Chapter 1: Introduction
1 Introduction

1.1 Test item design in higher education

Testing and assessment are of major importance for education because assessments and tests determine to a large extent what students learn, how they plan their studies, and the effort they expend (Bishop, 2002; Crooks, 1988; Miller & Parlett, 1974; Popham, 2003). The form and content of tests further influence how students study and how they perceive the nature of the courses and the profession (Entwistle, 1996; Gielen, Dochy, & Dierick, 2003; Snyder, 1971).

In higher education, many courses are concluded with achievement tests to measure the students’ attainment. Therefore, many teachers in higher education have to develop such end-of-course exams (Anderson, 1987; Mavis, Cole, & Hoppe, 2001; Parkes, Fix, & Harris, 2003). Often, tests developed by teachers are referred to as in-house or teacher-made tests (Jozefowicz et al., 2002). Today, a growing number of teachers provide students with digital online activating learning materials or quizzes that lead up to these end-of-course exams (Nicol, 2007).

For both purposes, especially for large enrolment courses in undergraduate education, quizzes and tests consisting of selected response test item formats (such as true-false, multiple choice, matching, ordering, drag-and-drop, and numeric) are developed by teachers to keep grading within reasonable time limits and be able to cover a fairly broad number of topics in a short time span. In this thesis, these selected response test items, whether hard copy or electronic, will be called simply test items.

Individual test items form the core building blocks of these tests and the test items determine the quality of those tests; in the words of Haladyna (1997): “Good tests consist of good test items.” It is the test item that determines which cognitive processes are elicited, whether more or less informed students are appropriately identified, and students’ perceptions of the nature of the subject domain.

Further, designing test items is a recurring task for teachers in higher education for two main reasons. First, especially in higher education, the content of courses
changes often due to new insights in the domain or curricular considerations, so
new test items need to be devised. Second, because tests are often made public after
having been administered to provide students with opportunities to learn or to
allow them to check the correctness of the scoring, those questions become unfit
for use in subsequent tests, forcing teachers to produce test items again and again.

It has been reported that test items developed by teachers in higher education
suffer from low quality (DiBattista & Kurzawa, 2011; Hansen & Dexter, 1997;
Holsgrove & Elzubeir, 1998; Jozefowicz et al., 2002). Sometimes teachers are not
aware of common construction errors or are not able to avoid common
construction errors. Such errors can confuse student leading them astray in
answering correctly or – in contrast – such errors can unintentionally lead students
to the correct answer without them knowing the materials. But also, teachers may
solely design test items that merely quote information verbatim from instructional
materials or only call for recognition or recall of information without any context
and therefore not for ‘higher-order thinking’ or low levels of cognitive demand.

Querying for higher-order thinking or learning outcomes such as application of
knowledge, insight or problem solving is a need often expressed by teachers and
learners. It is reported however that though teachers do try to query for higher-
order learning outcomes such as critical thinking or problem solving, they are not
able to design such items successfully and thus fall back on test items that indeed
merely elicit recall of factual information (Buckles & Siegfried, 2006). Test items
that only elicit recognition of factual information encourage students to use surface
learning approaches (Trigwell & Prosser, 1991; Vermunt, 2005) in which pure
memorization and rote learning are most effective (Beattie, Collins, & McInnes,
1997; Boud & Falchikov, 2007; Smith & Miller, 2005; Struyven, Dochy, & Janssens,
2003; Vermunt, 2005). Though for certain subjects such as language learning or
anatomy, this kind of knowledge acquisition and testing may be appropriate, such
studying strategies are considered inappropriate for higher education in general
(Biggs, 1996; Boud & Falchikov, 2007; Draper, 2009; Entwistle & Entwistle, 2003;
Struyven et al., 2003). Tests that contain solely test items that query
decontextualized information most probably reinforce the low appreciations of
students and the general public regarding the value of selected response test items
(see for example Beattie et al., 1997; Cizek, 2001) and the conviction that selected response test items cannot query for high cognitive demand.

For that matter however, Rodriguez (2003) argued that it is in particular the stimulus of the test item that is of most importance when thinking about cognitive demand. The review study of Rodriguez suggested that correlations between the scores on selected response test items (multiple choice answering mode) and constructed response test items (open ended question answering mode) were high when the stimulus (the task or problem presented) of the two different response modes were comparable. Therefore, generating problems or tasks with stimuli aiming at querying for understanding, critical thinking or problem solving requiring higher cognitive demand is essential for selected response test items to fulfill that role.

In this thesis, such higher-order learning outcomes are operationalized as outcomes in which learners need to demonstrate that they have knowledge and skill to use learnt facts, concepts, principles and procedures in familiar but new situations within the bounds of a course content. In this thesis, such new situations are considered to be test questions in which information has been reframed, added or excluded, or in which new problems, contexts or answering options are offered. Such new information requires the student to reevaluate, reinterpret and transform the learnt materials in order to deal with these new situations that are presented in the stimuli and answering options of the test item. The level of cognitive demand in those circumstances is a function of the examinee’s stance on the continuum from novice to expert. This continuum (novice-expert) is broad so the cognitive demand is never really known unless one questioned the examinee. Higher-order thinking in this thesis is thus related to the learner’s ability to show transfer (Bransford & Schwartz, 1999; Chi, Glaser, & Farr, 1988; Wiggins & McTighe, 1998) of knowledge and skill to comparable, yet different propositions, problems or situations.

It is therefore vital for higher education that teachers be able to generate a range of test items that are appropriate for the variety of goals for which test items are needed. Teachers should be able to design test items not only for rote learning
purposes but also and more importantly for querying higher-order learning goals such as application of knowledge, critical thinking, and problem solving.

Writing good-quality test items, however, is not a straightforward or easy task to execute. It has been reported that developing high-quality test items takes substantial amounts of time (Case, Holtzman, & Ripkey, 2001) and training (Downing & Haladyna, 2006); the task is regarded as difficult, especially in terms of developing sufficiently numerous plausible and distractors (Mayenga, 2009).

In practice of course, teachers are not specifically equipped and trained to design test items. Typically, teachers in higher education work alone when generating test items and rarely seek or receive support (Cizek, Fitzgerald, & Rachor, 1995; Conole & Warburton, 2005; DiBattista & Kurzawa, 2011; Downing, 2002; Jozefowicz et al., 2002; Popham, 2003; Sahari, 1999; Van der Vleuten et al., 2004). In practice, teachers are not quick to use an item-writing guide or handbook. Further, in formal one-time learning events such as workshops or training, the information obtained or skills learned can be forgotten if there is no context in which they are shared, fostered, and put into practice (Knight, Tait, & Yorke, 2006).

Optimal solutions to raise the expertise of teachers, the quality of test items, and the process of test item design entail more flexible approaches and organizational contexts such as communities of practice, peer review, and on-the-job training (Knight et al., 2006; Wenger, 1998). However, in the realities of higher education, instituting such learning contexts for teachers is seldom practical or affordable (Iramaneerat, 2012; Warburton, 2009). Thus, the present situation leads to test items in higher education produced in large quantities with the weaknesses reported above, which is highly problematic for ensuring the quality of assessment and learning.

Rather than simply stating unhelpful slogans that teachers need to be better trained or expend more effort, this thesis raises fundamental questions about what designing test items actually entails. Studying that craft at a more elemental level should offer opportunities to find meaningfully different and likely superior approaches to support teachers. Indeed, one of the key problems in test item design is that how teachers’ designing of test items actually functions remains largely
unknown (Downing & Haladyna, 2006; Hattie, Jaeger, & Bond, 1998; Welch, 2006); according to Haladyna, “the cognition that is used in writing the item is still a mystery” (personal communication, May 15, 2013). Analyzing and understanding the cognitive process of test item development is thus likely an important conditional step in being able to reduce problems regarding test item design, because in higher education it is to a large extent the triad of the teacher’s knowledge and skill regarding the subject matter, students’ abilities and perceptions, and test item design that determines the nature and quality of the test items developed. It is the extent and interrelation of the teachers’ knowledge and skill regarding these three factors and the effort they exert that combined determine the quality and quantity of the outcome of the design process.

Understanding the cognitive process by which teachers design test items can help to reveal and refine techniques to improve the design process. Studying and describing the process of test item design is therefore taken as a first, crucial step in this thesis.

1.2 Perspectives on test item design

Test item development has long been viewed as a psychometric measurement problem; all uncontrollable variables and errors in the process and outcome of developing such test items were regarded as undermining the tests’ psychometric properties and validity (Schuwirth & Van der Vleuten, 2006). This perspective has resulted in an emphasis on rational and often complex methods to develop test items (see, e.g., Gierl & Haladyna, 2012; Haladyna & Shindoll, 1989; Karamanis, Ha, & Mitkov, 2006; Kopriva & Winter, 2012; Oosterhof, Rohani, Sanfilippo, Stillwell, & Hawkins, 2008; Wilbrink, 1983a; Williams & Haladyna, 1982), an emphasis on structured training (Case et al., 2001; Downing & Haladyna, 2006; Iramaneerat, 2012), and an emphasis on using sets of guidelines to prevent test items from being flawed (Haladyna, Downing, & Rodriguez, 2002). These approaches stress the deductive nature of test item design, relying on a view that the learning goals and taxonomic categories of the course or instructional material lead to a single correct, objective test item answer. In this respect, convergent thinking (Guilford, 1967), which focuses on finding a single well-established
solution to a problem according to algorithmic rules. The literature implicitly presents test item design as a well-defined problem to solve.

As the deductive nature of testing and test item design is emphasized, teachers may be under the assumption that they should follow a logical line of thinking and be able to take a large number of restrictive criteria into consideration simultaneously. At the extreme, content expertise may even be viewed as secondary to form and construction.

In the view outlined above, quality is seen as the absence of mistakes in the design process or outcome. However, the absence of construction mistakes is merely a necessary and hardly a sufficient condition for a high-quality test item, which begins with a fundamentally sound idea of the test item, whether for verbatim recognition and recall or for test items that reach beyond that level. First, this idea must be based on an important topic or subject, such as one laid down in a test blueprint (Downing & Haladyna, 2006); this idea must then be reworked into an item with a reworded, paraphrased, or differently represented setup of the original instructional information (Anderson, 1987), for which new ideas are also needed. Second, in order to query for higher-order learning outcomes such as application of knowledge, critical thinking, and problem solving, high-quality test items should present variants, examples, counterexamples, or misconceptions of the facts, concepts, and principles contained in the instructional materials (Anderson, 1987; Crouch & Mazur, 2001; Haladyna, 1997). Alternatively, test items should contain new relations between facts, concepts, and principles to test how well students understand them (Anderson, 1987; Crouch & Mazur, 2001; Haladyna, 1997). In order to extend the meaningfulness of test items, they should present professionally or personally relevant problems and elicit student engagement (Biggs, 1996; Downing & Haladyna, 2006; Haladyna, 2004; Keller, 1983; Merrill, 2002; Struyven et al., 2003; Van der Vleuten & Schuwirth, 2005). Querying for higher-order learning outcomes in particular calls for the teacher to generate novel conceptions of existing information.

There is thus a need for teachers to be able to design test items that probe for a range of purposes, especially items that present novel material to students, which
demands of teachers the ability to produce fresh approaches to and representations of the tested material, not just once, but year after year. Manifold ideas are needed because teachers must design test items throughout their careers and because multiple solutions must be considered.

Intensive thought processes on the part of the teacher are needed for both the generation of verbatim recall test items and those that query higher-order cognition. Such cognitive processes necessarily involve, besides convergent thinking, *divergent* thinking and production (Acar & Runco, 2012; Guilford, 1967). Divergent thinking refers to cognitive processes that generate ideas by exploring many possible solutions and the relations between them. During and after the process of divergent thinking, ideas and information become better organized and, ideally, optimized by convergent thinking.

Though the aspect of creativity with respect to test item design has been noted as important by several authors (Osterlind, 1998; Popham, 1984; Rikers, 1988; Vale, 1995), the subject of divergent thinking and production regarding test item design has not yet received serious scholarly attention. The literature reveals that very limited attention has been paid to divergent thinking in test item design.

As it is clear that a more fundamental approach to the test item design process is needed to alleviate problems concerning test item design in higher education, the first step is to devote attention to the divergent production of test items in the design process. This approach forms the pivotal point of study in this thesis.

Further, if creativity is an essential part of test item design, that task can be regarded as solving a *creative design problem*, which implies solving *ill-defined* problems. This notion of an ill-defined problem is articulated when analyzing the task of test item design with Reitman’s (1964, 1965) classic three main distinguishing characteristics for solving ill-defined problems; if the start state, goal state, and transformation function are not well specified, a problem is ill-defined. The problem or goal of test item design simply cannot be clearly laid out; for example, achievement tests and quizzes are supposed to query students on courses’ important topics, but who decides, and how, what is important? Who decides, and
how, how many test items then should be presented? Who decides, and how, that a
given test measures student achievement well enough? Who knows what cognitive
processes students exhibit when answering test items? All these issues are either
negotiable (Van der Vleuten, 1996) or not directly observable (Dembo & Howard,
2007; Haladyna, 2004; Hibbison, 1991). Teachers have to make decisions on the
basis of their idiosyncratic and incomplete knowledge of the domain and the
specific contexts and effects of questioning, which results in a form of naturalistic
decision making (Zsambok & Klein, 1997) during the design process. Further, how
teachers design test items is not governed by algorithms; there are multiple
approaches to generating and designing test items.

There are thus boundaries regarding both the problem space of some specified
course content or curriculum and the solution space of a test consisting of selected
response test item to be produced, but these boundaries simply cannot be defined
precisely; what occurs between and across them, in the movement from problem to
solution, is open-ended.

Perceiving test item design as a design problem has been approached by Wilbrink
(1983), but he presented the process as solving a seemingly well-defined problem.
The concept of design and the design process of test items has not been explored by
scholars as solving an ill-defined problem. Further, expertise in test item design
plays a role in being able to design engaging and clear test items effectively. Item-
writing authors acknowledge this issue (Gronlund, 1998; Haladyna, 2004;
Osterlind, 1998; Wilbrink, 1983), but touch only very briefly upon the concept of
expertise regarding test item design. They do not explore this concept in
meaningful depth, observing only that it requires a practice-feedback loop to
become an expert (Downing & Haladyna, 2006); an important gap in research
exists here.

Teachers are in general left to their own devices in designing test items, while
traditional approaches to training or supporting them in this task have not worked
very well. Most importantly, despite the volume of research on test item design, it is
not at all clear what designing test items actually entails, especially from the
perspective of the cognitive processes that teachers employ to generate them. If we
are to support teachers more effectively in designing good test items, we must first understand more clearly the processes involved in designing them. This has some far-reaching consequences for the study and support of the design and development of test items.

This thesis thus has as its first main topic the goal of understanding – as a design problem – the cognitive processes exhibited by teachers involved in test item design. Designing test items is understood as an ill-defined problem-solving design task in which divergent production of test items is at least equally as important as convergent production of test items. Second, once this process is more adequately described and understood, the thesis outlines selected specific, novel teacher support interventions that can be designed and evaluated for both their appropriateness and effectiveness.

These two goals reflect my thinking and acting as an industrial design engineer. As a designer, I strive to solve problems for individuals by designing fit-for-purpose artifacts. Such artifacts must be designed based on a thorough understanding of the problem to solve, built and tested to assess whether these artifacts actually solve the problem in theory and practice. In analogy, I tried to understand the test item design task thoroughly, generate and design instruments to provide support for this task, develop prototypes and test the prototypes. Theory about designing test items is hence equally important as the development and evaluation of appropriate instruments.

1.3 Aim of this thesis

The aim of this study is to contribute to knowledge regarding the process of test item design, especially in the higher education context, by two main contributions. The first is to develop a cognitive process model for designing test items that provides a fuller understanding of the cognitive process and provides leads to generate supportive interventions for teachers, bearing in mind that “there’s nothing so practical as good theory” (Lewin, 1951, p. 169). The second contribution is to discover and develop support for teachers based on that new understanding of
the test item design process and evaluate the appropriateness and effectiveness of that support.

1.4 Research questions and methods

In view of the aim of the study, six research questions are put forward, with multiple research methods employed to answer them.

Chapter 2 contains the following research question in order to describe and understand more fully the cognitive processes involved in test item design:

1. *What are the characteristics of a process model for test item design that is based on the concept that the test item design task is an ill-defined creative design problem-solving task?*

This question is answered inductively by constructing a cognitive model for test item design based on research regarding the different approaches to ill-defined creative design problem solving. In the process model constructed, divergent and convergent cognitive processes involved in test item design are treated as the most important processes for designing such items; divergent processes receive more attention.

Second, as it is the divergent cognitive process of the teacher in higher education that is the focus of this thesis, it seeks out and describes practical divergent guidelines and examples. To that end, in Chapters 3 and 4, the following research questions are put forward:

2. *Which practical inspirational guidelines to support divergent production of test items for teachers in higher education can be identified, and what is their effectiveness?*

3. *Which design patterns to support divergent production of selected response test items can be identified in teacher-generated test items, and how can they be described?*

The primary concept of research questions 2 and 3 is to describe what ideas for divergent production of test items can be identified among teachers in higher
education. Such practical inspirational guidelines could serve as support for teachers in general. In terms of research methods, the author and teachers from a number of higher education institutions worked together, resulting in several case studies with multiple embedded units of analysis (Yin, 2003) involving teacher evaluation of the outlined guidelines as a form of participant validation (Paugh, 2004).

Third, research question 4 is derived from the premise that, given that teachers design test items largely on their own, computer support could be a valuable means to support them in the divergent and convergent production of test items. Research into computer support systems, however, does not provide clear-cut guidance on the design of such a system. In Chapter 5, therefore, the following research question is put forward:

4. What are the main requirements for a computer support tool to support divergent and convergent production of selected response test items, and how can these requirements be met by a computer application?

This study is carried out in the form of a “prototyping study” (Becker & Berkemeyer, 2002; Buxton, 2010; Ozenc, Kim, Zimmerman, Oney, & Myers, 2010; Tripp & Bichelmeyer, 1990). This type of study refers to a practice adapted from engineering and software design by which one or more designs (conceptual descriptions, drawings, mock-ups, global dimensions, wireframes, rationales, etc.) are developed to study and define features of a product, which is then used to elicit impressions and appraisals regarding the expected usefulness of such a product and inspection of the adequacy of the rationale underlying the design. The prototype itself is used in the subsequent experimental study presented in Chapter 6 of this thesis, providing evidence that the tool fulfilled its basic functionality of storing test

1 The term design study in Chapter 5 does therefore not refer to ‘design based research’ or ‘educational design research’ as often used in educational research (Barab & Squire, 2004; Van den Akker, Gravemeijer, McKenney, & Nieveen, 2006).
items and providing support information to the user in a clear but unobtrusive manner.

Finally, in Chapters 6 and 7, research question 5 and 6 are put forward. Because the previous research questions regarding test item design are answered with a qualitative research approach, an experimental quantitative approach is needed to establish to what extent interventions in the design process in controlled conditions do indeed result in better test items. A few specific guidelines, techniques and design process conditions, as identified in Chapter 3, are therefore worked out in detail. The following research questions are then put forward:

5. To what extent can divergent and convergent production of test items be improved by encouragement to diverge, developing a concept map, presenting item shells, and presenting classic test item writing guidelines?

6. To what extent does the presentation and use of an idea leaflet result in test items that are more original as compared to not using an idea leaflet,

a) when the idea leaflet is presented prior to a test item design task and used in the first phase of idea generation?

b) when the idea leaflet is presented and used after initial exhaustion of ideas?

To answer these research questions, two experimental between-subject design studies are conducted to determine the extent to which specific interventions intended to support teachers in the process of generating test items have an actual effect on the quality of test items. In the studies, participants in control and experimental conditions receive the task to generate test items based on a piece of instructional information. The participants must develop as many good test items as possible. Using regression analyses, the effects of the interventions are studied.

1.5 Significance

The literature regarding educational measurement pays scant attention to the process or theory of test item development, and a number of leading authors in this
field, such as Osterlind (1998) and Downing and Haladyna (2006), have highlighted this gap. In particular, the divergent production of test items has scarcely been studied. This thesis contributes to the understanding of the characteristic difficulties and support options regarding the process of designing test items.

The thesis also contributes to the growing body of knowledge concerning creative problem solving and design by studying a new problem type. The project fills a niche concerning general problem types. These well-studied problems are the well-defined and smaller problems such as math, reasoning, figural, Tower of Hanoi, and physics problems (Chi, Feltovich, & Glaser, 1981; Davidson & Sternberg, 2003), and problems such as the design of the large, complex physical and non-physical artifacts such as houses, warships, computers, instruction, services, political issues, or naturalistic decision-making problems like military problems (Cross, 2007; Jonassen, 1997; Jonassen & Mandl, 1990; Rowland, 1993; Zsambok & Klein, 1997). A gap in research exists regarding the design of relatively small, non-physical artifacts in an ill-defined context that must still comply with many restrictive rules and must be repeatedly generated, such as magazine columns, three-panel comics, or jokes. For that reason, designing test items takes a special position in the research into problem solving and design; it can in fact be regarded as prototypical for that category. The thesis further contributes to research regarding computer environments for creative problem solving (Hewett, 2005; Mulet & Vidal, 2008; Shneiderman, 2002), as a concept for a computer support tool is developed.

From a practical viewpoint, this study contributes to solving a widespread problem in higher education. The task of developing achievement tests, exams, and activating learning material consisting of test items is widely regarded as difficult. Research into the process of developing test items and means of supporting teachers in higher education has the potential to benefit many individuals. The developed process model could make the task of developing test items more insightful and raise the skill and interest of teachers in designing test items, leading to better test items. Eventually, this improvement could provide enhanced
opportunities to deliver assessment programs that fulfil their obligation to achieve maximum learning impact for learners.

1.6 Outline of this study

In Chapter 2, the importance of designing test items that are both original and technically sound is presented. It is argued in detail that constructing test items is a form of solving ill-defined problems; each test item to be produced presents its own small design problem. A process model is developed regarding the cognitive process that are most likely involved in designing test items. This model shows that divergent and convergent thinking and production are the main ingredients in test item design. The model also intends to be used prescriptively to support the test item design process. The implications and capabilities of the model are discussed.

In Chapter 3, a number of practical, inspirational guidelines that can be used as interventions to support teachers in higher education regarding the divergent production of content for test items are generated and classified. These guidelines are used in a number of small-scale test item development projects and evaluated for their usefulness.

Chapter 4 follows with a discussion about the possible capacities of the concept of design patterns, which refers to the fact that experienced designers have stored a number of solutions for recurring design problems that they can readily apply to new but similar problems. Design patterns are seen as a supplement for the process model in Chapter 2 to stimulate divergent production of test items. In Chapter 4, the relevant dimensions for design patterns and concrete design patterns are described and their capacities are discussed.

In Chapter 5, a concept for a support tool is presented, based on the process model in Chapter 2. The guiding principles for the design and requirements are identified and the main user interface elements, key functions, and resources incorporated into the tool are described. Arguments derived from the literature on computer tools for creative problem solving and user interface design are provided to support the expected appropriateness of the tool.
Chapter 6 presents an empirical study regarding the extent to which deliberate diverging and converging leads to more, more original, or technically higher-quality test items. Based on the process model developed in Chapter 2, a between-subjects design experiment is conducted. In particular, the effect of the development of a concept map prior to developing test items and the effect of providing test item developers with development prompts to aid in the design of test items are studied.

In Chapter 7, the effect of providing participants with a specific general creativity technique in the form of an idea leaflet to stimulate divergent production for test items is explored. Again, this creativity technique functions as an intervention derived from the process model in Chapter 2.

In the final chapter, a summary of the previous chapters and key findings are presented. The attainment and limitations of the cognitive process model developed and the findings of the studies will be integrated according to the main lines of the thesis and discussed in the appropriate level of detail.
1.7 References


Chapter 1


Introduction


