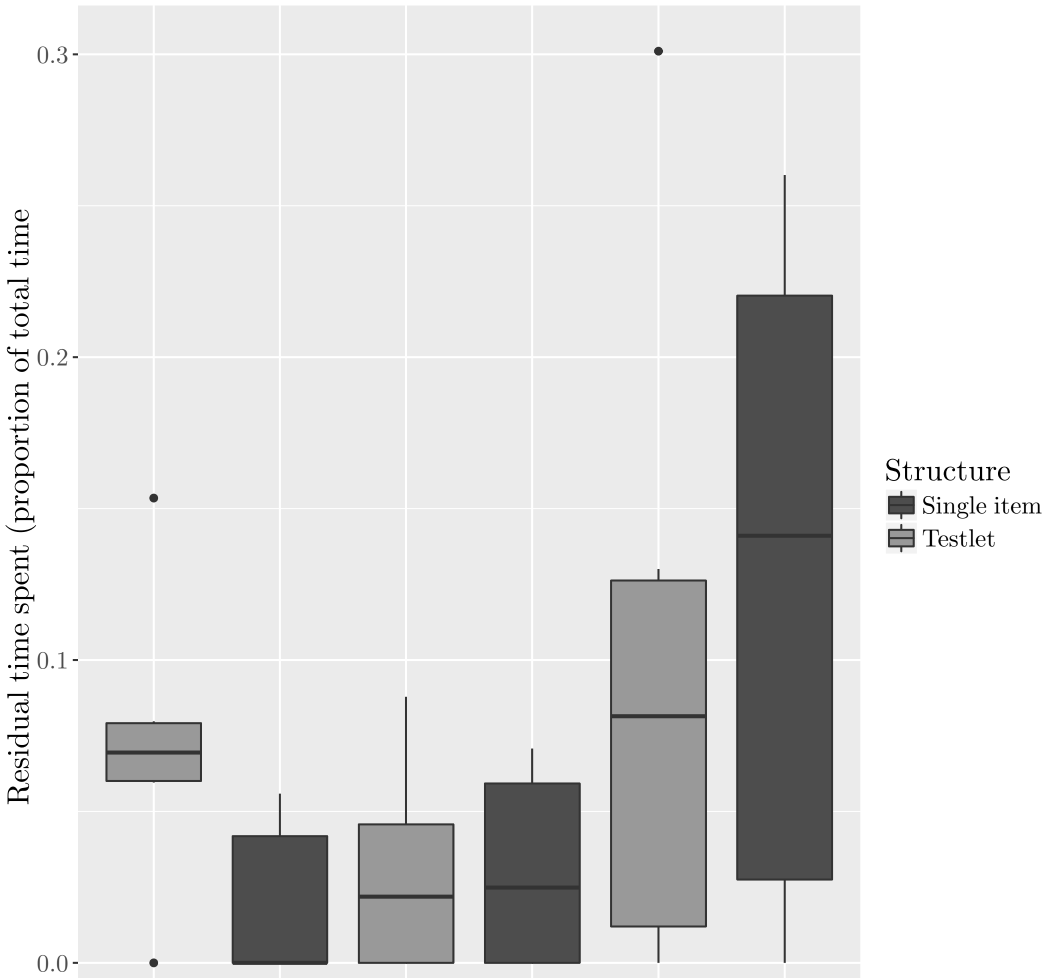
Note. The colour indicates whether the problem in question was a single item or a full testlet

Figure 4.4: Boxplot of time spent on item grouped by testlet problem.



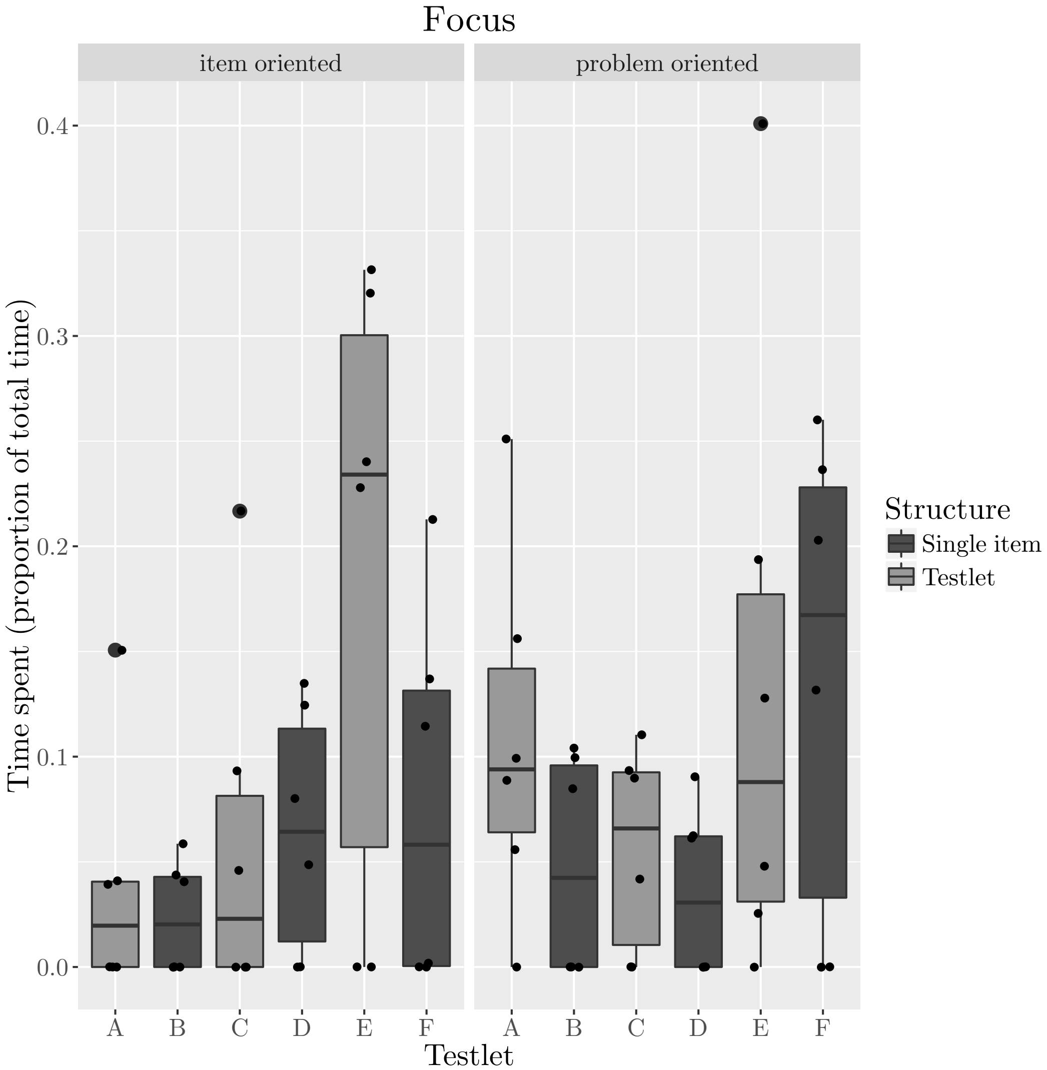
Note. The colour indicates whether the problem in question was a single item or a full testlet.

Figure 4.5: Boxplot of residual time in each testlet not accounted for by aggregating the items.



Note. The colour indicates whether the problem in question was a single item or a full testlet.

Figure 4.6: Boxplot of time spent focusing on either solving the whole problem or each item.



Note. The colour indicates whether the problem in question was a single item or a full testlet. As can be seen from the datapoints, not all participants had every focus category coded to every testlet category.

The unwillingness of some of the participants to guess can contribute to explaining the large amount of time spent on testlet E and F in Figure 4.2. They tried hard to find the right solution, but were unable to fully decipher the problem, hence the large amount of time spent focusing on the items (Figure 4.6. Figure 4.4 gives corroborating evidence of this behaviour, as the later items in testlet E has more time coded to them than the early ones, suggesting that most of the participants never really solved the problem fully before attempting to deal with the items.

4.3.3 Usability aspects and performance

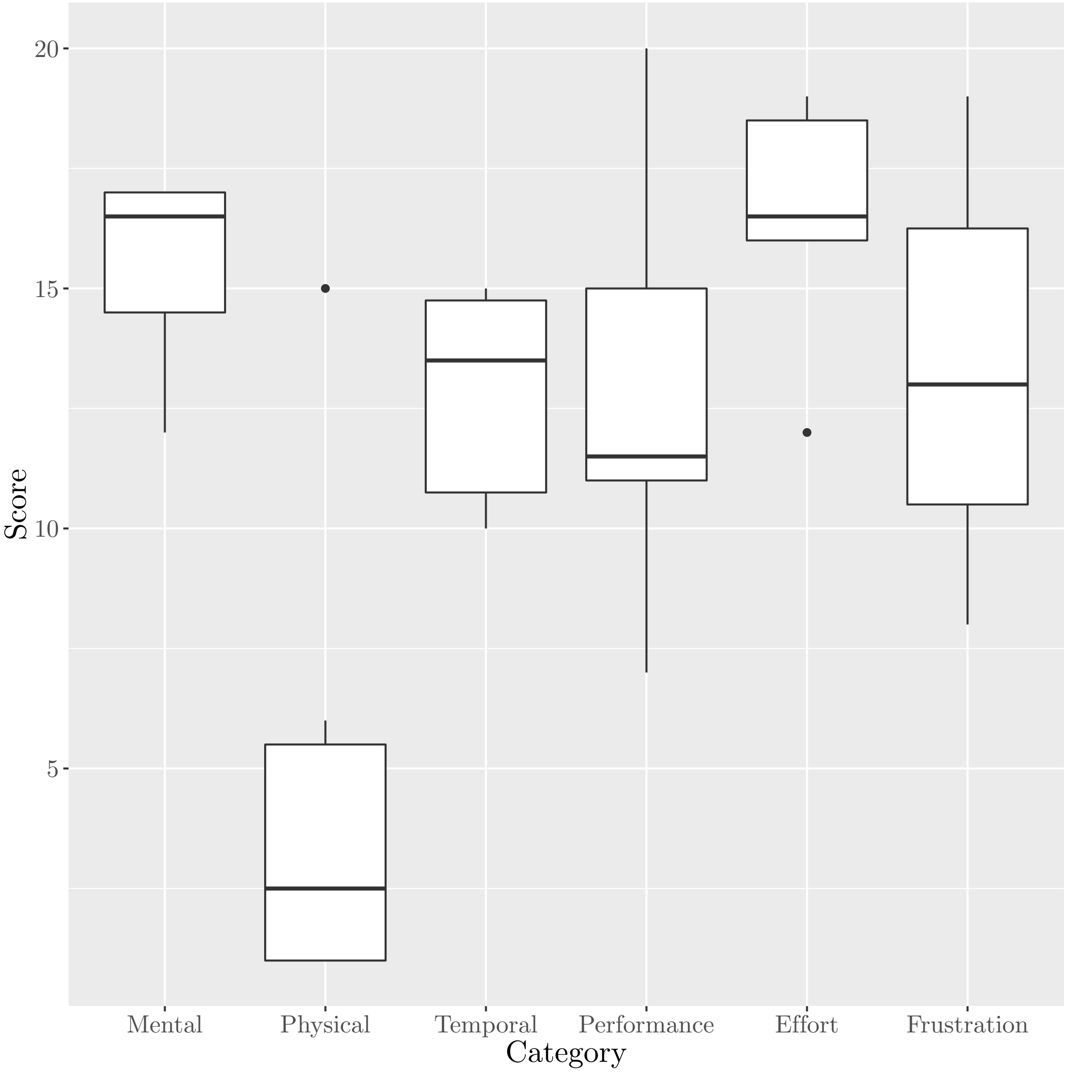
NASA task load index (NASA TLX)

Most participants reported feeling that solving the test was quite mentally demanding, with twelve as the lowest rating and an average of 15.5. Most also clearly felt a time pressure, with every participant rating ten or higher, and an average of 12.83. Physical strain was generally ranked low, with the notable exception of one participant (which was coincidentally also the person spending the shortest amount of time solving the test, as well as reporting the best self-perceived performance) (see Figure 4.7). These results are in line with Hart and Staveland (1988) who emphasise mental and temporal demand as the most important facets of the NASA TLX for simple cognitive tasks with time pressure. Scores on the performance factor centre around the middle of the scale, with quite a lot of spread. All but one of the participants report having to put in very much effort (the outlier still being above the middle of the scale). Their frustration was rated from the middle to the high end of the scale, also here with a fair amount of variation. The actual questionnaire can be found in Appendix D (please note that the performance factor has been inverted in Figure 4.7 for ease of interpretation).

Actual test performance

The three participants that spent the longest time on problem E and F in Figure 4.3, also reported amongst the lower self-reported performance (albeit this not reflected by actual performance). As we can see in Table 4.2, there is a positive relationship between

Figure 4.7: Boxplot of the results from the NASA TLX.



Note. The performance factor has been inverted for ease of interpretation.

the total time spent and the total score, with the two participants that spent less than twenty minutes achieving the lowest number of correct items. The performance variable form the NASA TLX seems a bit counter-intuitive, where we see a negative relationship between that the participants' rating of their own performance and their actual score. In other words, struggling with the problems for an extended time, negatively affected their feeling of success. It is reasonable to assume that this results from some of the participants taking the "no gambling" instruction overly seriously, thus breaking the time constraint to work on the problems until they were quite certain of the right answer (as we for instance can see from the interviews with Quote 19).

It is worth to note that Ole and Petra, who took notes from the very start, spent the longest time and *and* had the most items correct. Tor and Nina did not take notes, spending the least amount of time, but getting the highest amount of incorrect answers. Roar and Synne started taking notes at problem F and problem C respectively, and spent a fair amount of time, getting a fair amount of items correct. On a side note, Roar in particular is an interesting case, as the problems he spent the longest time on was also the ones with the fewest correct items.

Table 4.2: Table of total test score, duration and Performance

	Participant	Total score	Duration	Performance
1	Nina	4	17.70	11.00
2	Ole	14	55.28	11.00
3	Petra	13	49.22	7.00
4	Roar	8	38.18	12.00
5	Synne	9	28.83	16.00
6	Tor	7	17.92	20.00

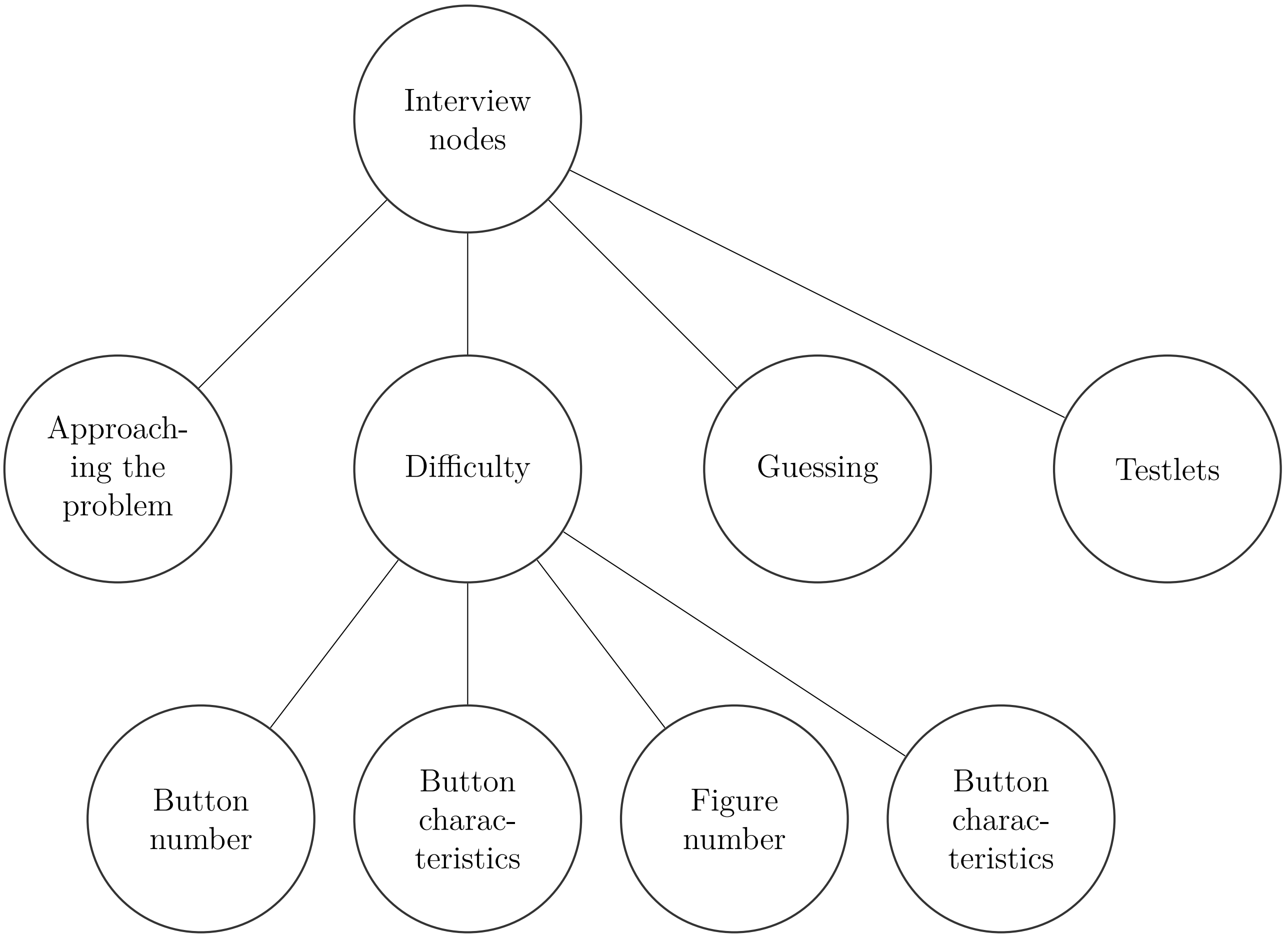
Note. Maximum attainable score on the test was 15. total time spent from the think-aloud is in minutes. The self-reported performance variable from the NASA TLX is inverted of the actual questionnaire.

4.3.4 Results from the interview

Coding tree

The majority of categories from the protocol (Appendix C) were present in all interviews. The interviews also included demographic questions, but the results of these are summed up in the methods section.

Figure 4.8: Code tree from the interviews.



Approaching the problem

The participants did follow a sequence when they approached the problems. Focusing on the buttons as key to understanding the problems, the most basic procedure for the easier items is to identify what buttons stood out (see e.g. Nina in Quote 1).

> **Nina:** First I concentrated on the button, I did not try to identify the figure features. I looked at the base figures and the end figures, looking at what changed, and then going back to the starting point. The objective was to get a grip on the F-buttons ... [to compare and see what changed]. (1)

Table 4.3: Table of descriptives of the number of codes used during the interviews.

	Grandparent node	Parent node	Node	Codes	Sources
1	Interview nodes	Difficulty	Button characteristics	3	3
2	Interview nodes	Difficulty	Button number	8	4
3	Interview nodes	Difficulty	Figure features	14	5
4	Interview nodes	Difficulty	Figure number	8	5
5	Interview nodes	Interview nodes	Approachig the problem	7	5
6	Interview nodes	Interview nodes	Difficulty	34	6
7	Interview nodes	Interview nodes	Guessing	5	5
8	Interview nodes	Interview nodes	Testlets	11	5
9	Interview nodes	Interview nodes	Interview nodes	57	6

Note. The text columns correspond to the model in Figure 4.8. The code and source numbers are aggregated into their parent node. One sequence too general for the child nodes were coded directly into the difficulty node, thus making the total bigger than its parts. The maximum number of sources is six

The participants then had to derive the rules by comparing the column of basic figures with the end figures (Tor gives a quite clear account on how he proceeded in Quote 2).

> **Tor:** If one button was present in all the rows, and one figure characteristic changed in all of them, then I knew that those must be connected. And then you compare the other buttons and figure changes, to see which ones belong where. You start with what is similar, and then narrow it down to what is different. (2)

Observing how the participants solved the test, this method was often quite straightforward with the easy items, but as the test got harder and they discovered new rules, they quickly started interpreting the buttons as "doing" something to the figures, and the problem solving adapted to a state where they tried out different rules to see who fit the problem best, in essence breaking it down into comparing different models in a trial and error fashion.

Although this was not obvious to the observer during think-aloud procedure, all of the participants claimed they mostly had the items in the back of their heads when solving the problems. The participants disagreed on whether this was directly useful, with both Roar (Quote 9) and Petra (Quote 8) pointing out the necessity to solve the whole problem to be able to know the whole rule scheme. Synne goes more into detail about how she incorporated the items into her problem solving strategy, by answering the most obvious items first, and only then moving to the less obvious ones (Quote 3). This is also a recognition of the varying likeness of each item to the problem.

> **Synne:** Yes. I started out solving the items from top to bottom, but I found out that it did not work, like "i can not get anything out of it, it can not be solved this simply. I must look at each of the items, and see what reoccurs, so that I can connect them ... which buttons make the changes".
>
> (3)

When there are salient buttons in the problems, this deciphering process is quite uncomplicated, allowing the participant to solve the problem rule by rule. However, whenever there is no one clear rule in which to start out from, the relations between the buttons in the problem become less obvious. In Quote 4, Nina talks about having to juggle many mental representations and compare them without having any certainty of the validity of each bit of theory.

> **Interviewer:** Did it matter, how many different F-buttons were present?
>
> **Nina:** When there were multiple F-buttons in different places, that's when it became difficult, because then I had to choose which ... I don't know if I was supposed to look at one of them specifically, but I started thinking: "which one of these F4-buttons fits the best here?".
>
> **Interviewer:** mhm?
>
> **Nina:** But of course, the more factors in play, the more I had to take into account, which made it difficult. When you only had one, it was a little ... [trailing off], you could look at even more relationships, with one button.
>
> (4)

One take from this participant is some evidence of increase of difficulty. The harder items are presumed to have less independent buttons, so in stead of working with one rule at a time, the participants having one that functions as a key starting point, she has no certainty of any one rule until she understands the whole problem. Paraphrased, the load on her mental capacity increased with the need to process the whole problem.

Testlets vs single items

When asked of the contrast between single item problems and testlets, responses were mixed. Some, like Tor (Quote 5), expressed that they felt the single items were more pleasant, with Tor adding that because you could move on, rather than having to respond to multiple items based on uncertain information.

> **Tor:** It was a bit easier to solve problems with only a single item, because then I was much more confident of my response, but when the problem had several items, I would solve one, and then perhaps find out I had erred on the next, so them I had to go back and think "what did I do here", and then go to the next one and do the same, and if they did not correspond, then I knew I had made a mistake. So it became easier in that I got some feedback, but then again if you discover mistakes in something you thought was right, then you become more insecure. (5)

This view was supported by Petra in Quote 6, who expresses that the full testlet made her focus more on deciphering the problem, and that when faced with single items, the focus became more pragmatic, in a sense that the participant only aimed to figure out enough of the problem to answer the relevant item.

> **Interviewer:** The number of items differed between each testlet. How did you think that worked? How did you change your strategy between when there was few and when there was many [items]?
>
> **Petra:** I was perhaps more focused on that [item], when it was only the one, but on the others I was more focused on moving onwards.
>
> **Interviewer:** So, more focus on the first screen ..?
>
> **Petra:** Yes.
>
> **Interviewer:** ... solving the problem ...
>
> **Petra:** Yes. (6)

Petra was one of the most ardent note-takers, which might have affected how she approached the test. However, others, like Roar (Quote 7) also stressed the importance of deciphering everything in order to be certain of their responses. He started to write the buttons down when solving the last of the sample problems, as the relationship between the buttons in the problem got quite complex. He mentions struggling with deciphering the whole problem, as this is necessary to identify the correct rules, and failing to crack the code in certain testlets could very well explain the amount of time spent and lack of success mentioned in Section 4.3.3. He did not see any reason to adapt his problem solving behaviour to accommodate the differences in structure for the very same reason.

> **Roar:** I felt that it was too much, that I couldn't figure it out.
>
> **Interviewer:** Did that concern all the testlets, or some in particular?
>
> **Roar:** I think I sometimes caused myself some difficulties, because in some [testlets] I saw it afterwards ... I looked at [the problem], I focused on it before I looked at what I was supposed to find out [the particular item].
>
> **Interviewer:** To you think it had any consequences?
>
> **Roar:** I don't know, you had to look at almost all [of the buttons] to eliminate anything.
>
> [going on about the testlets]
>
> **Interviewer:** Some items had also been removed from some of the sets [testlets]. Do you think it had anything to say?
>
> **Roar:** That was actually quite pleasant.
>
> **Interviewer:** You thought it was easier?
>
> **Roar:** I didn't really think about it.
>
> **Interviewer:** Did it have any consequences for how you solved the problems?
>
> **Roar:** No ... no I don't think so. (7)

As stated earlier, Petra was the participant sticking most to her strategy all the way through the test. She was asked whether she tried to get information from other responses within the same testlet, but said that it did not occur to her, at least not consciously (Quote 8). She was one of the participants most heavily coded for problem orientation, with a high success rate but also a lot of time spent, and she used a solving strategy that was close to the rational way.

> **Petra:** I didn't feel that was needed, since I had this [pointing towards the problem], since it should apply ... perhaps I should have seen them more in relation with each other, I don't know, I didn't do that. (8)

The importance of understanding the problem as a whole resurfaced in several interviews, with Roar summarising it in one sentence:

> **Roar:** I am having a bit of trouble with this. You cannot just look at F6 and F8: here's F6 and F8, you also have to know what F5 and F7 does to the figures. (9)

Some interesting insight about remembering the problem rules surfaced. When asked, Tor reported not having any issues with resetting his thinking when moving to a new problem. He did not take notes when solving the test, and in order to remember the problem rules, he used his own reasoning on previous items to inform the solving of subsequent items. Tor also reported feeling more insecure when solving the full testlets than the single items, because of the previously mentioned memorizing strategy.

> **Interviewer:** [asking about the problem rules] Was it difficult to reset between the problems?
>
> **Tor:** No, because I did not remember what I had thought. If I had a problem with four items [a full testlet], then I had to think back to what I had thought on the item above, so when I got to a new page [...] then I felt that I forgot them at once.
>
> **Interviewer:** [following up on the testlet information] ... you took information from items you had already answered?
>
> **Tor:** Yes. Or, I at least tried. If I was stuck, I went back [to a previous item] and thought "what did I think here? I have thought this and this, then that must be right", and if I thought something new at the last item, then I had to go back and "was it really like that"? And if it was, then both were right, and if it was not, then both were wrong. (10)

He was echoed by another participant, who took the feedback problem even further:

> **Nina:** Is it so that there really is no right answer, and so now I just made this all up? (11)

Radical and incidental elements

When being asked directly of the facets determining difficulty, the participants, like Tor in Quote 12, tended to bring up the number of figures in the rows of the problem screen.

> **Tor:** When multiple changes occurred to that single square or circle, I didn't think it was too difficult, but when it was like: circle square circle square, then I thought it was difficult.
>
> **Interviewer:** Now you are talking about the number of figures?
>
> **Tor:** Yes [...] the more figures, the more difficult, really. (12)

Some, like Synne (Quote 13) also mentioned the specific features of the different figures, focusing on the difficulty of and the characteristics of the rules involved.

> **Synne:** it was not really the number of figures that did it [made the problem difficult], but rather what it changed within the figure, if the figure changed, or if the number of lines changed. It was not the number of figures that caused difficulties, but rather what changes occurred. (13)

This emphasis on the actual changes was corroborated by most of the participants, and most of them struggled in one way or the other with identifying some figure characteristic changes (in essence the problem rules). The most notable cases mostly involved problems containing rules that adds or removes lines within each figure (like Quote 14 and 15), which was often confused with rules where a line shifted 90 degrees.

> **Interviewer:** what do you think about all the different [item features] changes, were some more difficult than others?
>
> **Synne:** [referring to a button adding lines in one of the problems] ... I do not know what it did at that last item ... I never managed to understand what it actually did. What [one specific button] did at the last item. I think I got the other ones, but [that button] does something I could not quite get. (14)

The button in question controlling a rule adding or removing one line was present in testlet F. Synne actually managed to answer the item correctly (as well as three out of four in testlet E containing similar rules), but clearly utilized an effective guessing strategy when not sure of the rational solution. Petra (Quote 15) also hit an obstacle when confronted with the lines rules, admitting to never understanding them in neither problem.

> **Petra:** I got very confused when more elements were added, that one had to take into consideration, the crosses and ...
>
> **Interviewer:** The figures on the side, or [the buttons] in the middle [of the problem]?
>
> **Petra:** No, within the circles or squares itself, where there was ...
>
> **Interviewer:** Here on the side, then, on the end figures and base figures?
>
> **Petra:** Yes ... in that crosses and lines started appearing ... I didn't manage to see the connection there.
>
> **Interviewer:** Yes, I noticed that you struggled a bit on [that one testlet][...] the one where I interrupted you [...] it was the crosses [rules involving lines] that got you?
>
> **Petra:** Yes.
>
> [...]
>
> **Interviewer:** But did you think that this testlet was more difficult than the one after?
>
> **Petra:** No, I had in a way just accepted that I would not be able to figure out the crosses [rules involving lines], so I cared less about getting them right.
>
> **Interviewer:** So you had in essence given up a bit on that last one?
>
> **Petra:** Yes, or I decided to rather focus on understanding [the rest of the problem], than the crosses [rules involving lines] that I did not understand in the last problem.
>
> **Interviewer:** So you utilized some way of educated guessing instead.
>
> **Petra:** A bit, maybe. (15)

When two line rules (horizontal and vertical) are present within the same problem, they lose salience, and can easily be confused with rotation or just a line switch, and also requires the person to be extra thorough when deciphering it.

Radicals regarding the specific items were not much emphasised explicitly by many of the participants, and they were mostly preoccupied with deciphering the problems. Synne was one of the few who briefly touched upon it (Quote 16). A six-sided figure (distinguishing itself from the regular squares and circles) that only appeared a few times within the sampled items contributed to some frustration.

> **Synne:** I thought the six-sided figure was a bit difficult, or irritating, but that was probably because it only appeared a couple of times, and I did not see it enough ... it is a special shape. (16)

For the instances where Synne was presented with a figure characteristic in the problem screen that was lacking in an item (and vice versa), she was forced to adapt her reasoning to a context that was slightly different from that for which the rule was derived. The six-sides could help distinguish rules involving figure rotation from rules involving adding and subtracting lines from the interior of the figures.

When it comes to item features that ought to be incidental, the button characteristics serves as an obvious example. Several participants had to ask whether the figure rules are transferable from one problem to another. There are identically named buttons across some of the testlets, which causes confusion about whether the testlets actually were independent. Most participants consequently referred to the buttons by name throughout the whole session, although this effect might have been amplified by their instructions to think aloud. It could however be a distraction, which was in fact brought up by Roar during the interview, where he complained that his reasoning was disturbed by the meaningful information embedded in the problem (Quote 17).

> **Roar:** It was harder than I had thought. It was difficult with the buttons. I have taken some IQ tests at Mensa and such, and those times I have gotten pretty decent scores, but I feel that the patterns in their tests have been much easier to catch. I got confused by the buttons [in this one].
>
> **Interviewer:** You got confused by the buttons, not the figures?
>
> **Roar:** I probably should't have looked at the buttons [thinking] yes, I had to do that. (17)

Guessing

All the participants admitted to having guessed at some point, but in this cognitive lab it seemed to be mostly made on an informed basis, where the participant had already managed to figure out some or most of the problem.

> **Tor:** There were situations where I reasoned that several responses could be right, and if I could not be sure, then I had to guess.
>
> (18)

Ole discloses that he had to guess if he was unable to decipher the whole problem, but that he in those cases used heuristics in order to attempt to get the problem right (19). Again the culprit is the problem with multiple lines.

> **Interviewer:** Did you have to guess sometimes?
>
> **Ole:** I made some qualified guesses ... one could question whether they were qualified, though. I tried not to gamble, like I was instructed: "no gambling", or I tried to avoid it in any case. When you are totally lost, then you have to make a choice, and I tried to make for the option with the highest number of right elements.
>
> [...]
>
> **Interviewer:** Was there something special that characterised the instances where you had to make an informed guess?
>
> **Ole:** For example that I did not understand what one particular button did, like those streaks or lines and stuff, I never really got those.
>
> (19)

4.4 Summary

It is reasonable to assume that real test takers, when being stuck on a problem, would resort to guessing when they realise they are unable to decipher it cleanly. Rules in the same problem that are very similar would probably be prime causes, as well as information dense problems where the button configuration is more complex.

It could be imagined that the full testlets would somehow force the participants to focus more on a deeper understanding the problem, as more hinges on getting the rules right. Some participants with a less thorough strategy were using items in the same testlets as reference points and sources of information when solving the test, in essence structuring their problem solving strategy around the entire testlet. This seems to confirm some relevance of the testlet structure for modelling of the assessment.

In the interviews, the participants reported feeling stressed about having to solve entire testlets based on insufficient understanding of the problem rules. This is corroborated by the NASA TLX, where we see that the test was quite demanding, and the participants felt decent amounts of frustration when solving the test. That most of the participants not necessarily related to performance, and this could be explained by the test being somewhat confusing, and that it is no way to straightforwardly know the right solutions.

4.4.1 Radicals and incidentals

The number of figures in each problem most likely contributes to multiple categories of radicals. In addition to enwidening the rule space and adding possibilities for more complex and abstract rules, both the think-aloud and the interviews indicate that more figures add to the information load of the problem.

The number of unique buttons in each problem most represent the sheer quantity of rules. Rationally, this facet is an important radical, and this was also the impression during the think-aloud. Interestingly, most of the participants played down the importance of the number of buttons when asked about it during the interview. There can be several explanations for this: a) they might have misunderstood the questions, referring to the raw quantity, disregarding uniqueness, b) other radicals were more salient, stealing their attention, c) lack of variation in number of rules made them neglect it as a cause, d) a pragmatic approach, where the buttons were simply perceived as a way of ordering the figure changes, or e) in this context, the quantity of rules is actually a weaker radical than expected. Anyhow, the urge of the participants to write down the button rules, suggest demand was indeed put on their working memory capacity. This speaks, in line with theory, to the merit of the number of rules as a radical.

The additional figures seemed to increase the load on rule finding by increasing the amount of cross-checking needed for each hypothesized rule, as it has to fit every single figure in that row. Having more figures in each row also allows introduction of more abstract rules into the rule space of that problem.

Button rules involving figure features in the same problem screen that are too perceptually similar, specifically the double line rule, confuse the participants.

Participants sometimes asked for feedback during the think-aloud. Later problems builds on knowledge acquired from earlier problems, as the participants are gradually introduced to new rules as they solve the problems, and later on they are required to be able to finely distinguish these from each other. If they misinterpreted the earlier problem, they were at a disadvantage when trying to solve later problems.

When higher order rules, like reordering or rotation, are present in the problem, they are easy to confuse with lower order rules, especially when the figure features are less perceptually salient. If one rule is misinterpreted or wrongly applied, the like one is often also wrong, risking that the whole testlet is answered incorrectly. There was however some indication that the rotation rule, at least the more visually salient variants, would be more tangible than theoretically expected.

Chapter 5

Validation of the item model

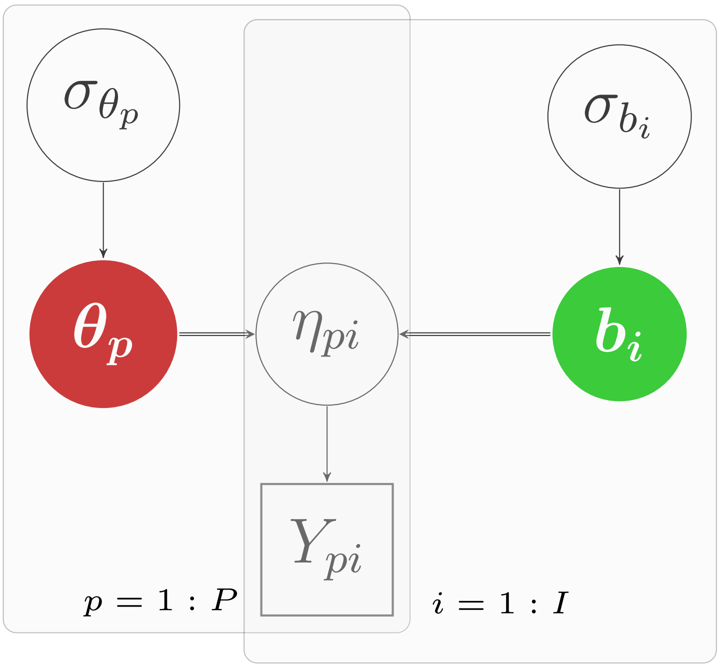
To be able to generate a tentative item model, the knowledge from the cognitive lab, will be put together with a logical-rational task analysis of the test. This will then be put to the test using explanatory IRT. First, the statistical procedures briefly presented in Chapter 1 will be elaborated on in detail, as well as the descriptive NULL model. Thereafter the item model itself - the research hypothesis of the study - is synthesised, with the operationalisation and results of the different explanatory models to follow in short order.

5.1 From Descriptive to Explanatory Item Response Models

5.1.1 Descriptive NULL model

Let Y_{pi} be the item response of person p ($p = 1, 2 \ldots P$ with P being the number of persons in the data: 6519) on item i ($i = 1, 2 \ldots I$ with I being the number of items in the test: 40). In item response theory the conditional probability of a correct answer is modelled as a function of characteristics of the persons on one side and properties of the items on the other side. Both sides come together in the so-called predictor component

Figure 5.1: Graphical representation of the NULL model.



Note. The predictor component η_{pi} of the item response Y_{pi} is a function of the person ability parameter θ_p and the item difficulty parameter b_i, each modelled as random effects following a distribution with variance across persons $\sigma^2_{\theta_p}$ or across items $\sigma^2_{\beta_i}$.

η_{pi}, which under a 1-parameter logistic model (see Equation (5.1)) takes on positive values if the person's ability θ_p exceeds the difficulty b_i of the item, and negative values the other way around. In the positive case, the probability of responding correctly will be higher than 50%; in the negative case, the probability of responding correctly will be lower than 50%, and finally the probability of responding correctly will be exactly 50% when the item's difficulty matches the person's ability (i.e., $\eta_{pi} = \theta_p - b_i = 0$).

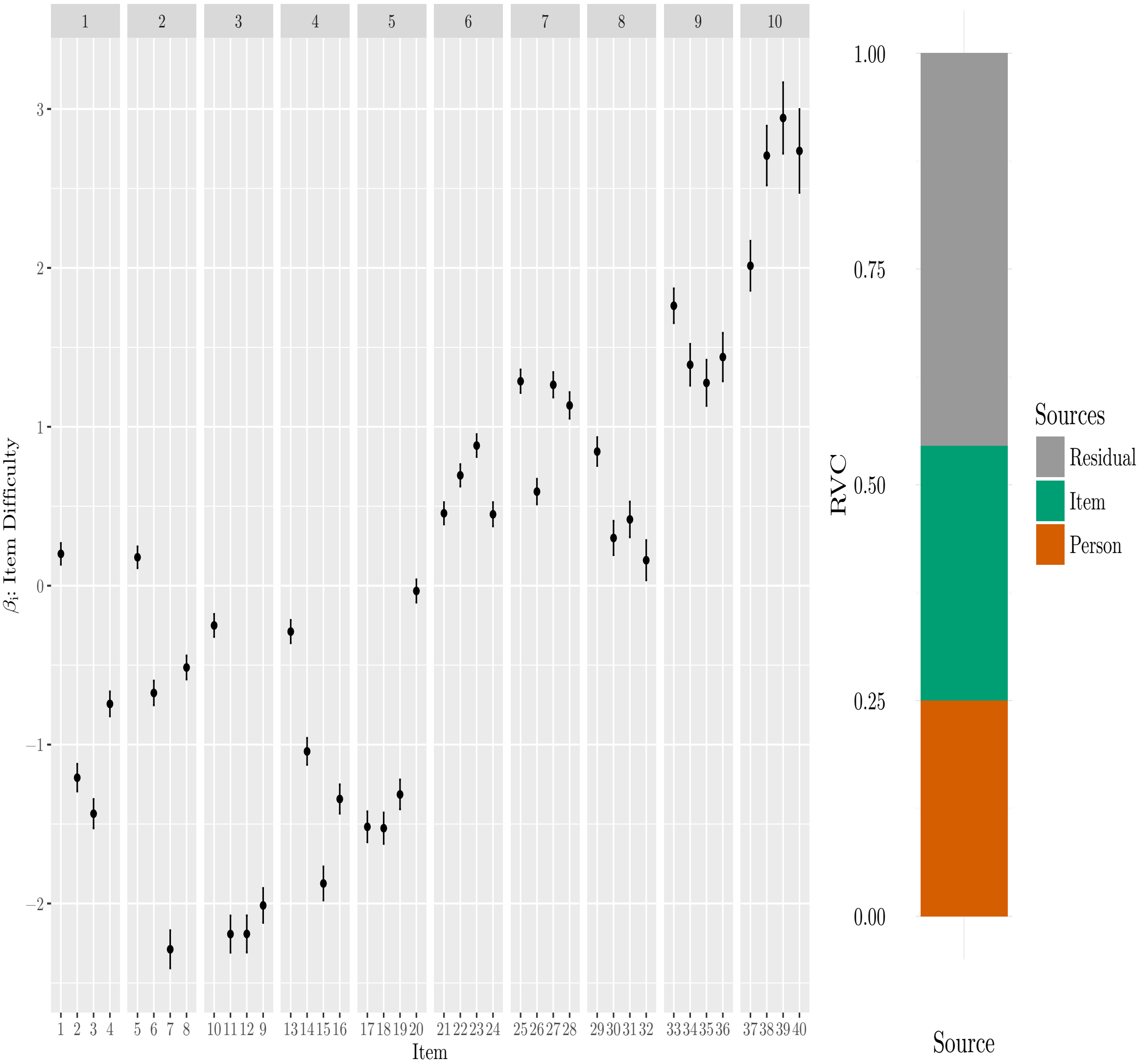
$$Pr(Y_{pi} = y_{pi}|\theta_p) = \frac{\exp[y_{pi}(\theta_p - b_i)]}{1 + \exp(\theta_p - b_i)} \tag{5.1}$$

In the diagram in Figure 5.1 the left side represent the person layer in the model, the red circle being the person unit, and the right side represent the item layer with the green circle being the item unit. A response is the result of a crossing of the two units, person and item, and as such this structure is sometimes called a cross-classified multilevel structure (De Boeck et al., 2011). Responses from the same person will hang together more strongly than responses from different persons, and at the same time, responses on the same item will hang together more strongly than responses on different items. Response variation between persons or between items is each modelled as random effects following a distribution with variance across persons $\sigma^2_{\theta_p}$ or across items $\sigma^2_{\beta_i}$. This so-called random-person-random-item response model can function as a descriptive baseline model.

The left panel of Figure 5.2 plots the 95% confidence intervals of the estimated item difficulties b_i as a function of the item position i in the test. The top-horizontal axis indicates which problem testlet the item is part of. We can see that the first five testlet problems were fairly easy, with most of the persons managing to get them correct. From problem six and onward, however, the probability of getting the items correct diminish quite a bit, and especially the last item is very difficult. The spread within each problem also seems to be greater for the early problems than the later ones. Take for example problem two and three: compared with the rest in their group, the third item in number two seems to be substantially easier, while the first item in problem 3 seems to be substantially more difficult. Differences like these most likely owe to factors pertaining each individual item, independent of the problem factors that are the focus of this study. It is easy to notice that the confidence intervals get increasingly wider the later you get in the test, and this is partially a consequence of less available observed responses on these items due to not all the persons managing to get all the way to the end of the test, but also a consequence of the more extreme difficulty of these items.

The right panel of Figure 5.2 plots the relative variance components in the random-person-random-item response model. These variance components allow to assess to what extent variation in response can be contributed to either the person or item side. Residual variation can then be seen as unknown idiosyncratic influences due to particular interactions between specific persons and specific items. This residual variation amounts to 45% of the total response variation in the current sample. Whereas 25% of the total response variation can be explained by individual differences between persons (σ_{θ_p}), up to 29% can be explained by individual differences between items (σ_{b_i}). Hence, there is a lot of variation in responses that can be ascribed to the particular items persons need to complete. Although we now have established that there is a lot to be explained at the item side, the question remains what the determining factors are behind these individual differences in the item pool!

Figure 5.2: Item difficulty, and variance components plot of the NULL model.



The points on the y axis are the estimated difficulty of the individual items on the x axis. The error bars represent the confidence intervals ($\alpha = .05$ two-tailed) around the estimates.

5.1.2 Explanatory extension

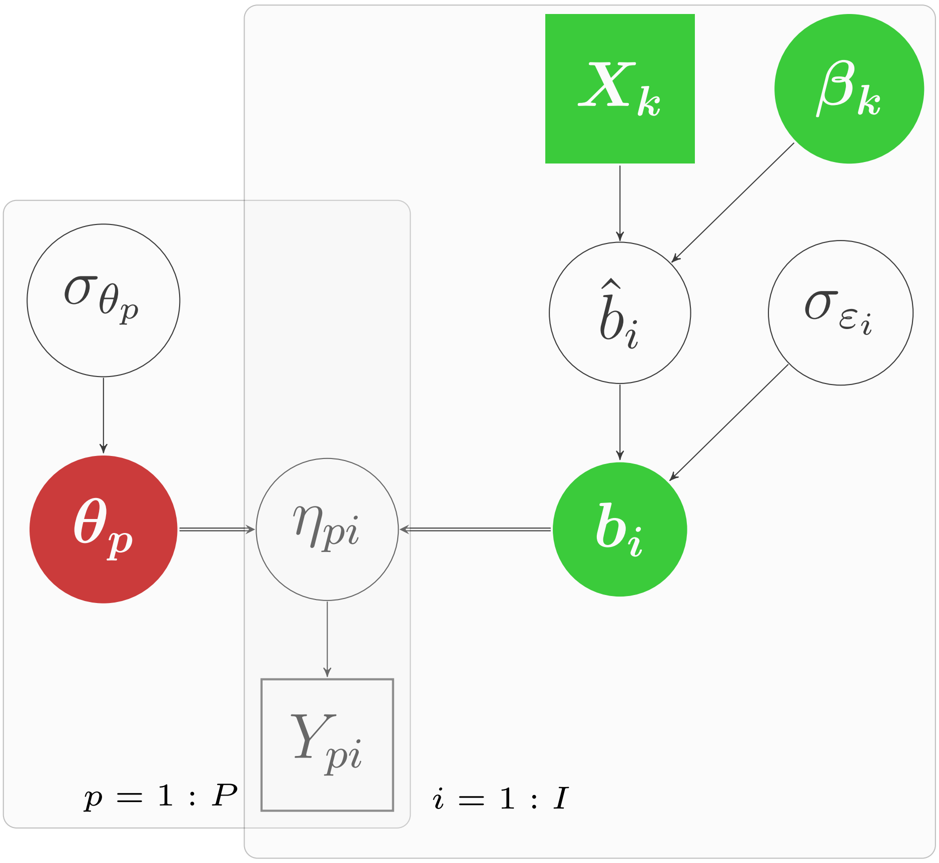
The descriptive model serves as the baseline, because none of the potential information on shared item properties has been used yet. From an item model perspective it can be considered to be a NULL model as it does not use any additional item information. To make use of this information and develop a formal item model, we can decompose the item difficulty in a part that can be predicted by item/problem properties and a residual part (see e.g. De Boeck & Wilson, 2004; Janssen, Schepers, & Peres, 2004).

An extra regression layer is added to the item side of the descriptive item response model (Equation (5.2)). The item difficulty parameter b_i is now modeled as a weighted function of item predictor values X_{ik} ($k = 1, 2 \ldots K$ with K being the number of predictors in the item model) with regression weights β_k representing the effect of predictor X_k on item difficulty. To accommodate for the fact that the predictors will most likely not predict the item difficulty perfectly, the regression layer also contains a residual item difficulty term ε_i.

$$\begin{array}{rcl} b_i & = & \sum_{k=1}^{K} \beta_k X_{ik} + \varepsilon_i \\ \sigma_b^2 & = & \sigma_{\hat{b}}^2 + \sigma_\varepsilon^2 \end{array} \tag{5.2}$$

The variation in item difficulty σ_b^2 is split up in explained variance σ_{hatb}^2 accounted for by the item predictors and residual variance σ_ε^2 that is left unexplained by these predictors. In the descriptive baseline model, the regression weights β_k can all be considered to be equalling zero, hence nullifying the explanatory predictors X_k. Also in the variance decomposition the total variance in item difficulty reduces to all residual unexplained variance ($\sigma_{b_i}^2 = \sigma_{\varepsilon_i}^2$). In other words, the descriptive model is nested within the explanatory model and can function as a NULL model for comparisons. The overall effect of all K item predictors in the explanatory model can be summarized by the percent explained item variance defined as $r_{item}^2 = 1 - \frac{\sigma_\varepsilon^2}{\sigma_b^2}$.

Figure 5.3: Graphical representation of the explanatory model.



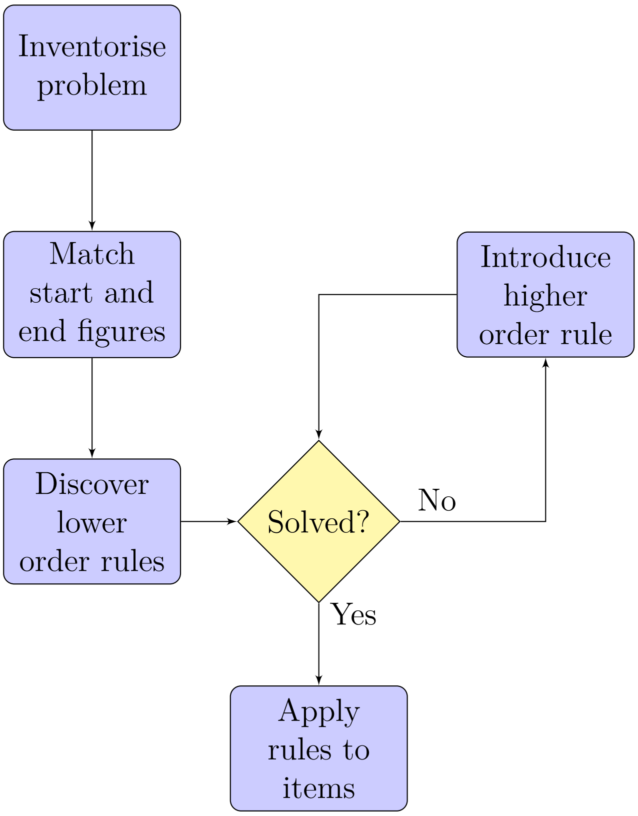
Note. The b_i will now be partially explained by weighted item properties $\sum_{k=1}^{K} \beta_k X_{ik}$. Anything left unexplained by the known item properties is considered $\sigma_{\varepsilon i}$ residual variance in item difficulty.

In our case, the item predictors in such an explanatory item response model will be hypothesized set of radicals based on characteristic features of the problem the item is part of. The effect of each predictor will be examined within their set, but also in isolation, to get a wider view of the situation.

5.2 Synthesizing the Radicals for an initial Item Model

As a starting point for structuring the item model a logical task analysis was performed that describes how the specific abstract reasoning problems in the test can be solved in a very systematical and rational way. This solution strategy was then also programmed in

Figure 5.4: Flow chart of the artificial intelligence.


Inventorise problem
Match start and end figures
Discover lower order rules
Solved?
No
Introduce higher order rule
Yes
Apply rules to items

Note. Perceptual ambiguity is assumed to impact all of the square nodes in the model.

the R programming language (R Core Team, 2015) as an artificial intelligence algorithm (AI). The task analysis is informed by the general theory from the content domain and focuses on the problem level as that came forward rather prominently in the cognitive lab. To make a comprehensive item model, the task analysis is also further supplemented by the findings from the cognitive lab, because in practice humans hardly ever act fully rational or as systematically as a computer. Human problem solving tends to rely slightly more on heuristics and trial & error.

The AI can be represented roughly by the diagram in Figure 5.4. Here the basic problem solving process is split into an inventorisation step and a matching step. The inventorisation of the problems can be said to pertain all the elements that will have to be stored in the working memory. The AI inventorized both the start and end set of figures in the problem by storing all directly observable visual features of the symbols (e.g., symbol type, color, ...) for both sets. To operationalise this inventorisation step into predictors for an item model, the sheer amount of information available in the problem needs to be quantified to be able to define the overall working memory load of the problem. More specifically, the inventorisation item predictors would include the number of figures and the number of rows in the start/end figure set of the problem, as well as the maximum number of features found in the figures of the problem. Note that the participants in the cognitive lab had also emphasised the number of figures as determining the perceived complexity of the problem.

The matching step involves matching the inventories of the start and end figures set, and aligning which inventory changes align with which buttons. In contrast to the participants, getting the AI to rationally do this basic matching was a fairly simple matter, and the noticeable part seemed to be the number of buttons, not minding the actual inventory configurations too much. The sheer number of feature changes might also be contributing to the information load of the problem; the logic being that the more changes need to be processed, the more demanding the matching should be.

For the simpler problems, the AI was able to directly connect buttons to inventory alterations in figure features, and as such derive the implied simple-change abstract reasoning rule. More complex problems, however, turned out to be unsolvable through normal operations alone, and in these cases the computer had to first try solving the problems through normal means and verify if this yielded the desired results, and if not, apply any combination of alternative abstract rules on the figures until it comes up with a solution that fits the desired outcome. These differences in rule complexity lead to an item predictor distinguishing between simple lower order rules and more complex higher order rules. Lower order rules would involve implied changes obvious to the bare eye (e.g., color change), with rule acquisition being a simple matter of inventorisation and matching. Rules of a higher order would have less obvious changes that lack a simple one-to-one relation and require the application of more involved operations (e.g., a reordering of inventory figures) on the problems to derive the abstract reasoning rule. Examples would be the reordering of inventory figures or the rotation of figures. The simplicity of the button configurations was also brought up as a factor in the cognitive lab. This was operationalised as an item predictor by counting the number of isolated buttons in the problem; if a button was constantly present in either a single or all the rows, it should be easier to isolate and derive the rule attached to it. As there was an indication that the participants had some problems with rules involving addition or removal of features (especially lines), the presence/absence of such a rule was also included as a predictor.

Perceptual ambiguity naturally did not show up in the rational task analysis, but the cognitive lab was able to shed some light on facets, although originally intended to be incidental, that might in practice impact item difficulty. Ambiguity factors as an error source stem from the human part of problem solving, so should impact both inventorisation, matching and rule derivation, hence it is not put as an independent node in Figure 5.4. Participants of the cognitive lab identified two primary factors as causing ambiguity: When a problem included multiple perceptually conflicting rules, that is rules affecting item features that were perceptually very similar to each other, most participants struggled to make sense of the problem. The higher-order rotation rule was at times also confusing when there were no obvious visual cues that a rotation happened. Each ambiguity factor was coded as a presence/absence dummy item predictor.

Table 5.1: Table of the predictors and expected effect

	Model	Source	Variable type	Expected effect
# figures	Inventory	Cognitive lab	Count	⇑ -
# features	Inventory	Task analysis	Count	⇑ -
# rows	Inventory	Task analysis	Count	⇑ -
Mean change	Matching	Task analysis	Fraction	⇑ -
# buttons	Matching	Task analysis	Count	⇑ -
Isolated buttons	Matching	Cognitive lab	Count	- ⇓
Rotate	Rule type	Task analysis	Categorical	⇑ -
Reorder	Rule type	Task analysis	Categorical	⇑ -
Add/ remove	Rule type	Task analysis	Categorical	⇑ -
Rule conflict	Ambiguity	Cognitive lab	Categorical	⇑ -
Non-visual cues	Ambiguity	Cognitive lab	Categorical	⇑ -

Note. The arrows show which direction the predictor is expected to affect the difficulty of the problem.

Table Table 5.1 provides an overview of the sets of potential item radicals and their hypothesized effects on item difficulty when they would be used as item predictors in the explanatory item response models for the abstract reasoning test under investigation. In general it is expected that inventory load, matching load, rule complexity and perceptual ambiguity all increase item difficulty.

5.3 Explanatory item response modelling Results

The different explanatory predictors will be walked through in this section, grouped into relevant explanatory models. Model one or two consist of continuous variables (at least potentially; several manifests themselves as binaries in this study), either simple counts or counted fractions. Model three and four only consist of dummy coded (see Cohen, Cohen, West, & Aiken, 2003) categorical variables (or names, to keep with Mosteller and Tukey, 1977, referenced in Hand, 1996). Each explanatory item response model will be presented in an item difficulty plot contrasting estimated item difficulty b_i to model predicted difficulty $\hat{b}_i$. A summary table of regression coefficients will be included for both the model with multiple predictors, as well as the models with the single predictors in isolation. At the end of the section is a summary table of comparisons between the different explanatory models and the descriptive NULL model.

5.3.1 Model one: inventorisation

Model one is the inventory model and includes three continuous predictors. The first predictor, the maximum number of figures in the problems, averaged at 2.2, with a minimum of one figure and a maximum of four (see Table 5.2). The SD of 1.40 is quite large, which is natural, since the first five problems only had one figure on them, and most of the other problems have at least one row with four figures. The second predictor, the number of features, was created by summarising the total number of features in the problem, and averaged at 6.47 different features, with a minimum of 5 and a maximum of 10 and a SD of 1.42. The third item predictor was simply the number of rows in the problem. There were no problems in the test with more than 4 rows and none with less than 3, with the majority having the latter amount.

The effect of each predictor within their set, as well as in isolation as a single predictor,

Table 5.2: Descriptive table of the predictors.

	Mean	SD	Min	Max	Variable type	Expected effect
# figures	2.20	1.40	1.00	4.00	Count	⇑ -
# features	6.47	1.42	5.00	10.00	Count	⇑ -
# rows	3.20	0.42	3.00	4.00	Count	⇑ -

Note. The arrows show which direction the predictor is expected to affect the difficulty of the problem.

Table 5.3: Explanatory Item Response model 1:
Regression coefficients of inventory-related predictors of item difficulty b_i.

	Multiple Predictors			Single Predictor Models			
Predictor X_k	β_k	SE	p	β_k	SE	p	r_b^2
constant	-6.70	0.84	<.001				
# figures	0.78	0.11	<.001	0.91	0.10	<.001	0.68
# features	0.08	0.09	.396	-0.21	0.12	.080	0.04
# rows	1.03	0.29	<.001	2.29	0.28	<.001	0.39
	multiple $r_b^2 = 0.75$						

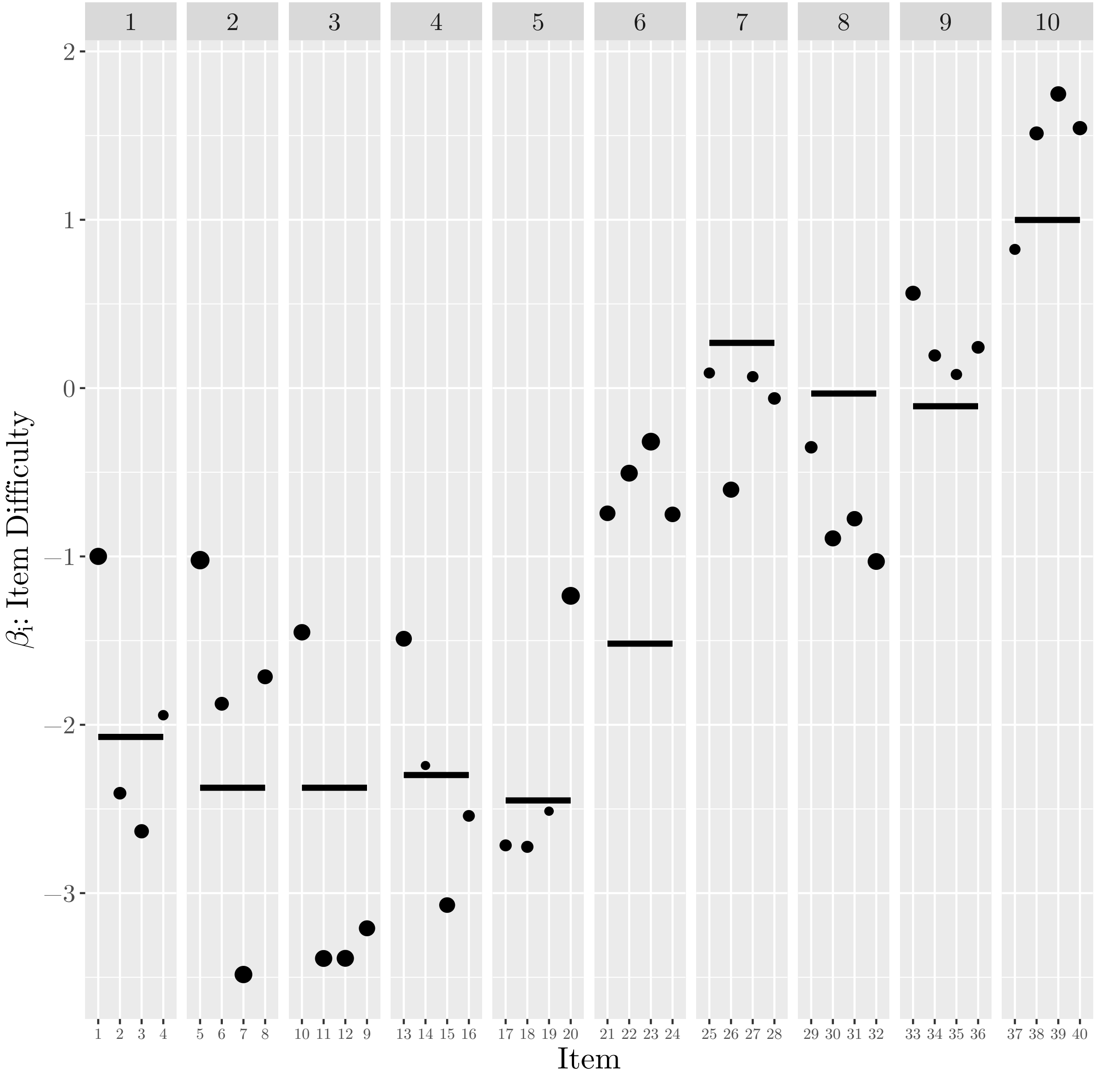
are summarized in Table 5.3. Both the number of figure columns and the number of rows in the problem show significant positive effects on item difficulty as expected (i.e. more figures or rows made the problem harder). However, the number of features surprisingly did not seem to have any effect at all (Table 5.3). The effects of the predictors did not change when combining them into a joined multiple regression like model, however there was some redundant information carried among the predictors, which we can see by comparing the combined sum of the r_b^2's of the single predictor model with the multiple r_b^2. This comes from the correlation between the number of figures and the number of features and number of rows, such that the predictors explain some of the same variance in b_i.

We can see in Figure 5.5 that the inventory model predicts the first five problems to be fairly easy as they have limited inventory load. From problem seven on out, the items should have higher inventory load and be quite difficult, and problem ten would be really hard.

Table 5.4: Correlation matrix of the predictors

	# figures	# features	# rows
# figures	1		
# features	−.35	1	
# rows	.49	−.05	1

Figure 5.5: Plot of predicted and estimated item difficulty in the inventory model.



Note. Horizontal lines represent predicted item difficulty $\hat{b}_i$, points represent estimated item difficulty b_i, with point size proportional to absolute residual item difficulty $|\hat{\varepsilon}_i|$.

5.3.2 Model two: matching

Model two is the matching model and consists of three continuous item predictors. The first item predictor, labelled the mean change, was constructed by counting the number of inventory changes going from the start to the end set of figures in the problem and dividing by the total number of features. This predictor essentially represents the proportion of the features that change in the problem. The mean across all items of this proportion was .28, with a *SD* of .09, which means that 28 % +- 9 % of the features would change on average across the whole test (see Table 5.5). The absolute least amount of change is 19 %, roughly one fifth, and at the most is 46 % that is almost half of the features in the problem changes. Each unique button in the problem represent a rule. The second item predictor is the number of unique buttons in a problem and has an average of 3.9, with a minimum of 3 and a maximum of 5 buttons, with a *SD* of 42. Most of the problems have four unique buttons. The third predictor, labeled isolated buttons, represents the simplicity of the button configuration, and is the only predictor that is expected to have a negative effect on item difficulty. The simpler the button configuration is, the easier the matching can be executed to correctly solve the problem. The predictor has an average of 1.7 with a *SD* of .67, and ranges from 1 to 3 isolated buttons. The expectation is that more isolated buttons will mean there are more independently derivable rules. This should in theory reduce the demand on the persons ability to mentally model the problem, by letting them partition it into more manageable chunks.

Table 5.5: Descriptive table of the predictors.

	Mean	SD	Min	Max	Variable type	Expected effect
Mean change	.28	.09	.19	.46	Fraction	⇑ -
# buttons	3.90	0.57	3.00	5.00	Count	⇑ -
Isolated buttons	1.70	0.67	1.00	3.00	Count	- ⇓

Note. The arrows show which direction the predictor is expected to affect the difficulty of the problem.

The effect of each predictor within their set, as well as in isolation as a single predictor, are summarized in Table 5.7. When all the predictors are part of the same model, the number of unique buttons has a significant positive effect on item difficulty, and the number of isolated buttons has a significant negative effect. Both of the former is as

Table 5.6: Correlation matrix of the predictors

	Mean change	# buttons	Isolated buttons
Mean change	1		
# buttons	$-.33$	1	
Isolated buttons	$-.09$	.20	1

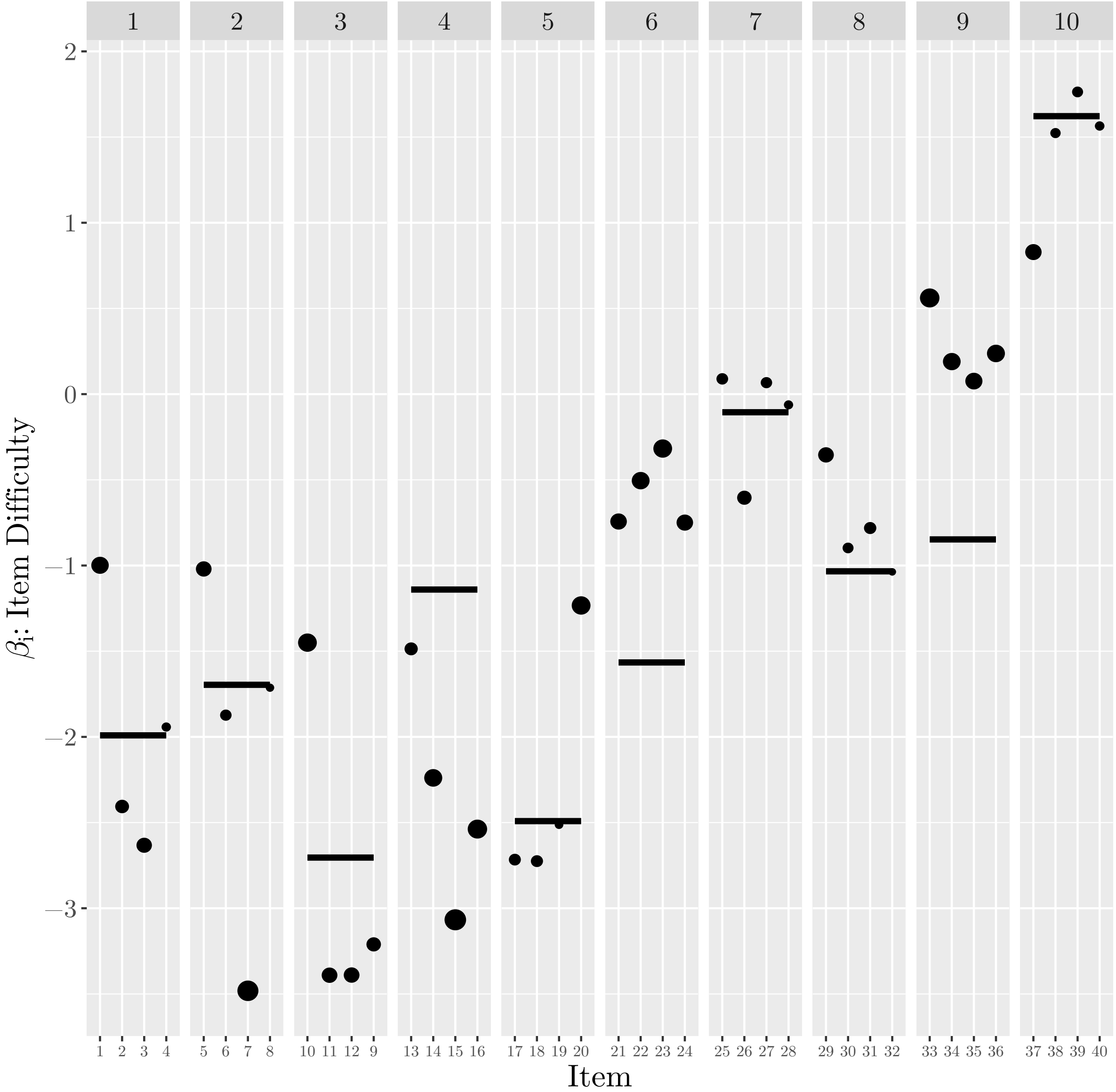
Table 5.7: Explanatory Item Response model 2:
Regression coefficients of matching-related predictors of item difficulty b_i.

	Multiple Predictors			Single Predictor Models			
Predictor X_k	β_k	SE	p	β_k	SE	p	r_b^2
constant	-6.17	0.82	<.001				
Mean feature change	-4.15	0.92	<.001	-7.55	1.19	<.001	0.20
# buttons	1.91	0.20	<.001	1.94	0.21	<.001	0.51
# isolated	-0.77	0.21	<.001	-0.39	0.38	.311	0.03
	multiple $r_b^2 = 0.66$						

expected, however, the mean changes of item features had a significant negative effect, when the expectation was for the opposite. The cause for this might be that the number of changes is dependent to other item predictors outside the current set of the Matching model. If you for instance take the reordering rule, it is logical that this rule in essence implies changes of many features in the inventory; however, a person would not necessarily interpret it as many changes occurring, but as one single operations affecting the whole figure. More features that change might also make it more easy to spot the particular pattern and connected rule, leading to a decrease in item difficulty.

Notice that there is also a suppression effect in this model. The effect of the number of isolated buttons is non-significant when it is the only explanatory predictor, but becomes significant when put together in context with the other item predictors, suggesting that the isolated buttons indeed matter if the mean feature change and number of buttons are kept constant. The number of isolated buttons is positively related to the number of buttons, which could imply that for problems with more buttons (which means more changes) the button configuration is easier to decipher when some of these are isolated, leading to a decrease in item difficulty. The resulting predicted and estimated item difficulties are plotted in Figure 5.6.

Figure 5.6: Plot of predicted and estimated item difficulty in the matching model.



Note. Horizontal lines represent predicted item difficulty $\hat{b}_i$, points represent estimated item difficulty b_i, with point size proportional to absolute residual item difficulty $|\hat{\varepsilon}_i|$.

5.3.3 Model three: rule type

Model three is the rule type model and consists of three dummy coded item predictors. The regular lower order rules serve as a baseline or reference category (cf. intercept/-constant), as the assumption is that rules that are characterised by simple matching of single feature changes should not make the problem more difficult. Also, these rules are present in basically all of the problems.

The item predictor Rotation was operationalised by flagging any problem that was deemed to contain elements that might be interpreted as rotation. Even though some of the rules could be construed as doing something different like only parts of the figure changing or similar (expressions by participants in the cognitive lab suggests this), it was decided that these cases should be bundled into the same category in order for the variable to be more robust. Rotation is expected to make the problem more difficult, since spatial rules in theory should be more difficult than simple ones (Primi, 2001), although the cognitive lab seemed to cast some doubt on this.

The item predictor Reordering was operationalised as any event where figures in a row changes places if a certain button is present. It should increase demand on the participants both by making the whole row unsolvable by simple matching until the rule is identified (as relevant figures can not be matched according to the baseline matching procedure), as well as requiring the performance of abstract operations to mentally model the problem.

If a rule is present that adds or removes a feature from the figures without some resemblance of changing properties, then that problem is said to have an add/remove rule. The addition of lines that was much talked about by the participants during the cognitive lab (Chapter 4) is a good example of this rule type. Addition of colour fill in an otherwise "blank" figure, was for instance classified as a change rule, since there was few indications that this rule was construed otherwise.

Both the rotation and reordering rule were present in four out of ten problems (see Table 5.8). Rules that controls feature addition and removal were present in seven out of ten problems. The inclusion of any of these additionally complex rules in the problem is expected to increase the item difficulty. The correlation matrix in Table 5.9 illustrates that there is no systematic pattern of co-occurrence of these rules, although rotation tends to be negatively correlated with the occurence of other more complex rules.

Table 5.8: Descriptive table of the predictors.

	Mean	SD	Min	Max	Variable type	Expected effect
Rotate	.40	.52	0	1	Categorical	⇑ -
Reorder	.40	.52	0	1	Categorical	⇑ -
Add/ remove	.70	.48	0	1	Categorical	⇑ -

Note. The arrows show which direction the predictor is expected to affect the difficulty of the problem.

Table 5.9: Correlation matrix of the predictors

	Rotate	Reorder	Add/ remove
Rotate	1		
Reorder	$-.25$	1	
Add/ remove	$-.36$	.09	1

The effect of each predictor within their set, as well as in isolation are summarized in Table 5.10. The presence of reordering rules is the only rule that had the expected significant positive effect on item difficulty. The presence of a rotation as well as the addition/removal rule had no significant effect on difficulty when part of the compound model. When checking the predictors in isolation, however, the rotation rule actually showed a significant negative effect on difficulty explaining 12 % of the variance. When looked at in context this effect disappears, implying that it can be attributed to the additional difficulty of a specific problem with both rotation and reordering present.

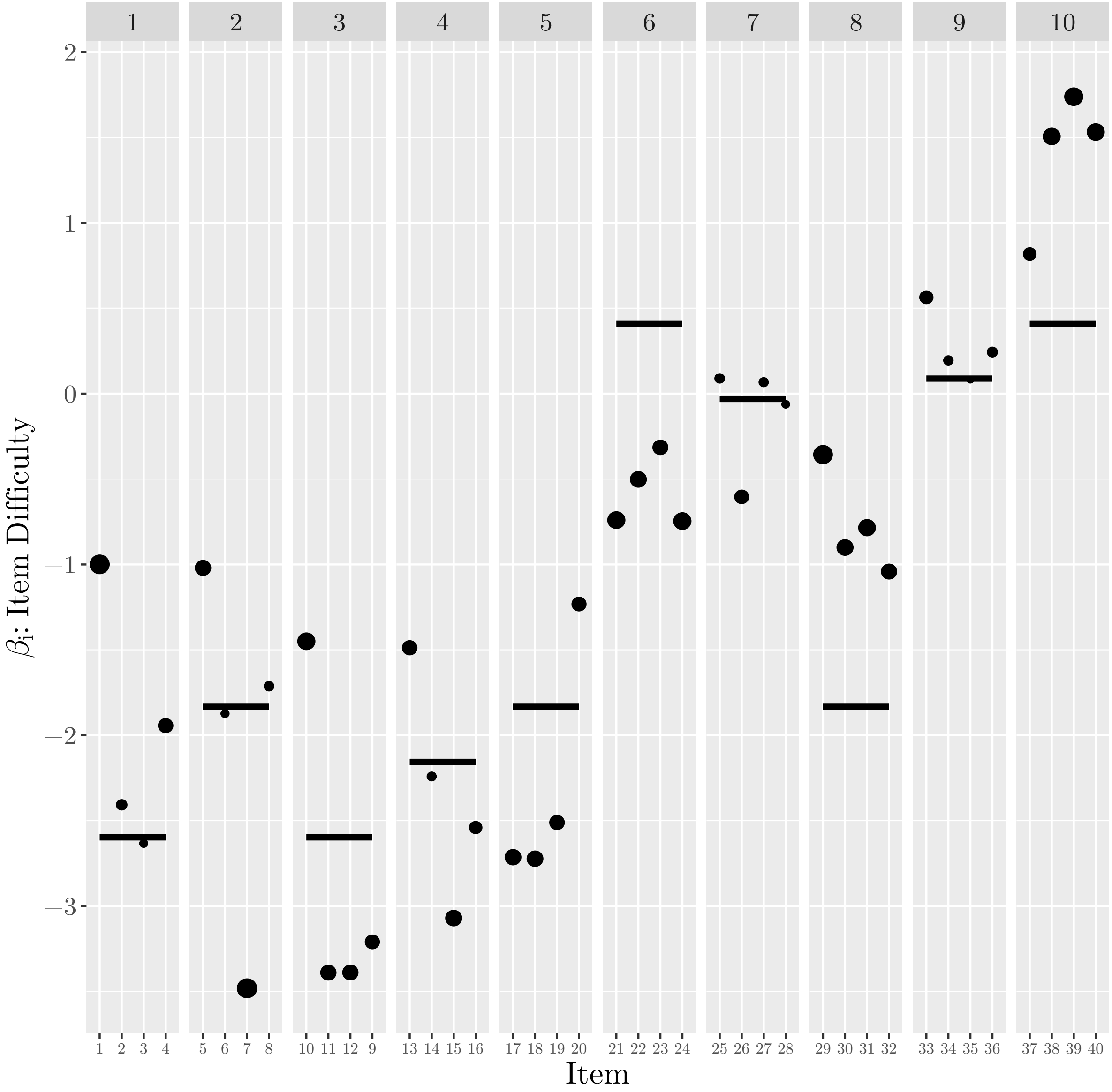
Model three has a fairly decent fit, with a multiple R^2 of .67. As we can see in Figure 5.7, there are some problems where the predicted difficulty diverges noticeably from the actual difficulty. Problem six is a good example of this; the model predicts it to be somewhat more difficult than it actually is, but if we look at Figure 5.5 and Figure 5.6, the inventory and matching models actually predict it to be easier. This seems to indicate that the effect of the presence of a higher order rule as a radical, really amplifies the difficulty of an already complex item. In problem eight, for instance, the model predicts it to be substantially easier than it actually is, which implies that there are few effective

Table 5.10: Explanatory Item Response model 3:
Regression coefficients of rule type-related predictors of item difficulty b_i.

	Multiple Predictors			Single Predictor Models			
Predictor X_k	β_k	SE	p	β_k	SE	p	r_b^2
constant	–2.28	0.34	<.001				
Rotation	-0.32	0.30	.274	-1.03	0.42	.015	0.12
Reordering	2.24	0.28	<.001	2.36	0.29	<.001	0.63
Add/remove	0.44	0.31	.155	0.78	0.53	.145	0.06
	multiple $r_b^2 = 0.67$						

higher order rules present in the problem, and that the difficulty is affected by other sources.

Figure 5.7: Plot of predicted and estimated item difficulty in the rule type model.



Note. Horizontal lines represent predicted item difficulty $\hat{b}_i$, points represent estimated item difficulty b_i, with point size proportional to absolute residual item difficulty $|\hat{\varepsilon}_i|$.

5.3.4 Model four: ambiguity

Model four is the perceptual ambiguity model and includes two item predictors. In this study, the ambiguity predictors are mainly based on unintended design factors that were picked up primarily during the cognitive lab, but also while making the AI. Hence, these predictors represent potentially radical elements, because more ambiguity should lead in theory to more difficult problems. The two predictors are dummy-coded and represent problems that include specific ambiguous rules: Rules implying a rule-feature conflict and a rotation rule without visual cues (see Table 5.11).

Five out of ten problems include a rule feature conflict of some sort. This conflict would mostly stem from the problems including multiple rules that manifested themselves as affecting item features that would be easy to confuse for the persons solving the problem. Only two out of ten problems had non-visual cues when a figure was rotated. Compared to the conflict predictor, this predictor only represent cases where one rule is ambiguous in itself, not in a relation with other rules (see Table 5.12). The idea is that the presence of this factor in a problem, however rationally solvable it is, should nevertheless trick the non-rational person by offering a multitude of different interpretation of what exactly changes in relation to the rule. Persons might for example interpret this rule as both regular rotation, changing direction of lines, or even multiple rule operations using one button (like for example both adding a horizontal and removing a vertical line). Several of these interpretations surfaced throughout the cognitive lab as the participants reasoned their way through the test. Although this predictor is not very variable, and not necessarily independent of the conflict rule, it was still included, as it looked a hypothetically promising factor, as well as to flesh out the ambiguity model.

Table 5.11: Descriptive table of the predictors.

	Mean	SD	Min	Max	Variable type	Expected effect
Rule conflict	.50	.53	0	1	Categorical	⇑ -
Non-visual cues	.20	.42	0	1	Categorical	⇑ -

Note. The arrows show which direction the predictor is expected to affect the difficulty of the problem.

The effect of each predictor within their set, as well as in isolation are summarized

Table 5.12: Correlation matrix of the predictors

	Rule visual conflict	Non-visual cues
Rule visual conflict	1	
Non-visual cues	0	1

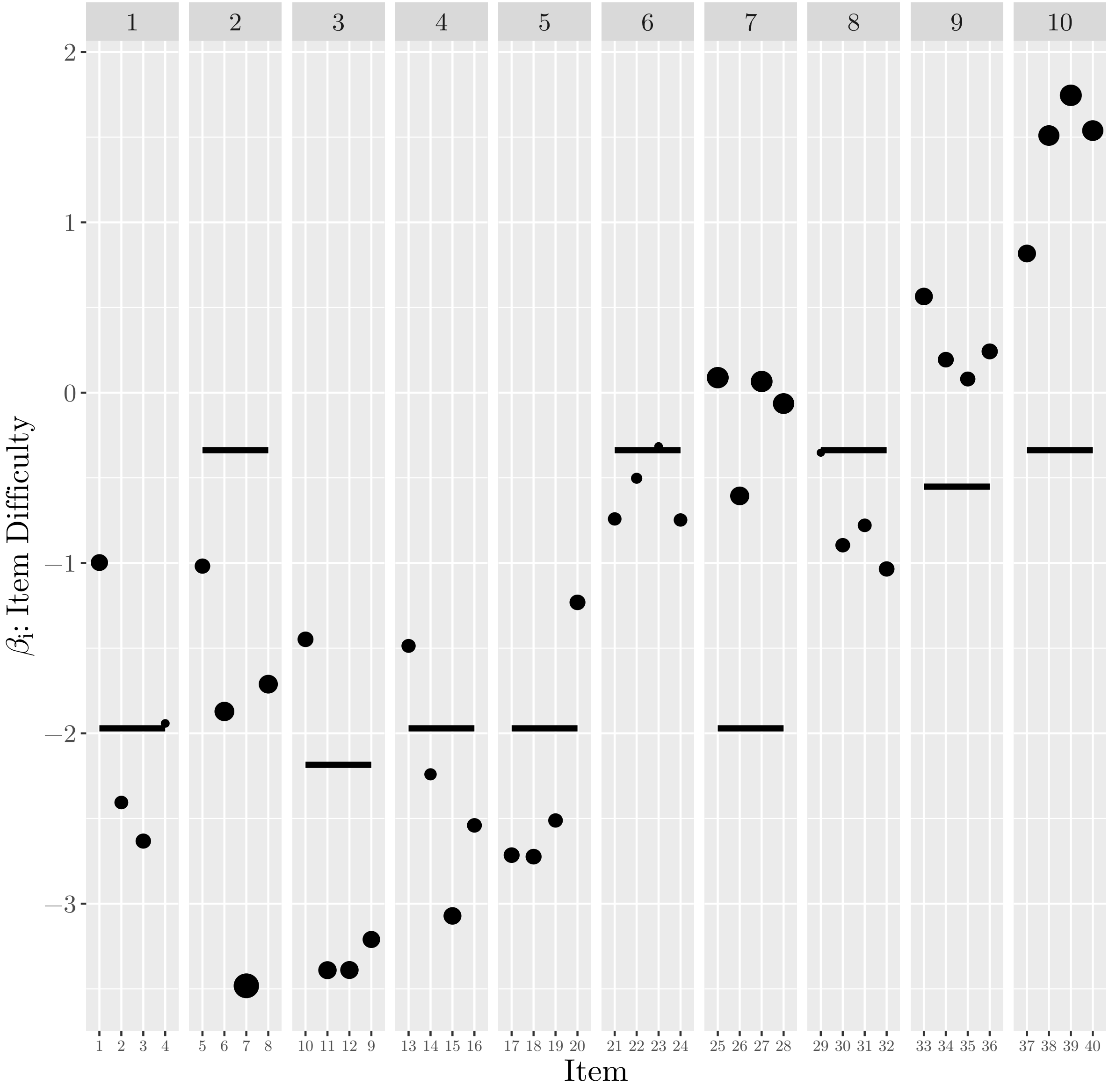
in Table 5.13. The presence of very similar rules, as represented by the rule conflict predictor, did have the expected negative effect on item difficulty. Rotation with no visual clues, however, showed no effect both as a single predictor as within the set. This suggests that in contrast to the expectations from the cognitive lab, persons had in practice no clear trouble in discovering the subtle version of the rotation rule. This might be a difference between the perceived difficulty and the effective difficulty of a problem. However, do note that the rotation without visual cues predictor is rather fragile as it reflects performance on only two problems.

Table 5.13: Explanatory Item Response model 4:
Regression coefficients of ambiguity-related predictors of item difficulty b_i.

	Multiple Predictors			Single Predictor Models			
Predictor X_k	β_k	SE	p	β_k	SE	p	r_b^2
constant	-1.97	0.26	<.001				
Rule conflict	1.63	0.34	<.001	1.63	0.34	<.001	0.31
Rotation no visual clues	-0.21	0.41	.600	-0.22	0.49	.660	<0.01
	multiple $r_b^2 = 0.32$						

The ambiguity model had the worst fit among all the models, with an R^2 of .32. Since one of the two predictors seems to have little to no effect on the difficulty of the items, there is very little information left in the model, reducing the prediction to a dichotomy; either the problem is easy, or it is hard. This means the model is less informative, but there are still some interesting artefacts that can be cause for speculation. If we look at the predicted and estimated item difficulties plot in Figure 5.8, problem two, has the largest spread in the test. The model predicts it to be a difficult problem because of ambiguousness; it turns out that the only item that does not have the ambiguousness factor present is evidently far more easy than the rest of the items in that problem, which could be an indicator that ambiguity or better the absence of it does make solving an item easier. Problem seven is an example in the opposite direction, where we see that the ambiguity model drastically underestimates it. There are especially few item features in this problem that could be manipulated to cause perceptual ambiguity compared to the other problems, the difficulty instead stemming from other sources, especially matching and rule complexity (Figure 5.6 and Figure 5.7).

Figure 5.8: Plot of predicted and estimated item difficulty in the ambiguity model.



Note. Horizontal lines represent predicted item difficulty $\hat{b}_i$, points represent estimated item difficulty b_i, with point size proportional to absolute residual item difficulty $|\hat{\varepsilon}_i|$.

5.3.5 Model comparison and summary

Each of these four explanatory models are extensions of, and thus can be formally compared to, the NULL model. This is done by utilising a likelihood ratio test as well as various penalised fit measures. The null hypothesis of the study is that the explanatory models fail to differ from the NULL model in terms of model fit. The NULL model is nested within each of the explanatory models, but the explanatory models do not followed a nested structure among themselves. It is possible to compare one explanatory model relatively to the others though the use of fit measures like the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). These measures balance the absolute fit and model complexity by including a penalty in terms of the number of parameters in the model (Burnham & Anderson, 2002), the logic being like that of Ockham's razor, where of two equally good models, the one with the fewer parameters should be chosen. Lower values indicate better balance between absolute model fit and model complexity.

The ideal case would be to present a full model which includes all the item predictors, but this turned out to be impossible with the quasi-experimental item design of the current study. The item set simply does not comply with the already identified theoretical framework for item radicals. For some combinations of properties we might have no items that match, and therefore a single item can be over-identified by considering all the predictors at once. This largely manifests itself as an issue of extreme multicollinearity between the predictors, causing the estimation of the model to fail to converge. When estimation was attempted, several predictors indeed had to been dropped from the estimation because of rank deficiency.

Table 5.14: Summary of the fit and explanatory power of the different models

Model	AIC	BIC	LL	LRT	df	p	r_b^2
0. NULL	151691	151722	-75843				
1. Inventory	151643	151704	-75815	55	3	<.001	.75
2. Matching	151654	151715	-75821	44	3	<.001	.66
3. Rule type	151653	151714	-75821	44	3	<.001	.67
4. Ambiguity	151680	151731	-75835	15	2	<.001	.32

The various fit measures for the item models in this study are summarised in Table 5.14. All in all, the inventory model (m1) seems to fit the data best, accounting for 75 % of the item-attributed variance of the item responses. The matching model (m2) accounted for 66 % and the rule type model (m3) 67 %. The ambiguity model (m4) explains only 32 % of the variance, and even fails to outperform the NULL model on the BIC, although not on the other measures. In general, it is possible to conclude that adding the extra predictors cause a model to significantly outperform the NULL model every time, and this gives merit to the included item properties and facets of all the assumed cognitive processes, even though further investigation clearly needs to be done.

There were some unexpected findings from the explanatory item response modelling analyses. None of the predictors involving anything related to figure rotation had any effect separating them from the simple change rules, contrary to what one should expect from theory (Primi, 2001). The feature addition or removal rule also showed no effect, although when multiple features were altered, like addition and removal of multiple lines discussed in the cognitive lab (Chapter 4), this seemed to cause an effect. Raw information load seemed to only increase difficulty when it was related to the size of the problem, as the number of features showed no effect and the mean features change even made an item easier.

Chapter 6

Discussion

6.1 Systematic item design

As this study is no true experiment, it is hard to say for certain what factors ultimately determine difficulty in the test. An obvious cause of this might be the layer-like structure of the test - the base of most of the problems consist of lower-order rules, with more advanced rules sprinkled on top, causing the sheer information load to grow with addition of higher order rules. Using higher-order rules naturally contributes to the massiveness of the problem, especially when it comes to rules requiring more figure columns (e.g. reordering).

Working memory capacity has generally been designated as the primary cause of item difficulty (Hosenfeld, 1997; Mulholland et al., 1980; Primi, 2001), and so it is in this case. The size of the problem, the number of rows, rules and figures, should act as a base. The rule taxonomy should act as an added layer, but more rules should be developed, especially simple ones, so that it would be possible to vary item size without introducing more complex rules. When developing a concrete item model for future automatic item generation of the assessment, information load, manifested in the number of figures and rules, should be emphasised as among the more powerful radicals.

The inclusion of some higher order rule should indeed affect the difficulty of the problem, especially if it already is fairly demanding from other sources. However, there was limited support for the impact of all the levels of the rule taxonomy found in Primi (2001), as especially the spatial part seemed to have no effect. This could stem from the test being structurally different from a matrix and completion test, hence the cognitive properties are simply not alike.

Addition and removal of concrete item features did seem to confuse the participants in the cognitive lab in Chapter 4, suggesting that the presence of that kind of rule makes the problem less intuitive. However, this did not show up in the statistical analysis, suggesting that the effect either did not show due to bad experimental design, or that the particular cases of this rule type manifest itself at a simpler level. This is also in line with the findings of Primi (2001), where feature addition or removal was disconfirmed to be among the more complex rules.

One of the perceptual ambiguity factors uncovered in the cognitive lab stood up to scrutiny during the explanatory IRT modelling, the rule conflict. Although there might yet be other potential sources of perceptual ambiguousness, this could be the main source in this type of test at this stage.

To tap into a person's working memory capacity to the purest extent would mean to focus on the information load (as discussed in the review by Arendasy & Sommer, 2005), which logically would imply limiting the impact of perceptual factors. However, it could be argued that tests of abstract reasoning has a problem solving aspect that transcends a narrow interpretation of the working memory. In this case, the "noise" elements could be a way to challenge other aspects of the cognition of the persons. Either way, ambiguity factors need to be known, so that they can be taken into account in the item model, either to exclude or embrace them.

6.2 Recommendation for further research

There are several aspects of the test that could be modelled. Currently only the one parameter logistic model (1PL) has been used, which implies that only the item difficulty is a function of the radicals in the item model. The IRT model could be expanded by for instance including a parameter for different discrimination, or even a lower asymptote (pseudo-guessing) parameter (Baker & Kim, 2004).

The impact of the testlet structure has not been directly addressed in itself at this point. Possibilities for modelling local dependencies between the items are multiple. Misconstruation of a rule in one testlet could potentially propagate to other testlets, especially if that testlet contains a similar rule. This could possibly lead to local dependencies between certain items with shared rules, which would be interesting to investigation further.

Given that many participants made informed guesses, it might be interesting to implement alternatives that resemble the correct answer, with the same number of rules, only that one or more rule are wrong. As things are now, persons would be able to eliminate some alternatives from the list without having to fully understand the problem. Eliminating sources of information among the alternatives milly similar, but wrong alternatives, arranging alternatives of varying likelihood of being correct, in essence recognising a heuristic solution of the item. The question of why persons guess has not been carried on from the cognitive lab at this point. Some findings indicate that guessing happens when rules are misconstrued, which for instance could be caused by rule complexity or ambiguousness. Guessing as a strategy have been pinpointed to influence performance on cognitive tests (Egan & Schwartz, 1979, referenced in Ericsson & Lehmann, 1996), so should be further investigated.

The item model is now based on problem characteristics only, and it is reasonable to assume that it can be improved by going into more detail at the item level. The main focus of the cognitive lab has also been on how participants analyse the problem screen, and how they understand their interaction with it. If the present test was to be investigated through further use of qualitative methods, an investigation of all the test items should be conducted. Exploring the qualities of each item in relation to the problem screen should be emphasised, for example through an item-by-item retrospective cognitive interview protocol (like for instance the one employed by Paap et al., 2015). This should be studied more closely looking at the particular items whose predictions based on problem models are especially off target.

Other improvements to the cognitive lab could possibly provide some better general insights into the person interaction with the test. Analysing the behaviour of the participants when answering the items was made difficult by only having audio data, as much contextual information depended on the memory of the observer, and his ability to synthesise the different data types. This was mitigated by notes of the observer taken during and after the procedure. However using video recordings or real eye-tracking technology (like Carpenter et al., 1990) might give a slightly more objective picture of perceptual attention involved when solving the test.

A more stringent verbal protocol analysis should be performed (Ericsson & Simon, 1993; Leighton, 2005; Ohlsson, 2012), to for instance investigate choice of words used when talking about the different rule types: whether a figure “becomes” something else, is different from figures that “gets” something. If there are systematic tendencies on how exactly persons talk about the rules, this might shed some light on the reasons for why spatial and complex rules like rotation as well as feature addition or removal seems to behave no different from simple feature change rules.

Whether the reasoning strategies are different for persons of different ability level, would be interesting to investigate. For matrix items we know from Arendasy and Sommer (2005) that for instance the perception rule impacts persons of low ability differently, and this effect would also be interesting to investigate for this assessment. If an effect shows, this would be a prime candidate for study through qualitative means.

The current item pool has been constrained by the fixed design of the existing test, hence it is not really a proper experimental design. As there is a lot of dependency issues between the predictor in this study, new items should be generated so as to investigate any combination of item properties. The current item pool large to analyse all the item properties. Making a completely new item bank with focus on experimental design should be considered. This would facilitate further development of the representational theory (Whitely, 1983), which is crucial to have if the assessment is to be modernised.

Most of the persons in the explanatory study had run out of time before reaching the end of the test, with only 13 % answering the first item of the last testlet. This is a problem for the validity of any results gained from modelling the properties of these items. It should be reasonable to assume that only the most able persons reach the end of the test, and as people of different ability levels react differently to the various radicals (for example like Arendasy & Sommer, 2005, fund with perceptual ambiguity), this might skew results. This could also be an indication that the test might be too long, thus rather inefficient. This is another reason why a strong item model is useful, as the test could be made adaptive and tailored to the individual person (Van der Linden & Glas, 2009; Wainer, 2000). This would make the test both more efficient and effective, whilst sparing the persons having to solve an unnecessarily large amount of items that does not add any information to improve their ability estimate.

The reaction time measures has not yet been utilised as other than a data cleaning tool. This is an untapped data source, and could provide a new perspective on person reaction to for instance rule complexity (like done by Primi, 2001).

Further investigation of process factors, like the accumulated rule space and the effect of misconstruation of rules (also related to ambiguous features) that could propagate through the test, both of which surfaced in the cognitive lab, has not been investigated further at this point. A misconstruation experiment could for instance be designed, where some persons get detailed feedback, some get limited and some get none. Within this test, the accumulated rule space couch the mum fairly early on. It woube interesting to see if adding new rules at a slower pace would impact the problem solving of the persons. Logically speaking, taking the limited cognitive capabilities of the person into account (Miller, 1956; Simon, 1972), having fewer possible rules to choose from should limit the total number of combinations to be considered.

6.3 Conclusion

The ability of the problem part of the testlets to explain this much of the item variance by it self is encouraging for further developing the test. There is still a lot of work that can be done on this existing assessment, and several concrete suggestions, both qualitative and quantitative, havbeen mentioned in this discussion. In order to modernise it however, strong theory as nded. Hence, a ne item bank with experimentally manipulated radicals, identified in this study (Chapter 5), should be generated, utilising systematic item design and the advanced technological and statistical tools described in Chapter 1. If the new item model still holds up, this could further prove the feasibility of modernising this assessment. A new vast population of precalibrated items can then be automatically generated, opening up for tailoring assessments and development of a CAT, with ahis offers. All this aside, the arguably most satisfying side to having a valid item model, is that it makes it possible to say with a fair amount of certainty that we actually know what we measure.

References

Arendasy, M., & Sommer, M. (2005). The effect of different types of perceptual manipulations on the dimensionality of automatically generated figural matrices. *Intelligence*, *33*(3), 307–324.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation (volume 2)* (pp. 89–195). New York: Academic Press.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory (vol. 8)* (pp. 47–89). New York: Academic Press.

Baker, F., & Kim, S.-H. (2004). *Item Response Theory: Parameter Estimation Techniques.* New York: Marcel Dekker.

Bastien, J. M. C. (2010). Usability testing: a review of some methodological and technical aspects of the method. *International Journal of Medical Informatics*, *79*(4), e18–e23. doi: 10.1016/j.ijmedinf.2008.12.004

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ, USA: Lawrence Erlbaum Associates Publishers.

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.). Springer-Verlag.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, *56*(2), 81–105.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures:

a theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological review*, *97*(3), 404–431.

Cattell, R. B. (1940). A culture-free intelligence test. I. *Journal of Educational Psychology*, *31*(3), 161–179.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. (2003). *Applied Multiple Regression / Correlation Analysis for the Behavioral Sciences* (3rd ed., Vol. Third Edit). London: Lawrence Erlbaum associates. doi: 10.2307/2064799

Cole, M. (1996). *Cultural Psychology: A Once And Future Discipline.* Cambridge, Massachusetts: The Belknap Press Of Harvard University Press.

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed., Vol. 4). Boston: Pearson.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. doi: 10.1037/h0040957

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The Estimation of Item Response Models with the lmer Function from the lme4 Package in R. *Journal Of Statistical Software*, *39*(12), 1–28.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* New York: Springer.

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471–517). Westport, CT, USA: Praeger Pub Text.

Ebbinghaus, H. (1913). *Memory; a contribution to experimental psychology* (C. E. Bussenius & H. A. Ruger, Eds.). New York: Teachers College, Columbia University. Retrieved from `http://catalog.hathitrust.org/Record/000360867http://hdl.handle.net/2027/mdp.39015014557006`

Embretson, S. E. (1994). Application of cognitive design systems to test development. In C. R. Reynolds (Ed.), *Cognitive assessment: A multidisciplinary perspective* (pp. 107–135). New York: Plenum Press.

Embretson, S. E. (1995). Developments toward a cognitive design system for psychological tests. In D. Lupinsky & R. Dawis (Eds.), *Assessing individual differences in human behavior* (p. 1748). Palo Alto, CA: Davies-Black Publishing Company.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests:

Application to abstract reasoning. *Psychological Methods*, *3*(3), 380–396.

Embretson, S. E. (2004, jan). FOCUS ARTICLE: The Second Century of Ability Testing: Some Predictions and Speculations. *Measurement: Interdisciplinary Research & Perspective*, *2*(1), 1–32. doi: 10.1207/s15366359mea0201_1

Ericsson, K. A., & Lehmann, A. C. (1996, jan). Expert and exceptional performance: evidence of maximal adaptation to task constraints. *Annual review of psychology*, *47*, 273–305. doi: 10.1146/annurev.psych.47.1.273

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: verbal reports as data* (Revised ed.). Cambridge: The MIT press.

Flynn, J. R. (2007). *What Is Intelligence?: Beyond the Flynn Effect.* Cambridge University Press.

Fonteyn, M. E., Kuipers, B., & Grobe, S. J. (1993). A description of think aloud method and protocol analysis. *Qualitative health research*, *3*(4), 430–441.

Gierl, M. J., & Haladyna, T. M. (2013). Automatic item generation: an introduction. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: theory and practice.* New York: Routledge.

Gierl, M. J., Zhou, J., & Alves, C. (2008). Developing a Taxonomy of Item Model Types to Promote Assessment Engineering. *The Journal of Technology, Learning, and Assessment*, *7*(2), 1–51.

Gilhooly, K. j. (2004). Working Memory and Reasoning. In J. P. Leighton & R. J. Sternberg (Eds.), *The nature of reasoning* (pp. 49–77). Cambridge: Cambridge University Press.

Glas, C. A., & van der Linden, W. J. (2003). Computerized Adaptive Testing with Item Cloning. *Applied Psychological Measurement*, *27*(4), 247–261.

Halpern, D. F. (1997). Sex differences in intelligence. Implications for education. *The American psychologist*, *52*(10), 1091–1102.

Hand, D. J. (1996). Statistics and the theory of measurement. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *159*(3), 445–492.

Hart, S. G. (2006). NASA-Task Load Index (NASA-TLX); 20 Years Later. In *Proceedings of the human factors and ergonomics society 50th annual meeting* (pp. 904–908). Santa Monica: HFES.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati

(Eds.), *Human mental workload.* Amsterdam: North Holland Press.

Hosenfeld, B. (1997). Constructing Geometric Analogies for the Longitudinal Testing of Elementary School Children. *Journal of Educational Measurement*, *34*(4), 367–372. doi: 10.1111/j.1745-3984.1997.tb00524.x

Hunt, E. (2011). *Human intelligence.* Cambridge: Cambridge University Press.

Hunt, E., Frost, N., & Lunneborg, C. (1973). Individual Differences in Cognition: A New Approach to Intelligence. *Psychology of Learning and Motivation - Advances in Research and Theory*, *7*(C), 87–122.

Irvine, S. H., & Kyllonen, P. C. (Eds.). (2002). *Item generation for test development.* Mahwa, NJ: Lawrence Erlbaum associates.

Jacobs, P. I., & Vandeventer, M. (1972). Evaluating the Teaching of Intelligence. *Educational and Psychological Measurement*, *32*(2), 235–248. doi: 10.1177/001316447203200201

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach.* (pp. 189–212). New York: Springer.

Johnson-Laird, P. N. (2004). Mental models and reasoning. In J. P. Leighton & R. J. Sternberg (Eds.), *The nature of reasoning* (pp. 169–204). Cambridge: Cambridge University Press.

Kane, M. T. (1982). A Sampling Model for Validity. *Applied Psychological Measurement*, *6*, 125–160.

Katz, I. R., Bennett, R. E., & Berger, A. E. (2000). Effects of Response Format on Difficulty of SAT-Mathematics Items: It's Not the Strategy. *Journal of Educational Measurement*, *37*(1), 39–57. doi: 10.1111/j.1745-3984.2000.tb01075.x

Katz, I. R., Moon, J. A., & King, T. C. (2015). Cognitive lab techniques: an overview, a framework, and some practice. *Workshop at the annual meeting of the National Council on Measurement in Education.*

Lai, H., Gierl, M., & Breithaupt, K. (2012). *Design Principles Required for Skills-Based Calibrated Item Generation* (Tech. Rep.). University of Alberta. Retrieved from `http://mcc.ca/wp-content/uploads/Technical-Reports-Lai-2012.pdf`

Leighton, J. P. (2004). The Assessment of Logical Reasoning. In J. P. Leighton & R. J. Sternberg (Eds.), *The nature of reasoning* (pp. 291–312). Cambridge: Cambridge University Press.

Leighton, J. P. (2005). Avoiding Misconception, Misuse, and Missed Opportunities: The Collection of Verbal Reports in Educational Achievement Testing. *Educational Measurement: Issues and Practice*, *23*(NOVEMBER 2004), 6–15. doi: 10.1111/j.1745-3992.2004.tb00164.x

Leighton, J. P., & Sternberg, R. J. (2003). Reasoning and problem solving. In A. F. Healy, R. W. Proctor, & I. B. Weiner (Eds.), *Handbook of psychology* (2nd ed., pp. 623–645). New York: Wiley.

Luecht, R. M. (2003). Assessment engineering task model maps, task models and templates as a new way to develop and implement test specifications. *Journal of Applied Testing Technology*, *14*.

Luecht, R. M. (2013). An Introduction to Assessment Engineering for Automatic Item Generation. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: theory and practice.* New York: TaylorFrancis/Routledge.

Marshalek, B., Lohman, D. F., & Snow, R. E. (1983). The complexity continuum in the radex and hierarchical models of intelligence. *Intelligence*, *7*(2), 107–127.

Meteorologisk Institutt. (n.d.). *Norske ekstremvær får navn.* Retrieved from [2016-03-11]`http://met.no/Meteorologi/A{_}varsle{_}varet/Varsling{_}av{_}farlig{_}var/?module=Articles;action=Article.publicShow;ID=246`

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, *101*(2), 343–352.

Mislevy, R., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, structures, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 61–90). Mahwah, NJ: Erlbaum.

Mislevy, R., Steinberg, L., Almond, R. G., Haertel, G. D., & Penuel, R. (2001). *Leverage points for improving educational assessment (PADI Technical Report 2)* (Tech. Rep.). Menlo Park, CA: SRI International.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A Brief Introduction to Evidence-centered Design. (July).

Mislevy, R. J., & Wu, P.-K. (1996). *Missing Responses and IRT Ability Estimation: Omits, Choice, Time Limits, and Adaptive Testing (Research Report No. RR-96-30)* (Tech. Rep.). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.1996.tb01708.x

Miyake, a., Friedman, N. P., Emerson, M. J., Witzki, a. H., Howerter, a., & Wager, T. D. (2000, aug). The unity and diversity of executive functions and their contributions to complex "Frontal Lobe" tasks: a latent variable analysis. *Cognitive psychology*, *41*(1), 49–100. doi: 10.1006/cogp.1999.0734

Mulholland, T. M., Pellegrino, J. W., & Glaser, R. (1980). Components of geometric analogy solution. *Cognitive Psychology*, *12*(2), 252–284.

Mullis, I., & Martin, M. (2013). *TIMSS 2015 assessment frameworks* (Tech. Rep.). Chestnut Hill, MA: Boston College.

Nielsen, J. (1994). *Estimating the number of subjects needed for a thinking aloud test* (Vol. 41). doi: 10.1006/ijhc.1994.1065

Nielsen, J., Clemmensen, T., & Yssing, C. (2002). Getting access to what goes on in people's heads? Reflections on the think-aloud technique. *Proceedings of the second Nordic conference on human-computer interaction - NordiCHI '02*, 101–110. doi: 10.1145/572020.572033

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. doi: 10.1037/0033-295X.84.3.231

OECD. (2016). *PISA 2015 Assessment and Analytical Framework*. OECD Publishing. doi: http://dx.doi.org/10.1787/9789264255425-en

Ohlsson, S. (2012). The Problems with Problem Solving: Reflections on the Rise, Current Status, and Possible Future of a Cognitive Research Paradigm. *The Journal of Problem Solving*, *5*(1), 101–128.

Paap, M. C. S., Lange, L., van der Palen, J., & Bode, C. (2015). Using the Three-Step Test Interview to understand how patients perceive the St. George's Respiratory Questionnaire for COPD patients (SGRQ-C). *Quality of Life Research*, 1–10. doi: 10.1007/s11136-015-1192-3

Pellegrino, J. W. (2003). Knowing What Students Know. *Issues in science and technology*, *19*(2), 48–52. Retrieved from `http://issues.org/19-2/pellegrino/`

Primi, R. (2001). Complexity of geometric inductive reasoning tasks contribution to the understanding of fluid intelligence. *Intelligence*, *30*(1), 41–70.

QSR International Pty Ltd. (2012). *NVivo qualitative data analysis Software, Version 10.* Retrieved from `http://www.qsrinternational.com/`

R Core Team. (2015). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from `http://www.r-project.org/`

Raven, J. (2000a). Psychometrics, cognitive ability, and occupational performance. *Review of Psychology*, *7*(1-2), 51–74. Retrieved from `http://eyeonsociety.co.uk/resources/CognitiveAbilityAndOccupationalPerformance.pdf`

Raven, J. (2000b). The Raven's Progressive Matrices: Change and Stability over Culture and Time. *Cognitive Psychology*, *41*(1), 1–48. doi: 10.1006/cogp.1999.0735

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin.

Simon, H. A. (1972). Theories of Bounded Rationality. *Decision and Organization*, *1*(1), 161–176.

Stanovich, K. E., Sá, W. C., & West, R. F. (2004). Individual Differences in Reasoning. In J. P. Leighton & R. J. Sternberg (Eds.), *The nature of reasoning* (pp. 375–409). Cambridge: Cambridge University Press.

Sternberg, R. J. (1985). *Beyond IQ: A Triarchic Theory of Human Intelligence.* CUP Archive. Retrieved from `http://www.google.no/books?hl=en{&}lr={&}id=jmM7AAAAIAAJ{&}pgis=1`

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. doi: 10.1016/0364-0213(88)90023-7

Tantau, T. (2015). *The TikZ and PGF Packages.* Retrieved from `http://sourceforge.net/projects/pgf`

The Psychometrics Centre. (n.d.). *Concerto platform for the development of on-line adaptive tests.* Retrieved from [2016-06-13]`http://www.psychometrics.cam.ac.uk/newconcerto`

Van der Linden, W., & Glas, C. A. W. (2009). *Elements of adaptive testing.* New York: Springer.

Vygotsky, L. S., Hanfmann, E., & Vakar, G. (2012). *Thought and Language.* MIT Press. Retrieved from `https://books.google.com/books?hl=en{&}lr={&}id=B9HClB0P6d4C{&}pgis=1`

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A Primer* (2nd ed.). Mahwah, NJ: ELawrence Erlbaum Associates.

Wainer, H., & Kiely, G. L. (1987, sep). Item Clusters and Computerized Adaptive Testing: A Case for Testlets. *Journal of Educational Measurement*, *24*(3), 185–201.

doi: 10.1111/j.1745-3984.1987.tb00274.x

Whitely, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*(1), 179–197.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag. Retrieved from `http://ggplot2.org/book/`

Winter, P. C., Kopriva, R. J., Chen, C.-S., & Emick, J. E. (2006). Exploring Individual and Item Factors that Affect Assessment Validity for Diverse Learners: Results from a Large-Scale Cognitive Lab. *Learning and Individual Differences*, *16*(4), 267–276. doi: 10.1016/j.lindif.2007.01.001

Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex problem solving - More than reasoning? *Intelligence*, *40*, 1–14.

Part III

Appendix

Appendix A

Protocol for the think-aloud study

(Norwegian version only)

Protokoll for thinkaloud

Forberedelser

- Test
  - Oversette instruksjoner
  - Alle deltakerne får den samme testen
    - Testlet 1, 5, og 8
    - Ett enkeltitem fra hvert enkelt av de resterende problemene
- Utprøving av prosedyre
- Finne deltagere
  - n = 6
- Lokaler
  - Uforstyrret
  - Nærhet til jaktmarker
- Tid
  - 20 min test
  - 10 min intervju
  - 30 min margin
  - Implisitt tidspress («jeg har estimert ca 1 min pr. spm»)

Prosedyre

1. Hei, takk, velkommen
2. Audio på
3. Informert samtykke – muntlig
4. «Du vil bli presentert med en rekke resonneringsproblemer som du skal løse på begrenset tid. Jeg er interessert i hvordan du resonnerer når du løser problemene, så det er veldig viktig at du snakker høyt mens du holder på. Dette er ikke alltid like enkelt å huske på, så hvis du eventuelt skulle slutte å snakke, vil jeg signalisere deg for å minne deg på å fortsette snakkingen. Jeg har regnet med at du vil bruke ca 1 minutt per spørsmål. De deltakerne som gjør det best vil bli belønnet etterpå. Noen informasjonsvinduer har kun en enkelt oppgave, mens andre har fire.»
   a. Presentasjon
   b. Avtale signal for å minne deltakerne på å snakke
   c. Estimert tid: 1 min per oppgave
   d. Belønning
   e. Deltakeren kan ikke få hjelp – ikke en dialog
   f. Hvis deltaker ikke gir adekvate data ➔ avbryte og irettesette
5. Presentert med en øvelsesoppgave på tid.
   a. 2 minutter
   b. Trenger ikke snakke høyt
6. Testen
   a. Husk å snakke høyt
   b. To faser: testlet og enkeltitems
   c. Mengde tid på hvert spørsmål: 1 minutt
7. Intervju ➔

Utstyr:

Papir til meg selv	Notisblokk
Skrivesaker	4 penner
Testen (utskrifter av de relevante oppgavene)	10 + 1
Flaxlodd	6
Lydutstyr	Pc-mikrofon
Protokoll	
Samtykkeskjema	
Kontaktskjema	

Appendix B

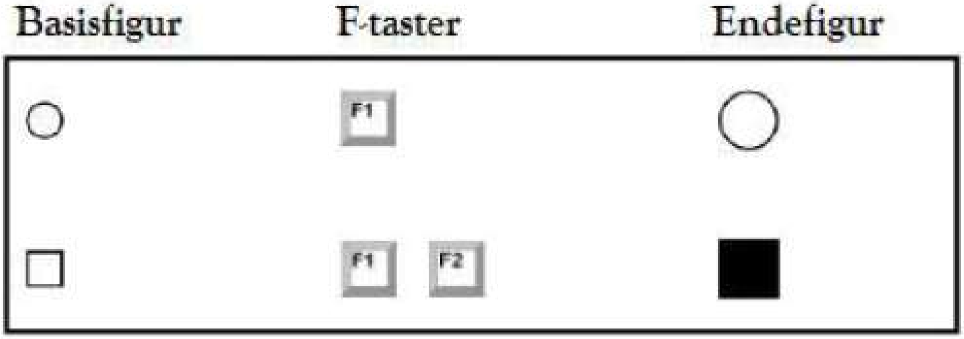
Instructions presented to the candidates before the think-aloud

(Norwegian version only)

INSTRUKSJONER

Denne øvelsen består av et informasjonsvindu med 2 figurer; en basisfigur og en endefigur. Mellom begge figurene befinner det seg noen F-taster. Settet med F-taster bestemmer hvilken operasjon som utføres på én eller flere basisfigurer. Din oppgave er å sammenligne basis-, og endefigurene med hverandre for å finne ut effekten av hver F-tast. Deretter må du svare på spørsmålene etter informasjonsvinduet.

LA OSS FØRST SE PÅ EKSEMPLET NEDENFOR



Den lille sirkelen er blitt en stor sirkel. Den lille firkanten forvandles til en stor, farget firkant. På både basis- og endefiguren er dermed størrelsen på symbolet endret. Bare det ble endret rekke, forandret fargen seg. Siden begge figurene modifiseres av F1, er denne F-tasten ansvarlig for endring av størrelse. F2-tasten endrer derfor fargen.

Husk at hver tast i et informasjonsvindu medfører en operasjon. Denne operasjonen kan påvirke en eller flere basisfigurer. Når en F-tast gjentatte ganger opptrer i samme informasjonsvindu forblir operasjonen den samme. Vær oppmerksom på at betydningen av F-tasten kan endres mellom de ulike informasjonsvinduene. Før du starter får du noen testspørsmål slik at du kan øve. På denne måten kan du gjøre deg kjent med hvordan du løser dem.

Gå nå til side 4 og 5 i spørsmålsheftet og løs eksempelspørsmålene. Angi svarene på svararket.
- Marker svarene på svararket.
- Visk fullstendig ut svar du ønsker å endre.

Er det et svar du er usikker på, angi det beste valget, men unngå å gamble.
Hvis du har fullført oppgaven, kan du levere inn svarheftet. Lykke til!

Appendix C

Protocol for the interview

(Norwegian version only)

Protokoll for intervju

Think-aloud ➔

1. Vanskelighetsgrad
   a. Var det noen oppgaver som var vanskeligere enn andre?
   b. Hvilke elementer er fremtredende?
   c. (Presentere min idé – gir den mening, kontrast med deltakerens idé)
2. Hva skal til for å løse oppgavene?
   a. Hvilke strategier?
   b. Hjelpemidler?
   c. Huskeregler?
3. Testlet vs enkeltitem
   a. Gambling
   b. Informasjon
   c. Hvordan taklet de endringene mellom en eller flere items per problem
4. Bakgrunnspørsmål
   a. Studieprogram
   b. Erfaring med spill
   c. Har de vært i en lignende situasjon tidligere?
5. Audio av
6. Takke for deltagelsen

Appendix D

NASA task-load index questionnaire

Figure 8.6

NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name	Task	Date

Mental Demand How mentally demanding was the task?

Very Low Very High

Physical Demand How physically demanding was the task?

Very Low Very High

Temporal Demand How hurried or rushed was the pace of the task?

Very Low Very High

Performance How successful were you in accomplishing what you were asked to do?

Perfect Failure

Effort How hard did you have to work to accomplish your level of performance?

Very Low Very High

Frustration How insecure, discouraged, irritated, stressed, and annoyed wereyou?

Very Low Very High

Appendix E

Correspondance with the Norwegian Social Science Data Services (NSD)

MELDESKJEMA

Meldeskjema (versjon 1.4) for forsknings- og studentprosjekt som medfører meldeplikt eller konsesjonsplikt
(jf. personopplysningsloven og helseregisterloven med forskrifter).

1. Intro

Samles det inn direkte personidentifiserende opplysninger?	Ja ● Nei ○	En person vil være direkte identifiserbar via navn, personnummer, eller andre personentydige kjennetegn. Les mer om hva personopplysninger. NB! Selv om opplysningene skal anonymiseres i oppgave/rapport, må det krysses av dersom det skal innhentes/registreres personidentifiserende opplysninger i forbindelse med prosjektet.
Hvis ja, hvilke?	□ Navn □ 11-sifret fødselsnummer □ Adresse ■ E-post ■ Telefonnummer □ Annet	
Annet, spesifiser hvilke		
Skal direkte personidentifiserende opplysninger kobles til datamaterialet (koblingsnøkkel)?	Ja ○ Nei ●	Merk at meldeplikten utløses selv om du ikke får tilgang til koblingsnøkkel, slik fremgangsmåten ofte er når man benytter en databehandler
Samles det inn bakgrunnsopplysninger som kan identifisere enkeltpersoner (indirekte personidentifiserende opplysninger)?	Ja ○ Nei ●	En person vil være indirekte identifiserbar dersom det er mulig å identifisere vedkommende gjennom bakgrunnsopplysninger som for eksempel bostedskommune eller arbeidsplass/skole kombinert med opplysninger som alder, kjønn, yrke, diagnose, etc. NB! For at stemme skal regnes som personidentifiserende, må denne bli registrert i kombinasjon med andre opplysninger, slik at personer kan gjenkjennes.
Hvis ja, hvilke		
Skal det registreres personopplysninger (direkte/indirekte/via IP-/epost adresse, etc) ved hjelp av nettbaserte spørreskjema?	Ja ○ Nei ●	Les mer om nettbaserte spørreskjema.
Blir det registrert personopplysninger på digitale bilde- eller videoopptak?	Ja ● Nei ○	Bilde/videoopptak av ansikter vil regnes som personidentifiserende.
Søkes det vurdering fra REK om hvorvidt prosjektet er omfattet av helseforskningsloven?	Ja ○ Nei ●	NB! Dersom REK (Regional Komité for medisinsk og helsefaglig forskningsetikk) har vurdert prosjektet som helseforskning, er det ikke nødvendig å sende inn meldeskjema til personvernombudet (NB! Gjelder ikke prosjekter som skal benytte data fra pseudonyme helseregistre). Dersom tilbakemelding fra REK ikke foreligger, anbefaler vi at du avventer videre utfylling til svar fra REK foreligger.

2. Prosjekttittel

Prosjekttittel	Explanatory item response modelling of an abstract reasoning assessment	Oppgi prosjektets tittel. NB! Dette kan ikke være «Masteroppgave» eller liknende, navnet må beskrive prosjektets innhold.

3. Behandlingsansvarlig institusjon

Institusjon	Universitetet i Oslo	Velg den institusjonen du er tilknyttet. Alle nivå må oppgis. Ved studentprosjekt er det studentens tilknytning som er avgjørende. Dersom institusjonen ikke finnes på listen, har den ikke avtale med NSD som personvernombud. Vennligst ta kontakt med institusjonen.
Avdeling/Fakultet	Det utdanningsvitenskapelige fakultet	
Institutt	Institutt for pedagogikk	

4. Daglig ansvarlig (forsker, veileder, stipendiat)

Fornavn	Johan	Før opp navnet på den som har det daglige ansvaret for prosjektet. Veileder er vanligvis daglig ansvarlig ved studentprosjekt. Veileder og student må være tilknyttet samme institusjon. Dersom studenten har ekstern veileder, kanbiveileder eller fagansvarlig ved studiestedet stå som daglig ansvarlig. Arbeidssted må være tilknyttet behandlingsansvarlig institusjon, f.eks. underavdeling, institutt etc. NB! Det er viktig at du oppgir en e-postadresse som brukes aktivt. Vennligst gi oss beskjed dersom den endres.
Etternavn	Braeken	
Stilling	Førsteamanuensis	
Telefon	22844826	
Mobil	22844826	
E-post	johan.braeken@cemo.uio.no	
Alternativ e-post	johan.braeken@cemo.uio.no	
Arbeidssted	Centre for Educational Measurement	
Adresse (arb.)	Molkte Moes vei 35	
Postnr./sted (arb.sted)	0371 OSLO	
Sted (arb.sted)	OSLO	

5. Student (master, bachelor)

Studentprosjekt	Ja ● Nei ○	Dersom det er flere studenter som samarbeider om et prosjekt, skal det velges en kontaktperson som føres opp her. Øvrige studenter kan føres opp under pkt 10.
Fornavn	Fredrik	
Etternavn	Helland	
Telefon	41083931	
Mobil	41083931	
E-post	fredrik@helland.org	
Alternativ e-post	fredrhel@mail.uio.no	
Privatadresse	Lakkegata 66 B	
Postnr./sted (privatadr.)	0562 OSLO	
Sted (arb.sted)	OSLO	
Type oppgave	● Masteroppgave ○ Bacheloroppgave ○ Semesteroppgave ○ Annet	

6. Formålet med prosjektet

Formål	The research goal is to reverse engineer an existing abstract reasoning test with the purpose to construct a modern test design framework that can be used to automatically generate items and computerized tailored or adaptive tests for abstract reasoning. In a first stage, a cognitive task analysis and think-a-loud-procedure (involving video recording and/or eye trackers) are imposed on the existing abstract reasoning test to further investigate and define important cognitive mechanisms as well as potentially relevant structural item properties. Later stages will mainly consist of explanatory IRT-modeling of an existing, anonymized, dataset, using information extracted from stage one.	Redegjør kort for prosjektets formål, problemstilling, forskningsspørsmål e.l.

7. Hvilke personer skal det innhentes personopplysninger om (utvalg)?

Kryss av for utvalg	□ Barnehagebarn ■ Skoleelever □ Pasienter □ Brukere/klienter/kunder □ Ansatte □ Barnevernsbarn □ Lærere □ Helsepersonell □ Asylsøkere ■ Andre	
Beskriv utvalg/deltakere	Enten skoleelever over 15 år på en ungsdomskole i Oslo, laveregradstudenter ved Universitetet i Oslo eller andre tilfeldige fra eget nettverk.	Med utvalg menes dem som deltar i undersøkelsen eller dem det innhentes opplysninger om.
Rekruttering/trekking	Enten rekruttering av tilfeldige interesserte på campus/i forelesninger eller fra egne nettverk/tidligere arbeidsplass (elever over 15 år). Rekrutteringen vil bli gjort av studenten. Randomisert utvalg er ikke interessant i forbindelse med hva det søkes om godkjenning for.	Beskriv hvordan utvalget trekkes eller rekrutteres og oppgi hvem som foretar den. Et utvalg kan trekkes fra registre som f.eks. Folkeregisteret, SSB-registre, pasientregistre, eller det kan rekrutteres gjennom f.eks. en bedrift, skole, idrettsmiljø eller eget nettverk.
Førstegangskontakt	Er det elever som brukes, vil førstegangskontakten bli formidlet via lærere og ledelse. Øvrige vil bli kontaktet av studenten. Insentivering kan forekomme hvis nødvendig.	Beskriv hvordan kontakt med utvalget blir opprettet og av hvem. Les mer om dette på temasidene.
Alder på utvalget	■ Barn (0-15 år) ■ Ungdom (16-17 år) ■ Voksne (over 18 år)	Les om forskning som involverer barn på våre nettsider.
Omtrentlig antall personer som inngår i utvalget	3-15	
Samles det inn sensitive personopplysninger?	Ja ○ Nei ●	Les mer om sensitive opplysninger.
Hvis ja, hvilke?	□ Rasemessig eller etnisk bakgrunn, eller politisk, filosofisk eller religiøs oppfatning □ At en person har vært mistenkt, siktet, tiltalt eller dømt for en straffbar handling □ Helseforhold □ Seksuelle forhold □ Medlemskap i fagforeninger	
Inkluderes det myndige personer med redusert eller manglende samtykkekompetanse?	Ja ○ Nei ●	Les mer om pasienter, brukere og personer med redusert eller manglende samtykkekompetanse.
Samles det inn personopplysninger om personer som selv ikke deltar (tredjepersoner)?	Ja ○ Nei ●	Med opplysninger om tredjeperson menes opplysninger som kan spores tilbake til personer som ikke inngår i utvalget. Eksempler på tredjeperson er kollega, elev, klient, familiemedlem.

8. Metode for innsamling av personopplysninger

Kryss av for hvilke datainnsamlingsmetoder og datakilder som vil benyttes	□ Papirbasert spørreskjema □ Elektronisk spørreskjema □ Personlig intervju □ Gruppeintervju ■ Observasjon □ Deltakende observasjon □ Blogg/sosiale medier/internett □ Psykologiske/pedagogiske tester □ Medisinske undersøkelser/tester □ Journaldata	Personopplysninger kan innhentes direkte fra den registrerte f.eks. gjennom spørreskjema,intervju, tester, og/eller ulike journaler (f.eks. elevmapper, NAV, PPT, sykehus) og/eller registre (f.eks.Statistisk sentralbyrå, sentrale helseregistre). NB! Dersom personopplysninger innhentes fra forskjellige personer (utvalg) og med forskjellige metoder, må dette spesifiseres i kommentar-boksen. Husk også å legge ved relevante vedlegg til alle utvalgs-gruppene og metodene som skal benyttes. Les mer om registerstudier her. Dersom du skal anvende registerdata, må variabelliste lastes opp under pkt. 15
	□ Registerdata	
	■ Annen innsamlingsmetode	
Oppgi hvilken	Filmet talk-aloud-prosedyre og eye-tracking i forbindelse med løsing av utvalgte deler av en abstrakt resoneringstest.	
Tilleggsopplysninger		

9. Informasjon og samtykke

Oppgi hvordan utvalget/deltakerne informeres	□ Skriftlig ■ Muntlig □ Informeres ikke	Dersom utvalget ikke skal informeres om behandlingen av personopplysninger må det begrunnes. Les mer her. Vennligst send inn mal for skriftlig eller muntlig informasjon til deltakerne sammen med meldeskjema. Last ned en veiledende mal her. NB! Vedlegg lastes opp til sist i meldeskjemaet, se punkt 15 Vedlegg.
Samtykker utvalget til deltakelse?	● Ja ○ Nei ○ Flere utvalg, ikke samtykke fra alle	For at et samtykke til deltakelse i forskning skal være gyldig, må det være frivillig, uttrykkelig og informert. Samtykke kan gis skriftlig, muntlig eller gjennom en aktiv handling. For eksempel vil et besvart spørreskjema være å regne som et aktivt samtykke. Dersom det ikke skal innhentes samtykke, må det begrunnes.
Innhentes det samtykke fra foreldre for barn under 15 år?	Ja ○ Nei ●	Les mer om forskning som involverer barn og samtykke fra unge.
Hvis nei, begrunn		
Innhentes det samtykke fra foreldre for ungdom mellom 16 og 17 år?	Ja ○ Nei ●	Les mer om forskning som involverer barn og samtykke fra unge.
Hvis nei, begrunn		

10. Informasjonssikkerhet

Spesifiser		NB! Som hovedregel bør ikke direkte personidentifiserende opplysninger registreres sammen med det øvrige datamaterialet.
Hvordan registreres og oppbevares personopplysningene?	□ På server i virksomhetens nettverk □ Fysisk isolert PC tilhørende virksomheten (dvs. ingen tilknytning til andre datamaskiner eller nettverk, interne eller eksterne) □ Datamaskin i nettverkssystem tilknyttet Internett tilhørende virksomheten □ Privat datamaskin ■ Videoopptak/fotografi □ Lydopptak ■ Notater/papir ■ Mobile lagringsenheter (bærbar datamaskin, minnepenn, minnekort, cd, ekstern harddisk, mobiltelefon) □ Annen registreringsmetode	Merk av for hvilke hjelpemidler som benyttes for registrering og analyse av opplysninger. Sett flere kryss dersom opplysningene registreres på flere måter. Med «virksomhet» menes her behandlingsansvarlig institusjon. NB! Som hovedregel bør data som inneholder personopplysninger lagres på behandlingsansvarlig sin forskningsserver. Lagring på andre medier - som privat pc, mobiltelefon, minnepinne, server på annet arbeidssted - er mindre sikkert, og må derfor begrunnes. Slik lagring må avklares med behandlingsansvarlig institusjon, og personopplysningene bør krypteres.
Annen registreringsmetode beskriv		
Hvordan er datamaterialet beskyttet mot at uvedkommende får innsyn?	Datamaskiner er passordbeskyttet. Filene vil i tillegg bli beskyttet med et eget passord. Filene lagres på en ekstern harddisk. Det resterende vil låses inne.	Er f.eks. datamaskintilgangen beskyttet med brukernavn og passord, står datamaskinen i et låsbart rom, og hvordan sikres bærbare enheter, utskrifter og opptak?
Samles opplysningene inn/behandles av en databehandler?	Ja ○ Nei ●	Dersom det benyttes eksterne til helt eller delvis å behandle personopplysninger, f.eks. Questback, transkriberingsassistent eller tolk, er dette å betrakte som en databehandler. Slike oppdrag må kontraktsreguleres.
Hvis ja, hvilken		
Overføres personopplysninger ved hjelp av e-post/Internett?	Ja ○ Nei ●	F.eks. ved overføring av data til samarbeidspartner, databehandler mm. Dersom personopplysninger skal sendes via internett, bør de krypteres tilstrekkelig. Vi anbefaler for ikke lagring av personopplysninger på nettskytjenester. Dersom nettskytjeneste benyttes, skal det inngås skriftlig databehandleravtale med leverandøren av tjenesten.
Hvis ja, beskriv?		
Skal andre personer enn daglig ansvarlig/student ha tilgang til datamaterialet med personopplysninger?	Ja ○ Nei ●	
Hvis ja, hvem (oppgi navn og arbeidssted)?		
Utleveres/deles personopplysninger med andre institusjoner eller land?	● Nei ○ Andre institusjoner ○ Institusjoner i andre land	F.eks. ved nasjonale samarbeidsprosjekter der personopplysninger utveksles eller ved internasjonale samarbeidsprosjekter der personopplysninger utveksles.

11. Vurdering/godkjenning fra andre instanser

Søkes det om dispensasjon fra taushetsplikten for å få tilgang til data?	Ja ○ Nei ●	For å få tilgang til taushetsbelagte opplysninger fra f.eks. NAV, PPT, sykehus, må det søkes om dispensasjon fra taushetsplikten. Dispensasjon søkes vanligvis fra aktuelt departement.
Hvis ja, hvilke		
Søkes det godkjenning fra andre instanser?	Ja ● Nei ○	F.eks. søke registereier om tilgang til data, en ledelse om tilgang til forskning i virksomhet, skole.
Hvis ja, hvilken	Eier av datasettet som skal brukes i IRT-delen har godkjent bruken av dette.	

12. Periode for behandling av personopplysninger

Prosjektstart Planlagt dato for prosjektslutt	01.08.2015 31.07.2017	Prosjektstart Vennligst oppgi tidspunktet for når kontakt med utvalget skal gjøres/datainnsamlingen starter. Prosjektslutt: Vennligst oppgi tidspunktet for når datamaterialet enten skalanonymiseres/slettes, eller arkiveres i påvente av oppfølgingsstudier eller annet.
Skal personopplysninger publiseres (direkte eller indirekte)?	□ Ja, direkte (navn e.l.) □ Ja, indirekte (bakgrunnsopplysninger) ■ Nei, publiseres anonymt	NB! Dersom personopplysninger skal publiseres, må det vanligvis innhentes eksplisitt samtykke til dette fra den enkelte, og deltakere bør gis anledning til å lese gjennom og godkjenne sitater.
Hva skal skje med datamaterialet ved prosjektslutt?	■ Datamaterialet anonymiseres □ Datamaterialet oppbevares med personidentifikasjon	NB! Her menes datamaterialet, ikke publikasjon. Selv om data publiseres med personidentifikasjon skal som regel øvrig data anonymiseres.Med anonymisering menes at datamaterialet bearbeides slik at det ikke lenger er mulig å føre opplysningene tilbake til enkeltpersoner. Les mer om anonymisering.

13. Finansiering

Hvordan finansieres prosjektet?	Det vil eventuelt bli søkt om midler gjennom masterstipendordningen ved UV-fakultetet, eller gjennom forskergruppen LEA.	

14. Tilleggsopplysninger

Tilleggsopplysninger		

Norsk samfunnsvitenskapelig datatjeneste AS
NORWEGIAN SOCIAL SCIENCE DATA SERVICES

NSD

Harald Hårfagres gate 29
N-5007 Bergen
Norway
Tel: +47-55 58 21 17
Fax: +47-55 58 96 50
nsd@nsd.uib.no
www.nsd.uib.no
Org.nr. 985 321 884

Johan Braeken
Institutt for pedagogikk Universitetet i Oslo
Postboks 1092 Blindern
0317 OSLO

Vår dato: 28.08.2015	Vår ref: 44083 / 3 / MSS	Deres dato:	Deres ref:

TILBAKEMELDING PÅ MELDING OM BEHANDLING AV PERSONOPPLYSNINGER

Vi viser til melding om behandling av personopplysninger, mottatt 21.07.2015. Meldingen gjelder prosjektet:

44083	*Explanatory item response modelling of an abstract reasoning assessment*
Behandlingsansvarlig	*Universitetet i Oslo, ved institusjonens øverste leder*
Daglig ansvarlig	*Johan Braeken*
Student	*Fredrik Helland*

Personvernombudet har vurdert prosjektet og finner at behandlingen av personopplysninger er meldepliktig i henhold til personopplysningsloven § 31. Behandlingen tilfredsstiller kravene i personopplysningsloven.

Personvernombudets vurdering forutsetter at prosjektet gjennomføres i tråd med opplysningene gitt i meldeskjemaet, korrespondanse med ombudet, ombudets kommentarer samt personopplysningsloven og helseregisterloven med forskrifter. Behandlingen av personopplysninger kan settes i gang.

Det gjøres oppmerksom på at det skal gis ny melding dersom behandlingen endres i forhold til de opplysninger som ligger til grunn for personvernombudets vurdering. Endringsmeldinger gis via et eget skjema, http://www.nsd.uib.no/personvern/meldeplikt/skjema.html. Det skal også gis melding etter tre år dersom prosjektet fortsatt pågår. Meldinger skal skje skriftlig til ombudet.

Personvernombudet har lagt ut opplysninger om prosjektet i en offentlig database, http://pvo.nsd.no/prosjekt.

Personvernombudet vil ved prosjektets avslutning, 31.07.2017, rette en henvendelse angående status for behandlingen av personopplysninger.

Vennlig hilsen

Katrine Utaaker Segadal

Marie Strand Schildmann

Kontaktperson: Kjersti Haugstvedt tlf: 55 58 29 53
Vedlegg: Prosjektvurdering

Dokumentet er elektronisk produsert og godkjent ved NSDs rutiner for elektronisk godkjenning.

Kopi: Fredrik Helland fredrik@helland.org

Personvernombudet for forskning

Prosjektvurdering - Kommentar

Prosjektnr: 44083

Ifølge prosjektmeldingen skal utvalget informeres muntlig om prosjektet og samtykke til deltakelse. For å tilfredsstille kravet om et informert samtykke etter loven, må utvalget informeres om følgende:

- at UiO er ansvarlig for studien
- prosjektets formål / problemstilling
- hvilke metoder som skal benyttes for datainnsamling, og en beskrivelse av instrumentene som anvendes
- hvilke typer opplysninger som samles inn
- at det er frivillig å delta og at man kan trekke seg når som helst uten begrunnelse
- dato for forventet prosjektslutt
- at data anonymiseres ved prosjektslutt
- at enkeltpersoner ikke vil kunne gjenkjennes i den ferdige oppgaven
- kontaktopplysninger til student/veileder

Forventet prosjektslutt er 31.07.2017. Ifølge prosjektmeldingen skal innsamlede opplysninger da anonymiseres. Anonymisering innebærer å bearbeide datamaterialet slik at ingen enkeltpersoner kan gjenkjennes. Det gjøres ved å:

- slette direkte personopplysninger (som navn/koblingsnøkkel)
- slette/omskrive indirekte personopplysninger (identifiserende sammenstilling av bakgrunnsopplysninger som f.eks. bosted/arbeidssted, alder og kjønn)
- slette digitale lyd-/bilde- og videoopptak

Fredrik Helland <fredrik@helland.org>

Prosjektnr: 44083. Explanatory item response modelling of an abstract reasoning assessment

1 e-post

Kjersti Haugstvedt <kjersti.haugstvedt@nsd.uib.no> 1. oktober 2015 kl. 15.15
Til: fredrik@helland.org, johan.braeken@cemo.uio.no

Personvernombudet viser til endringsmelding mottatt 28.09.15. Personvernombudet har registrert at det foretas lydopptak, og ikke bildeopptak i prosjektet. Det vil videre bli anvendt et kort spørreskjema. Vi forutsetter at informasjonsskrivet til utvalget oppdateres med hensyn til endringene, og viser ellers til vår tilråding av studien den 28.08.15.

--
Vennlig hilsen
Kjersti Haugstvedt
Spesialrådgiver
(Special Adviser)

Norsk samfunnsvitenskapelig datatjeneste AS
(Norwegian Social Science Data Services)
Personvernombud for forskning
Harald Hårfagres gate 29, 5007 BERGEN

Tlf. direkte: (+47) 55 58 29 53
Tlf. sentral: (+47) 55 58 81 80
Email: kjersti.haugstvedt@nsd.uib.no
Internettadresse www.nsd.uib.no/personvern