Design of a competence-based assessment system for air traffic control training

Citation for published version (APA):

Oprins, E. (2008). *Design of a competence-based assessment system for air traffic control training*. [Doctoral Thesis, Maastricht University]. Maastricht University. <https://doi.org/10.26481/dis.20081002eo>

Document status and date:

Published: 01/01/2008

DOI:

[10.26481/dis.20081002eo](https://doi.org/10.26481/dis.20081002eo)

Document Version:

Publisher's PDF, also known as Version of record

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy

If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

**Design of a competence-based
assessment system
for air traffic control training**

Luchtverkeersleiding Nederland
Air Traffic Control the Netherlands

The work presented in this thesis was conducted at the Human Factor department of
Air Traffic Control the Netherlands, Schiphol Airport.

Photos: Rob Berghege (Luchtverkeersleiding Nederland)
Printed by: Datawyse | Universitaire Pers Maastricht

ISBN 978-90-5278-736-7

© E.A.P.B. Oprints, Luchtverkeersleiding Nederland, Schiphol 2008.

Design of a competence-based assessment system for air traffic control training

Ontwerp van een competentiegericht beoordelingssysteem
voor de opleiding van luchtverkeersleiders

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit Maastricht,
op gezag van de Rector Magnificus,
Prof. mr. G.P.M.F. Mols,
volgens het besluit van het College van Decanen,
in het openbaar te verdedigen
op donderdag 2 oktober 2008 om 14.00 uur

door

Esther Anna Petronella Barbara Oprins

geboren op 29 mei 1970
te Gilze-Rijen

Promotor

Prof. dr. R.A. Roe

Beoordelingscommissie

Prof. dr. W.H. Gijssels (voorzitter)

Prof. dr. J.J.G. van Merriënboer (Open Universiteit Heerlen)

Prof. dr. M.J. Waller

Prof. dr. F.R.H. Zijlstra

Contents

Voorwoord.....	9
1 Introduction.....	11
1.1 Problem description	11
1.2 Objective of the study	15
1.3 Scope of the study	15
1.4 Design methodology	16
1.5 Design cycle	16
1.6 Outline.....	20
PART 1: ANALYSIS	23
2 Performance and competence acquisition	25
2.1 Performance	25
2.2 Competence	31
2.3 Competence acquisition	34
2.4 Conclusions	40
3 Competence acquisition in ATC	41
3.1 The air traffic control task	41
3.2 ATC competence acquisition.....	46
3.3 Training design for ATC.....	48
3.4 Conclusions	54
4 Assessment design for ATC training	55
4.1 Assessment and learning	55
4.2 Performance assessment.....	60
4.3 Assessment in ATC-related learning environments	65
4.4 Conclusions	69
5 ATC competence analysis	71
5.1 Task and competence analysis	71
5.2 The ATC Performance Model.....	76
5.3 Description of the components of the ATC Performance Model.....	79
5.4 Conclusions	81

6	Analysis of the design problem: Program of Requirements	83
6.1	Methods	83
6.2	The redesign problem.....	85
6.3	Program of Requirements	88
6.4	Conclusions	92
PART 2: DESIGN & IMPLEMENTATION		95
7	Redesign of the assessment system	97
7.1	The design process	97
7.2	Competence-based assessment.....	98
7.3	Training design	102
7.4	Properties of the assessment system	107
7.5	Design of a web-based assessment tool.....	118
7.6	Conclusions	127
8	Implementation	129
8.1	The process of implementation	129
8.2	Training of coaches and assessors.....	131
8.3	Conclusions	133
PART 3: EVALUATION		135
9	Evaluation of psychometric quality	137
9.1	Interrater agreement.....	137
9.2	Rating errors	150
9.3	Test reliability	169
9.4	Conclusions	174
10	Evaluation of learning processes: analysis of learning curves	175
10.1	Learning curves	176
10.2	Methods	187
10.3	Results	188
10.4	Discussion.....	205
10.5	Conclusions	211
11	Evaluation of learning processes: competence development	213
11.1	Development of competences	213
11.2	Methods	215
11.3	Results	217
11.4	Discussion.....	225
11.5	Conclusions	228

12	User evaluation	229
12.1	Methods	229
12.2	Results	230
12.3	Conclusions	232
13	Evaluation of Program of Requirements.....	235
13.1	Methods	235
13.2	Results	236
13.3	Discussion.....	241
13.4	Conclusions	242
14	General conclusions	243
14.1	Assessment systems for complex skill acquisition	243
14.2	The LVNL assessment system	247
14.3	The design process	250
14.4	Implications for further research	251
	References	253
	Summary	263
	Samenvatting	269
	Appendix A: ATC work and training at LVNL.....	277
	Appendix B: Assessment reports ACC	284
	Appendix C: Data structure.....	295
	List of abbreviations	297
	Glossary of ATM terminology	299
	Index	301
	Curriculum Vitae.....	305

6	Analysis of the design problem: Program of Requirements	83
6.1	Methods	83
6.2	The redesign problem.....	85
6.3	Program of Requirements	88
6.4	Conclusions	92
	PART 2: DESIGN & IMPLEMENTATION	95
7	Redesign of the assessment system	97
7.1	The design process	97
7.2	Competence-based assessment.....	98
7.3	Training design	102
7.4	Properties of the assessment system	107
7.5	Design of a web-based assessment tool.....	118
7.6	Conclusions	127
8	Implementation	129
8.1	The process of implementation	129
8.2	Training of coaches and assessors.....	131
8.3	Conclusions	133
	PART 3: EVALUATION	135
9	Evaluation of psychometric quality	137
9.1	Interrater agreement.....	137
9.2	Rating errors	150
9.3	Test reliability.....	169
9.4	Conclusions	174
10	Evaluation of learning processes: analysis of learning curves	175
10.1	Learning curves	176
10.2	Methods	187
10.3	Results	188
10.4	Discussion.....	205
10.5	Conclusions	211
11	Evaluation of learning processes: competence development	213
11.1	Development of competences	213
11.2	Methods	215
11.3	Results	217
11.4	Discussion.....	225
11.5	Conclusions	228

Voorwoord

Dit proefschrift zou niet bestaan hebben zonder de hulp en bijdragen van vele anderen, die ik middels dit voorwoord hiervoor wil bedanken. Het zal mij helaas niet lukken om volledig te zijn, en daarom wil ik beginnen met diegenen die niet expliciet vermeld worden. Er gaat veel (onzichtbaar) werk achter dit proefschrift schuil dat onder andere te maken heeft met de praktische uitvoering van de opleidingen bij Luchtverkeersleiding Nederland. Heel hartelijk bedankt voor alle ondersteuning, adviezen, middelen, productie- en regelwerk, enzovoort!

Als eerste wil ik graag mijn promotor noemen, Prof. dr. Rob Roe. Rob, ik heb ontzettend veel geleerd in afgelopen jaren van onze gesprekken, die ik als heel plezierig heb ervaren. Het viel voor mij niet altijd mee om het praktische werk naar een hoger, wetenschappelijk niveau te tillen, maar uiteindelijk hebben we samen toch een manier gevonden om dit in een proefschrift uiteen te zetten. Ik heb ook veel gehad aan het zéér zorgvuldige en gedetailleerde commentaar (inclusief het Engels!) op alles wat ik inleverde. We hadden niet alleen met elkaar te maken in de verhouding promotor – promovendus, maar we werkten ook heel prettig samen in het herontwerp van de selectie. Ik ben dan ook erg blij dat we in een nieuw project onze samenwerking kunnen voortzetten. En wie weet ontmoeten we elkaar nog eens in onze gezamenlijke hobby: de klassieke muziek. Rob, erg bedankt voor alles!

Luchtverkeersleiding Nederland (LVNL) heeft mij de mogelijkheid geboden om mijn werk als Expert te combineren met het schrijven van dit proefschrift. Ik wil de organisatie hier heel hartelijk voor bedanken. Binnen LVNL zijn er een aantal personen geweest die een belangrijke bijdrage geleverd hebben. Dit zijn degenen die ik steeds bedoel met 'we' in dit proefschrift: we hebben het beoordelingssysteem echt *samen* ontworpen. Allereerst zijn dit de (toenmalige) Training Managers van Unit Training, Ernst Burggraaff en Hans van Weerdenburg.

Ernst, je hebt voor mij een hele bijzondere rol gespeeld tijdens de totstandkoming van dit proefschrift. Zonder jouw waardevolle input, creatieve ideeën, hulp en steun was het niet gelukt om een dergelijk systeem praktisch werkbaar en geaccepteerd te krijgen bij jouw achterban op zaal. Ik heb onuitwisbare herinneringen aan onze samenwerking in de ACC opleidingen, aan alle presentaties die we samen hebben gegeven op congressen, OJT-dagen, CuCo-dagen, enzovoort. Je hebt ook nog concrete bijdragen geleverd zoals de aanlevering van plaatjes en de woordenlijst, en je hebt elk paper kritisch gereviewed. Jouw (persoonlijke) feedback heeft me steeds erg geholpen. Ik kon altijd en voor alles bij je terecht . . . heel erg veel dank!

Hans, we hebben heel wat bereikt in de TWR/APP opleidingen. De invoering van het nieuwe beoordelingssysteem betekende een grote verandering, en dankzij jouw gedrevenheid en enthousiasme is dit heel succesvol verlopen. Je mag nog steeds trots zijn op de competentiewerkshops! We hebben veel beleefd in onze periode van intensieve samenwerking, waar ik met een heel goed gevoel op terug kijk. Dankjewel voor je betrokkenheid, ondersteuning en gezelligheid!

Ook andere verkeersleiders behoren tot 'we' in dit proefschrift. Dit zijn de (toenmalige) Cursus Managers van ACC (o.a. Emiel, Manfred, Diko, Marc & Mark) en van TWR/APP (o.a. Hans, Eric, Allard, Arjen). De CuCo's horen ook in dit rijtje thuis. Jullie zijn degenen geweest die de 'inhoud' hebben aangeleverd, bijvoorbeeld in Handboeken en beoordelingsformulieren. Nog belangrijker, jullie waren altijd enthousiast en betrokken bij nieuwe ontwikkelingen in je eigen cursus. Mannen, ik heb extreem genoten van alle discussies en (informele) gezelligheid bij CuMa-vergaderingen, CuCo-dagen, enzovoort. Ik heb ontzettend gelachen... erg bedankt!

Daarnaast zijn de coaches en leerlingen onmisbaar geweest, ook als proefpersonen in het evaluatieonderzoek hoewel dit niet bij iedereen bekend is. Zonder noemenswaardige problemen, hooguit wat tegenstribbeling bij het gebruik van het LVS, konden veranderingen in de opleiding doorgevoerd worden. Dankjewel voor jullie medewerking en positieve feedback!

Niet alleen verkeersleiders, maar ook 'kantoormensen' van LVNL zijn belangrijk geweest voor mij en het proefschrift. Esther Geven en Marloes Bos-Valkenburg: jullie zijn niet voor niks mijn paranimfen! Esther, we zijn samen gestart met het project herontwerp selectie en dat was vooral in het begin een spannende, maar leuke periode. Niet alleen op werkinhoudelijk gebied (en qua naam) passen we bij elkaar, maar bovenal ben je een erg fijne vriendin bij wie ik altijd kan aankloppen! Superbedankt voor alle persoonlijke gesprekken, je lieve steun en wijze raad! Marloes, je begon als stagiaire waarin je een stukje van mijn onderzoek hebt gedaan en nu ben je mijn opvolgster in Unit Training. Het is heerlijk om met jou samen te werken, en je was een ontzettend fijne kamergenoot! Ik mis onze onderonsjes soms wel... Bedankt voor al je hulp, vertrouwen, en gezelligheid!

De andere Experts (o.a. Hans, Marian, Inge, Lilian, Dori) en overige collega's op de Human Factor afdeling wil ik bedanken voor de reviews, feedback, collegialiteit, prettige samenwerking, en voor het voortzetten van ontwikkelingen in de opleiding zoals beschreven in dit proefschrift.

Tenslotte ben ik speciale dank verschuldigd aan de General Manager van deze afdeling, Jurgen van Avermaete. Jurgen, ik heb jouw persoonlijke betrokkenheid altijd erg gewaardeerd. Dankjewel voor je steun bij de realisatie van dit proefschrift, en voor het bieden van nieuwe mogelijkheden en uitdagingen in mijn werk!

Dit proefschrift heb ik voornamelijk in mijn vrije tijd geschreven. Ik dank mijn familie en (muziek)vrienden en vriendinnen voor het feit dat ze mij steeds weer doen inzien hoe belangrijk andere dingen in het leven zijn. Tot slot, Joost, dankjewel voor je voortdurende enthousiasme en trots!

Esther Oprins

1 Introduction

The study described in this thesis addresses a problem faced by organizations that are highly dependent on the performance of human operators of complex human-machine systems, working in safety-critical environments. Examples of such organizations are providers of air traffic control services, airlines, rail and shipping companies, nuclear and chemical plants, and armed forces. Any mistake of the individual operator may cause an unsafe situation that can ultimately result in great damage to the environment or to human life and property. Consequently, the standards for the operators' competences are usually very high. Only a minority of candidates appears to be able to acquire these competences within a reasonable training period.

The application of high standards for individual competences in recruitment, selection and training results in low numbers of operators, sometimes too small to provide the required operational capacity. Many of the organizations that suffer from operator scarcity look for possibilities to raise the number of operators without lowering the performance standards. Over the years, various possibilities have been explored. One of them is the search to highly sophisticated systems to protect human performance against its limitations. Another option, investigated in this study, is based on the assumption that a potential solution lies in the improvement of training so that more candidates are able to attain the required competences.

The field of research addressed here is air traffic control (ATC), and the focus of this study is on the design of a competence-based assessment system for ATC training. The study is based on a project that has been carried out at Air Traffic Control the Netherlands (LVNL).

1.1 Problem description

In this section, we define the problem to be solved in this study and we explain how this problem has led to the redesign of the assessment system at LVNL (Oprins, Burggraaff & Van Weerdenburg, 2005, 2006a, 2006b, 2007a, 2008¹).

¹ Parts of the text in this section (and in other sections) are literally copied from these journal articles (2006a, 2008) or conference papers (2005, 2006b, 2007a).

The nature of the ATC task

The primary aim of air traffic control (ATC) is to expedite and maintain a safe and orderly flow of air traffic (International Civil Aviation Organization, 2004). To prevent collisions, controllers keep aircraft separated from each other by providing pilots with headings, levels, and speeds. Besides safety, controllers must meet efficiency requirements to avoid unnecessary delay for flights. Other tasks are providing traffic information and navigational assistance. Nowadays, advanced communication systems and radar equipment are available for controllers to do their job (Nolan, 1999)².

Like in other process control tasks in transportation (e.g., aviation, shipping, railways) or process industries (e.g., chemical, nuclear plants), the ATC task is considered highly complex and dynamic. Processing large amounts of dynamically changing information calls for complex cognitive processes. Therefore, ATC is also called a complex cognitive skill or high-performance skill, especially in relation to learning processes (Schneider, 1990; Van Merriënboer, 1997). The strict safety requirements make the performance standards even higher because any (human) error caused by incompetence of controllers must be avoided. Growing traffic intensity and environmental regulations have increased the workload for controllers during the last decade.

Due to the nature of the task, its learnability is limited. Selection requirements are high in the most air navigation service providers (ANSP's), but the output from training often remains low, especially in organizations serving busy and complex airports such as Schiphol Airport (LVNL). This may result in a shortage of operational controllers. High failure rates in training are undesirable because of time and costs. Training normally starts in expensive high-fidelity simulators, followed by time-consuming on-the-job training (OJT). Trainees are usually trained one-by-one by operational controllers, whose availability is limited.

Solutions to increase the training output have been sought in different areas. First, various research projects have been launched to reduce the complexity of the air traffic management (ATM) system with the help of more advanced technology (CAST, 1999; SESAR, 2007). However, it seems that human performance in this area cannot easily be replaced by computer systems (Shebilske, Goettle & Garland, 2000). In addition, the amount of traffic in the airspace keeps increasing. This makes it unrealistic to think that the task difficulty will decrease much in the near future. Second, all over the world attempts have been made to optimize the selection of controllers (e.g., MRU, 1993; Ramos, Heil & Manning, 2001; Eissfeldt, Heil & Broach, 2002). This research is also in progress since it appears to be difficult to make valid predictions about a candidate's future performance.

The third possible solution is the subject of this study: the improvement of training. The quality of both selection and training has to be very high to ensure the procurement of successful controllers in the most cost-effective manner (Wickens, Mavor & McGee, 1997). A more efficient and effective training system should gener-

² See Appendix A for a more detailed description of the job of air traffic controllers with the focus on the situation at Air Traffic Control the Netherlands.

ate a higher number of well-trained controllers (Salden, Paas, Broers & Van Merriënboer, 2004).

Design of ATC training

The complex cognitive processes involved in ATC result in two main problems for the design of ATC training. The first problem concerns the question which competences are required to become an 'expert'. The idea is that experts apply more sophisticated strategies needed to perform the predominantly cognitive tasks (Anderson, 1995). However, the automated and unconscious character of the cognitive processes in ATC makes it very difficult to explicate these strategies and to define the required competences. Many cognitive task analyses in ATC have been done (EATCHIP, 1997, 1999, 2002; Lamoureux & Cox, 1992; Hadley, Guttman & Stringer, 1999; Redding & Seamster, 1994), but analyses of the competences that have to be acquired to perform these tasks are rare. Besides, the trainability of competences versus innate personal characteristics is still an issue (Ackerman, 1988, 1989, 1990; Regian & Schneider, 1990). Identifying the limitations of trainability is important because the emphasis in training design should be on the learnable aspects, while selection design should aim at prerequisite cognitive abilities, personality, and other features that are needed to learn the job.

The second problem involves the question how these competences are acquired. An integral learning theory about the acquisition of complex cognitive skills such as ATC does not exist yet. There is a large body of evidence and theory about the acquisition of simple skills (Anderson, 1982, 1983), but the findings cannot be transferred to complex skill acquisition immediately (Schneider, 1990; Taatgen & Lee, 2003). Nevertheless, some attempts have been undertaken to investigate the acquisition of ATC skills in the last decades. For instance, the Kanfer-Ackerman task, a simplified simulated ATC task, is used for modeling learning processes (Ackerman, 1988, 1989; Browne & Du Toit, 1991; Lee & Anderson, 2001; Taatgen, 2001; Taatgen & Lee, 2003). Other research has focused on dynamic task selection in ATC training, also based on a simulated ATC task (Salden et al., 2004; Salden, 2005), or on strategy learning (Schneider, 1990; Nunes & Mogford, 2003). Although this kind of research can be used to get more insight into learning processes in ATC, it remains difficult to translate its findings into practical guidelines on how to train trainees.

In this context, attention should be paid to the specific learning environments in which ATC competences are acquired. Especially worth mentioning, is the use of simulators, which allows the training of candidates until the highest possible levels of performance without any risk of fatalities, before they enter into on-the-job training (OJT). Theories about the design of complex skill training in simulators, mainly developed for the military domain, have to be developed further (Farmer, Van Rooij, Riemersma, Jorna & Moraal, 1999). For example, discussion exists about appropriate task selection (e.g., part-task versus whole-task training; Van Merriënboer, 1997), and the amount of guidance by the coach (Petraglia, 1998). On-the-job training often relies on informal learning (Bolhuis & Simons, 1999; Onstenk, 1997). Learning at the workplace in which a coach explicates his knowledge to a pupil has

always existed, but growing needs for competence development of employees in organizations has increased the interest in 'structured on-the-job training' (Rothwell & Kazanas, 1994; Jacobs & Jones, 1995).

Assessment in ATC training

The subject of this study is assessment, which is an important part of training design. Assessment in simulator and on-the-job training is usually referred to as *performance assessment*. This can be characterized by direct observation of the learner's behavior in a realistic context (Berk, 1986).

A valid and reliable assessment system may contribute to an increase in the number of competent controllers as output from training in several ways. First, assessment is a base for feedback that supports the trainee's learning process. Assessment makes it possible to adapt the training to the trainee's needs and to define appropriate training interventions, based on the trainee's strengths and weaknesses (deficiencies). In this way, assessment helps to increase the chance for trainees to successfully complete the training. Assessment should make clear which competences have been developed and which do not yet meet the performance standards of the training phase. Second, more accurate judgments of assessors lead to better founded pass-fail decisions. This may help to restrict the number of prospectless trainees at expensive training positions (false positives) and to reduce the number of failed trainees who could have had a chance to become competent controllers (false negatives). Assessment often has a selective character due to the high task demands and the safety requirements. Third, predictive validity of selection may be improved by using more reliable and valid training results in validation studies. The selection and training system (including assessment) should preferably be an integrated system, with coherence between the different parts, and with feedback-loops from training to selection.

Unfortunately, despite its importance, scientific research on assessment in ATC training appears to have been very limited. This study is one of the first studies that examined this subject for this particular domain. The scarce literature is restricted to descriptions of existing assessment methods from a practical point of view (Hopkins, 1995; EATCHIP, 2000). Some studies on ATC selection seem relevant, because they involve assessment in work samples or criterion development for validation (e.g., Ramos, Heil & Manning, 2001; Eissfeldt, Heil & Broach, 2002). Within the domain of aviation, a substantial amount of research has been done on assessment of aircrew (e.g., Flin & Martin, 2001; O'Connor, Hormann, Flin, Lodge & Goeters, 2002). The ATC task, however, is different.

Despite of the lack of domain-specific research on assessment, other research on performance assessment in educational and organizational settings has been conducted over the last decades. Within organizational contexts, performance assessment by means of simulators has been applied for training in various domains, including process control (e.g., Sanders, 1991; Vreuls & Obermayer, 1985). A great deal of research on performance assessment at the workplace has focused on performance appraisal (e.g., Murphy & Cleveland, 1995). Literature about assessment

design in on-the-job training is very scarce, probably because learning on the workplace usually has an informal character.

The LVNL project

This study is the result of a project that was performed at the training department of Air Traffic Control the Netherlands (LVNL). In this project, the assessment system in simulator and on-the-job training was redesigned. The main organizational problem to be solved at LVNL was to reduce the existing shortage of operational controllers by an increase of the training output. A more valid and reliable assessment system, based on a newly defined set of competences for air traffic controllers, should improve the efficiency and effectiveness of training, also for the benefit of trainees and coaches. Other measures to improve the quality of training were made in coaching, task selection, training content, training structure, length of training phases, and learning climate. An important aim was to make the expensive and safety-critical on-the-job training more efficient by providing a better structure. The selection system was redesigned simultaneously. This was a separate project, although selection and training are closely related.

1.2 Objective of the study

The objective of the project was to design an assessment system that optimally supports the acquisition of air traffic controller competences in simulator and on-the-job training. This thesis describes its design. We present both the *product* characteristics of the assessment system, and the design *process* itself. This study also aims at providing insight into assessment design in a broader context. General characteristics of the assessment system, the design process, or the evaluation methods may be applicable in other ANSP's, in related domains with complex skill acquisition in safety-critical environments (e.g., aviation, process industry), and in comparable learning environments (simulators, OJT).

1.3 Scope of the study

LVNL has an internal training program for various air traffic controller functions. Selected candidates are engaged as employees of LVNL from the start of the training period; training and working are alternated. The project described in this study involves Unit Training defined in European standards (EATCHIP, 2000) at Schiphol Airport³.

Training at LVNL begins with Initial Training. This focuses on prerequisite theory and basic ATC skills in accordance with European regulations (ESARR, 2000). Its content is generic for ATC in Europe and is designated as Common Core Content

³ In Appendix A the LVNL training system is described in more detail.

(CCC). Unit Training is specific for the different ATC functions at various airports. Unit Training results in a 'rating' (cf. license) for a specific ATC function and is divided into two main parts. Trainees are first submitted to preOJT, which mainly consists of simulator training besides some theory lessons. Next, trainees enter on-the-job training (OJT). The assessment system was primarily designed for Unit Training as a combined system of preOJT and OJT, but the main properties are also applied in Initial Training. The assessment system was not only implemented at Schiphol Airport, but as well at other airports in The Netherlands (Beek, Eelde, Rotterdam). Unit Training is divided into various subsequent courses. We focus on one course as main case in the study: Area Control (ACC), preOJT and OJT. We selected this course because it was the first course in which we implemented the new system.

1.4 Design methodology

A universal method to (re)design assessment systems does not exist. It depends on the typical properties of assessment design which methods are appropriate, for instance, when designing performance appraisal systems, assessment for training, or assessment in selection systems. Assessment for training is usually integrated in training design; it is not considered as a distinct product. The emphasis of the assessment design literature is typically on the product of assessment rather than on the process of designing itself. This implies that desirable assessment properties are described, while it is left open how the assessment system should be designed. Furthermore, the practical context for which assessment is designed is often ignored, although the context determines the final properties of the system strongly. In order to arrive at a methodology that is appropriate for the design of an assessment system embedded in training redesign as carried out at LVNL, we consider the generic design cycle of Roozenburg and Eekels (1991) as well as the training design methodology of Moonen (2000). We combine them in a specific design methodology that seems most suitable for the purpose of the present study.

1.5 Design cycle

Roozenburg and Eekels (1991) consider 'design' to be a form of problem solving, because it involves a search for means to achieve specific goals. The design cycle is based on the general empirical cycle for problem solving, but it has been made more specific for the design of products. Although the design cycle was originally proposed for industrial design, it can also be applied in other design contexts. The design cycle has been used to design products related to assessment such as personnel selection systems (Roe, 1989, 2005), simulations used as job samples in selection systems (Ridderbos, 1992), and personnel appraisal systems (Roe & Daniels, 1994). It can be applied in design as well as in redesign processes (Roe, 1989). The design cycle is presented in Figure 1.1.

Figure 1.1. Design cycle (Roozenburg & Eekels, 1991)

Analysis

The first step is to establish the purpose and functions of the new product. Designers have to formulate the design problem. Design assignments are often not well defined yet, but contain only global indications of the final product. A *Program of Requirements (PoR)* contains criteria based on an analysis of the functions to be fulfilled. These criteria have to be formulated as concretely as possible, and have to specify the required and desirable properties of the product. Besides requirements, the list may include constraints, for instance, time restrictions or costs. This PoR guides possible solutions for the design problem. It is important that all those people concerned with the product ('stakeholders') are involved in the establishing of the PoR: end users and others who have interest in the product.

Synthesis

The second step in the design process is to make a provisional design of the product, which accords with the PoR. Creative thinking processes form the central part. Possible solutions for the design problem are presented in the form of a global description, sketch, or model, depending on the type of product and the phase in the design process. The value of a solution has to be proven in the next phases of simulation and evaluation. This step is called 'synthesis', because separate or known parts are combined to a new whole. Creativity is needed to be able to generate new

ideas for the design. Three categories of creative methods can be used: associative methods (e.g., brainstorming techniques), creative confrontation methods (e.g., making use of analogy), and analytical systematic methods (e.g., function analysis). The assumption is that there is no single best solution.

Simulation

In the next step, expected properties of the product are deduced by reasoning and modeling. This is called 'simulation'. This deductive process precedes the realization and use of the final product. Simulation is strongly related to prediction, because the result of simulation involves a prediction about the properties of the product. Usually, models are made that simulate the working of the final product. This may vary from abstract mathematical models to high fidelity simulations. Many methods can be used, including scientific research methods and theories. Experiential knowledge often plays an important role, for instance, because scientific methods cannot be applied yet.

Evaluation

Afterwards, evaluation against the PoR is needed. This implies that the value and quality of the provisional design are investigated by a comparison between the expected properties and the required properties in the PoR. Differences with the expectations will always exist; therefore, judgments are needed whether the differences are acceptable or not.

Decision

Finally, a decision is made whether the (provisional) design can be elaborated upon further or whether an alternative has to be generated. Usually, more attempts are needed before a design proposal will be accepted and realized finally. One may return to the step of 'synthesis' (create a new design proposal) or return to the step of 'analysis' (redefine the Program of Requirements). First versions may be adjusted or extra requirements may appear to be important.

According to Roozenburg & Eekels (1991), a design process always consists of a sequence of intuitive (reductive) and deductive steps. There exist many alternative solutions and this results in a necessarily iterative process. The design cycle can be considered as an iterative 'trial and error' process: the cycle is usually passed through more than once. The different steps do not necessarily follow each other in time; they are principally distinguished by methodological characteristics. Intermediate products that are the result of a 'phase' in the design process often reflect the number of times that the cycle is passed through. For instance, provisional design one is a prototype that is tested with potential users; next, a more elaborated design two, based on revisions, is further evaluated and so on.

Domain-specific methodologies are available to describe the different phases in the design process. They indicate which intermediate products have to be made and which domain-specific methods have to be used. For instance, for training design purposes, task analysis is needed to determine the training objectives, while for the design of technology functional specifications have to be described.

Training design methodology

Instructional design theories are often used as a framework for training design (e.g., Romiszowski, 1988; Gagné, 1985; Merrill, 1994). The majority of assessment systems developed in this context are part of the design of training as a whole. Assessment design as such plays a minor role, and the design process itself is hardly given any consideration. Methodologies for the design of information systems in education and training seem to be applicable. Regardless whether the design involves web-based training applications or assessment for simulator training, the same steps need to be taken. The design process usually comprises five main phases: (1) analysis, (2) design, (3) development, (4) implementation, and (5) evaluation.

Moonen (2000) presents an alternative view. He focuses on the design of educational information systems and argues: "there is no single right approach and therefore no single right solution for design and development problems (. . .). From practical experiences, it often seems to be that the design and development process of each specific artifact is an unique experience for which only global methodological support can be given" (p. 23). This conclusion is concurrent with the perspective of Roozenburg and Eekels (1991) on design processes. Generally, design processes are very complex and influenced by the context, such as the existing learning situation and the interests of end users.

Moonen (2000) developed the '3-space design strategy' to deal with these problems. The 3-space design strategy distinguishes between three *activity spaces* to emphasize the cyclic character of designing: a *consensus space* focused on negotiating temporary agreements, a *task space* focused on prototyping, and a *implementation space* focused on end-user tailoring. These are called 'spaces' instead of 'phases' to avoid the suggestion that they might be executed in a linear order. Each space is a combination of the *structured or product-oriented* design approach, which can be used when the design problem is well-defined and when there is agreement about the end product, and the *associative or process-oriented* design approach, which emphasizes the relationship between product and intended user in an iterative process. This view of the design process is more in line with modern developments towards social-constructivism emphasizing social interaction processes.

The design strategy of Moonen (2000) illustrates that the design process is complex and cannot be executed in a structured linear order. Prototypes are tested, while users are negotiating about the advantages and disadvantages. This results in improving the prototype and so on. The 'implementation space' in which this product is developed, presented, and implemented in the real context with users, seems to be a crucial part in training design processes.

Design methodology in this study

In our project, like in other redesign projects, intermediate products and steps-in-between are not clearly distinguished. A possible explanation is that the design process was executed within the organization itself. Designers and users (cf. controllers) were working together closely. In addition, the assessment system was gradually redesigned. Some parts were renewed first and implemented directly, followed

by gradual improvements of other parts. This makes it difficult to indicate how many times and in what way the design cycle of Roozenburg & Eekels (1991) was passed through. In terms of the design methodology of Moonen (2000), it implies that the processes in the 'consensus space', the 'task space', and the 'implementation space' overlapped.

We propose to combine and integrate the views of Roozenburg & Eekels (1991) and Moonen (2000) and to model the design process as depicted in Figure 1.2. We describe the redesign process of the assessment system from a methodological point-of-view adopting the design cycle of Roozenburg & Eekels (1991). But unlike Roozenburg & Eekels we incorporate *implementation* as part of the redesign process in accordance with Moonen (2000). This results in three main phases in the redesign process in the present study as illustrated in Figure 1.2; the phases of Roozenburg & Eekels are displayed in cursive text. Figure 1.2 shows that the design process is divided into three main phases, being a cyclic process: (1) *analysis*, (2) *design & implementation*, and (3) *evaluation*. The phases 'synthesis' and 'simulation' in the design cycle of Roozenburg & Eekels (1991) are part of the second phase, and 'decision' is incorporated in the third phase. We added 'implementation' to the second phase, based on the ideas of Moonen (2000).

Figure 1.2. Design cycle in the present study

1.6 Outline

The outline of the study is based on the three-phase view of the design process presented in Figure 1.2. We used this division into phases to describe the design of the assessment system at LVNL.

Part 1: Analysis

The part that reflects the analysis phase of the design process comprises three activities. Chapter 2, 3 and 4 provide a *literature review* covering theoretical theories and domain-specific applications that may be useful for the design of the assessment system for ATC training in the project. Chapter 5 describes a *competence analysis*, based on findings from the literature and judgments generated by experts

from LVNL. This resulted in a set of competences that should be acquired in ATC training and therefore should be assessed. The subject of chapter 6 is the construction of a *Program of Requirements* (Roozenburg & Eekels, 1991) as a result of documentation studies, observations and interviews with stakeholders. Extensive analysis of the specific training and assessment situation at LVNL was needed to identify the purposes and functions of the new assessment system.

Part 2: Design and implementation

This phase of the design process focuses on the steps of 'synthesis' and 'simulation' in the design cycle of Roozenburg & Eekels (1991). These steps were alternating and integrated in the process. Creative processes were dominating in the step of 'synthesis'. Therefore, we applied typical methods such as brainstorming with end users and analogies with comparable assessment systems in other ANSP's. We discussed the expected properties of the product, being predicted in the 'simulation' step of the design cycle, with end users. We organized try-outs to test parts of the products (first versions) in order to find out whether the expected properties were present. There did not exist a clear chronological order of steps. We could say that this phase is most illustrative for the intersection between the 'consensus space', the 'task space', and the 'implementation space' according to Moonen (2000). Chapter 7 describes the main properties of the assessment system and how it is positioned in the training system. This chapter also addresses the activities of the *design* process itself. *Implementation* activities, considered being a specific 'space' according to Moonen (2000), are subject of chapter 8. They include the introduction of the assessment system to the user group and the development of a special training program for assessors.

Part 3: Evaluation

Chapters 9 to 13 describe the evaluation of the assessment system. Based on the results, we made decisions about adjustments on the system in accordance with the design cycle of Roozenburg & Eekels (1991). The evaluation is divided into three parts: (1) evaluation of *psychometric quality* (chapter 9); (2) evaluation of learning processes, sub divided into analysis of *learning curves* (chapter 10) and analysis of *competence development* (chapter 11); (3) *end user evaluation* (chapter 12). The evaluation part concludes with an overall evaluation against the *Program of Requirements* (chapter 13). Finally, chapter 14 presents the conclusions of this study.

Area controllers at work in the operational room of LVNL.

Area controller at work in the operational room of LVNL.

PART 1: ANALYSIS

This part refers to the *analysis* phase in the design cycle. This phase comprises three activities:

- *Literature review* (chapter 2, 3 and 4) aims at getting more information about issues related to the subject of this study that can be used in the design of the assessment system. First, general theories about performance, competences, and learning are described (chapter 2). Second, domain-specific literature about competence acquisition in air traffic control is reviewed (chapter 3). Finally, general theories about assessment and applications in ATC-related environments are discussed (chapter 4).
- *Competence analysis* (chapter 5) aims at analyzing what has to be acquired in ATC training. This competence analysis resulted in the so-called 'ATC Performance Model' that was chosen as a framework for training and assessment design at LVNL.
- *Analysis of the design problem* (chapter 6) aims at defining the purposes, requirements and constraints of the assessment system to be designed. The result of this analysis was the construction of a Program of Requirements.

2 Performance and competence acquisition

This chapter provides a review of the literature about performance, competence and competence acquisition, including the relations between them. We discuss general theories in organizational and educational psychology with our objective of designing an assessment system for training in mind. First, we define performance and we address several influences on performance (section 2.1). Second, we pay attention to the concept of competence, especially in relation to performance and learning processes (section 2.2). Third, we describe the process of competence acquisition, based on general learning theories (section 2.3).

2.1 Performance

Performance is a central notion in organizational, educational, and cognitive psychology. We summarize the most important literature. We explain the meaning of performance and its relations with other notions (e.g., task, information processing, strategies) as far as they are relevant for our study.

Definition

Research on performance has a long history in the literature, especially in work and organizational psychology. The concept of performance is used in many different contexts. This makes a uniform definition rather difficult to make. We adopt a definition of Roe (1999, p. 234) that consists of two components, *process* and *outcome*:

- "Performance is the process by which people try to achieve a given work goal.
- Performance is the congruence between the work goal and the outcome of the process by which people try to achieve that work goal."

This implies for individual performance in work situations that the goal corresponds to the work task and that the process is a series of work actions, including mental and physical activities; the outcome is the individual work result. The definition of Hockey (1997, p. 77), which is comparable with the definition of Roe (1999) but adds cognitive aspects, accentuates underlying mental processes: "It is normally understood to refer either to (a) the effectiveness of specific skills in meeting (typically externally-imposed) cognitive goals, or (b) the underlying mental operations associated with such behavior".

Task

Performance is supposed to be inherently connected with the task to be executed. The level of performance is dependent on task specification, task demands, task difficulty, task extensiveness, task complexity, task temporality and so on (Roe, 1999). In action regulation theory, performance always implies some goal-directed activity. A distinction is made between the 'objective task' and the 'subjective task' (Hacker, 1998; Roe, 1999). The objective task is the 'task-as-given', the subjective task is the 'task-as-taken'. Tasks are generally not executed as they are given, but they are interpreted and redefined by performers themselves. Personal and situational factors make the task rather subjective.

Information processing

According to human information processing theory, actions are mainly guided by information processing. Many information-processing models have been developed in the last decades (e.g., Wickens, 1992; Rasmussen, 1986; Pribram & McGuinness, 1992; Card, Moran & Newell, 1986). These information-processing models make use of the analogy with a computer system: information is processed in different stages. In each stage, specific transformations of information are performed.

Most models of information processing have in common that they have an input and an output system. The input system is concerned with perception of incoming stimuli; the output system is concerned with motor responses. Decision-making and response selection determine the execution of responses, dependent on attentional resources (Wickens, 1992). This process in-between is regulated by the 'cognitive system' as opposed to the 'perceptual system' (input) and 'motor system' (output) (Card, Moran & Newell, 1986; Garland, Stein & Muller, 1999).

Memory is supposed to play a central role in information processing. One of the original theories says that it is divided into 'short-term memory', 'working memory', and 'long-term memory' (cf. Atkinson & Shiffrin, 1968). Short-term memory relates to the perception of information. It is held in sensory stores from which it is lost unless attended. Attention, followed by rehearsing, is needed to store information in the relatively permanent long-term memory. According to Atkinson & Shiffrin, working memory is the central processing unit, which includes rehearsal. Working memory has limited capacity as opposed to long-term memory; for instance, the maximum size of memory span is assumed to be seven units (Anderson, 1995).

Cognitive psychologists generally assume that information processing may occur in different modes, dependent on task demands, task novelty, and the amount of learning and experience. Most cited are Schneider and Shiffrin (1977) and Shiffrin and Schneider (1977). They make a distinction between two modes in information processing: controlled and automated processing. They argue that automated processing requires little or no attention, occurs rather unconsciously and results in fast performance. The process is not restricted by short-term memory. In contrast, controlled processing takes time and effort, requires a lot of attention, does not occur unconsciously and results in slow performance. Automated processing takes place in well-

known situations whereas controlled processing is needed in novel, unfamiliar situations. Information processing develops into the direction of automation as a result of learning processes, since learning according to general learning theories such as Anderson (1982, 1983, 1995) diminishes the number of novel situations (this will be further explained in section 2.3).

In the theory of action regulation (Hacker, 1998), three levels of regulation are distinguished. These levels relate to the controlled and automatic mode of information processing; this is also called 'cognitive regulation' (Roe, 1999). The regulation levels differ in the amount of attention and degree of consciousness involved. They agree with the classification of Rasmussen (1986), originally developed for process industry but often applied in aviation psychology. Rasmussen argues that controlled information processing occurs at the 'knowledge-based' level, automatic processing at the 'skill-based level' and a combination of both at the 'rule-based level'. The German terminology (Hacker, 1998, p. 160) reflects directly which type of information processing is involved at each regulation level: "bewusstseinspflichtige Heuristiken, Strategien, Pläne" at the "intellektuelle Regulationsebene" or intellectual level (knowledge-based), "bewusstseinsfähige Handlungsschemata" at the "perzeptiv-begriffliche Regulationsebene" or perceptive-conceptual level (rule-based), "nichtbewusstseinsfähige Stereotypen/Routinen (Fertigkeiten) Bewegungsentwürfe" at the "automatisierte Regulationsebene" or sensimotor level (skill-based).

The idea is that novel situations require processing at the knowledge-based level while frequently practiced situations are mainly guided by skill-based and rule-based behavior. In operational performance a strict separation between the three levels does not exist but behavior continuously switches between the three levels (Rasmussen, 1986). An experienced job performer is handling routine tasks at the lowest levels, but unexpected events force him/her to problem solving activities at the highest (knowledge-based) level. In addition, controlled processes at the highest level control behavior at the lowest level. It is argued that it is necessary that a certain amount of skills is automated to keep enough space for problem solving activities in novel situations, which require effort, memory capacity and attention (Rasmussen, 1986).

Emotional-motivational regulation

It is assumed that several underlying regulation processes influence the performance process and its outcome (see for an overview: Roe, 1999). Factors such as stress, fatigue, motivation and self-efficacy are supposed to be implied in 'emotional-motivational regulation' (Hockey, 1997; Hockey, Coles & Gaillard, 1986). Models to describe this type of regulation are based on an integration of biological and information processing theories. Examples of environmental factors are colleagues, working environment, and task complexity.

Hockey (2003) uses the term 'Operator Functional State' (OFS) in the specific context of operators, that is, operators of industrial control processes such as chemical and nuclear plants, aviation, air traffic control, shipping and railways. OFS refers to the intrinsic relationship between human task performance and the background state of the individual. Hockey (2003, p. 3) defines OFS as follows: "the variable capacity

of the operator for effective task performance in response to task and environmental demands, and under the constraints imposed by cognitive and physiological processes that control and energize behavior".

A general idea in psychology is that emotions, stress or motivational influences can affect the regulation of information processing, usually into the direction of automatic processing (e.g., Hettema, 1979, 1989; Hockey, Coles & Gaillard, 1986). Human factors psychologists such as Reason, Isaac and Rasmussen are particularly interested in this phenomenon in their investigation of human errors in the aviation domain. They argue that a considerable amount of human errors occur because experienced job performers keep working on their routine in novel situations while controlled processing is needed (Rasmussen, 1986).

Underlying this applied psychological approach, more fundamental psychological theories try to explain why emotional-related factors affect information processing. The Yerkes-Dodson law is the first well-known attempt to explain the influence of stressors on performance (Yerkes & Dodson, 1908). The law assumes that there is an inverted-U relationship between arousal (motivation, tension) and the level of performance. Optimal performance occurs at a moderate level of arousal while this optimal arousal level is inversely related to task difficulty. However, this law has been criticized because it does not explain why there are these relationships; the law seems to be too generalistic and it is unclear which factors make a task difficult (Hockey, Coles and Gaillard, 1986; Sanders, 1986).

The cognitive energetic approach has sought an explanation in energy (called 'energetic regulation' by Roe, 1999). Several emotional factors such as motivation and stress are supposed to influence the person's 'psychophysiological state' (Hockey, Coles & Gaillard, 1986). They manifest themselves in physiological reactions (EEG, heartbeat, pupil and so on) that are related to the regulation level of information processing. Hockey, Coles and Gaillard (1986) generalize these processes as 'energetics', including concepts such as motivation, arousal, activation and stress. In contrast with the arousal concept of Yerkes and Dodson (1908), it is assumed that three mechanisms are involved instead of only one: arousal, effort and activation. An often-cited information-processing model that includes these three mechanisms is the model of Pribram and McGuiness (1992). In this model, arousal relates to activation of the perceptual system and has the function of familiarizing changing inputs. The activation system is concerned with the output, preparing motor performance. The effort system has a regulation function between input and output, which is related to attention. The three mechanisms are supposed to have different effects on information processing.

Some theorists say that the effort mechanism is responsible for self-regulation. It is supposedly activated when a person gets 'out of control', due to stressors (Mulder, 1986; Hettema, 2001), and tries to regain control or adapt to the new environment. Under such conditions, individuals will react intuitively, do not think clearly anymore, and may have a narrowing field of attention (Hettema, 2001). It is assumed that regulation at the higher (knowledge-based) level requires more energy than at the lower levels; effort is needed to switch from automatic to controlled processing (Mulder, 1986). In the energetic approach, it is assumed that additional effort is

needed to adjust the person's actual state to the state required to execute the task (Hockey, Coles & Gaillard, 1986). More effort enables controlled information processing, and more attention is paid to the task.

The concept of effort is further elaborated by Hockey (1997, 2003) in his state regulation model of compensatory control. In contrast with the energetic approach, the emphasis is on the motivational control of action. This assumes that "1) behavior is essentially goal-directed; 2) control of goal states is normally a self-regulatory process, and 3) regulatory activity attracts costs to other parts of the system" (Hockey, 1997, p. 75). Hockey focuses on performance under stressful conditions. Effective performance under stress remains remarkably stable. He considers this an active process under control of the individual. The idea is that "management of effort allows individuals to control the effectiveness of task behavior in relation to competing concurrent goals, changing demands, and current levels of energetic resources" (Hockey, 1997, p. 78). The person's actual psychophysiological state may require additional effort to be able to execute the same task under unfavorable circumstances. Besides extra effort, adjustment of performance targets (goals) is possible as a response to excessive task demands, for instance, by reducing required levels of accuracy or speed, by adopting less demanding strategies, or by paying less attention to subsidiary activities (Hockey, 1997, 2003).

The concepts of 'arousal' and 'activation' are often related to emotions that are non-intentional and spontaneously followed by 'emotional regulation' according to Roe (1999). From a point of view of personality, Hettema (2001) and Hettema, Leidelmeijer and Geenen (2000) performed experiments using physiological measurements of the three mechanisms arousal, effort and activation, which are supposed to occur each in the controlled and automatic mode. The studies showed that positive emotions tend to result in controlled processing while negative emotions are more likely to lead to processing at the lower (partly) automated levels. The idea is that negative emotions close the neurological 'gates' to controlled processing in conformance with the model of Pribram and McGuinness (1992). In this way, they explain why performance may be negatively affected by emotions.

Hettema, Leidelmeijer and Geenen (2000) try to explain how personality and information processing are connected. The assumption is that people react differently to situations because of person-related differences in preferences for controlled versus automatic arousal, effort, or activation. Kuhl (1994) also focuses on similar stable differences between individuals. He argues that individuals differ in state-orientation versus action-orientation, showing preference for one of both. In emotional conditions, which enhance automatic processing, some individuals are more likely to revert to state-orientation than to action-orientation. State-orientation is characterized by passivity, indecisiveness and hesitation whereas action-orientation is characterized by activity, decisiveness and initiative.

Strategies

The ability to use different strategies in varying situations is especially important for optimal performance in complex cognitive tasks (e.g., Anderson, 1995). There exist different concepts of strategy.

According to Hacker's action theory (Hacker, 1998) 'work strategies' refer to specific working methods, (sequence of) procedures or heuristics to perform a task. People usually dispose of several strategies, but they have their preferences that are dependent on specific task situations. This relates to the distinction between 'objective' and 'subjective' task as described earlier: the subjective task determines which strategy the individual applies. In this respect, the notion of 'efficiency' is relevant; an individual usually chooses a strategy that is most 'efficient'. Zijlstra (1993, p. 15) defines 'psychological efficiency': "Psychological efficiency relates to the individual's perception and/or experience of what is the most efficient way to carry out the task or, to be more specific, which strategy (or action alternative) is required to obtain the goal (i.e. 'benefits' of work behaviour) that involves the lowest 'costs' ". The concept of 'costs' relates to the amount of effort the individual has to invest while 'benefits' refer to the goal he intends to obtain; efficiency is considered a ratio between costs and benefits.

The cognitive energetic approach supports these ideas (Hockey, 1997; Cnossen, Meijman & Rothengatter, 2004). When operators are confronted with increased task demands, they may use less demanding strategies and this implies a reduction of effort. These adaptive strategies may result in the achievement of the primary task goal, while skipping a less essential secondary task may help to protect the main task goal. Cnossen et al. (2004) support this view for car drivers. Car drivers give more priority to tasks that serve goals of the driving task itself such as route finding, than tasks not directly relevant for driving such as radio tuning. Reduction of effort (caused by strategy adjustment) usually means a shift towards a less resource-intensive mode of control. This implies a shift from controlled processing ('knowledge-based') to a more automated mode ('rule-based') of information processing (Hockey, 1997).

Hettema (1979; 1989) has developed a slightly different view on the concept of 'strategy'. According to Hettema (1979), every person tries to get control over his environment by transforming a situation (S1) to another situation or goal (S2) by using different strategies (R). He defines strategy as follows: "a structure at the cognitive-symbolic level with the function to direct and guide behavior in concrete situations. It contains a collection of broad behavioral opportunities as well as behavioral preferences existing before the actual execution of behavior is started" (Hettema, 1979, p. 72). He assumes that these strategies do not work anymore when control over the environment is lost (Hettema, 1979, 1989). In that case, the 'tactical subsystem' takes it over, which results in intuitive, automatic reactions to situations. Thus, tactics have the function of maintaining and restoring control over the situation. We could say that Hettema's notion of tactics relates to the notion of adaptive strategies of Hockey (1997). According to this latter view, a distinct 'tactical' subsystem does not exist, but regulation processes at different levels determine person's actions continuously.

Cognitive psychologists suggest a perspective on strategies that relates to learning and problem solving. Anderson (1995) argues that a strategy is concerned with the organization of the solution of a problem. Experts are supposed to perform better than novices because they apply more effective problem solving strategies. It is assumed that different domains vary in optimal strategies, that individuals have

their own preference for strategies, and that strategies are dependent on acquired expertise.

In our study, we will make a distinction between two types of strategies. The first type of strategies refers to 'problem solving strategies' (Anderson, 1995): a specific method, heuristic or procedure to execute a task. The second type refers to adaptive strategies in the meaning of 'energetic demands' (Hockey, 1997, 2003). It is important to notice that personal efficiency (called 'psychological efficiency' by Zijlstra, 1993) can be in contrast with efficiency of the outcome. For instance, strategy A can be more effective and efficient than strategy B concerning the outcome of task execution (e.g., quality, quantity). Due to factors such as fatigue and stress the individual may choose strategy B (e.g., applying standard procedures) that is less effortful for himself (see section 3.1 for examples in ATC).

2.2 Competence

Since the 1990's the concept of competence has been become more and more popular in several disciplines. Several studies show that the concept of competence is applied in many contexts, such as education (e.g., Bereiter, 2002; Onstenk, 2000), training in organizations (e.g., Bolhuis & Simons, 1999) and human resource management (e.g., Hoekstra & Van Sluijs, 2000; Spencer & Spencer, 1993).

Definition

A universal definition of competence does not exist; the meaning depends on the context in which competences are used. Competences have been defined at individual and organizational levels (Van der Heijden, 1998; Hoekstra & Van Sluijs, 2000).

A relationship is supposed to exist between individual and organizational competences (Hoekstra & Van Sluijs, 2000; Roe, 2002): learning at the individual level enhances the competences of the organization as a whole, and analyzing the competences at the higher organizational level helps with formulating the competences of individual employees and trainees. For instance, in ATC organizations the issues 'safety' and 'efficiency' are central issues in both individual and organizational competences.⁴ We focus on individual competences for the purpose of this study. We use our own definition: "Competence refers to the ability to effectively apply acquired knowledge, skills and attitudes while performing tasks in realistic settings". This definition is based on various definitions and theories and will be further explained and elaborated by reviewing the literature.

⁴ At LVNL, the concept of "VEM" (Safety, Efficiency, Environment) is used for development of operational procedures at the highest organizational level, and these issues are also applied in individual competence profiles.

Components of competences

In educational contexts and in training situations, it is commonly assumed that competences consist of skills, knowledge and attitudes that are applied together in a realistic learning environment (Onstenk, 2000). A learning environment that corresponds strongly with reality is generally called an 'authentic learning environment' (Onstenk, 2000; Petraglia, 1998). People are supposed to be competent to perform a task when they are able to apply the acquired skills, knowledge and attitudes all integrated in the working environment.

A general idea is that different kinds of knowledge and skills require different instruction methods (Gagné, 1985; Romiszowski, 1988). Most educational theories restrict their classifications to the cognitive part; the non-cognitive part is often ignored. Bereiter (2002) maintains that competences cover everything. To illustrate this idea, he distinguishes six kinds of knowledge:

- *Statable knowledge*: Knowledge that is usually called 'declarative knowledge' (Anderson, 1983): knowledge that you can state or declare, knowledge about facts and things.
- *Implicit understanding*: Knowledge that people have acquired by expertise; knowledge people have but cannot state it. It is knowledge about something and stays always implicit.
- *Episodic knowledge*: Remembered events, cases or episodes. It relates to theories about 'case-based reasoning' and experiential learning.
- *Impressionistic knowledge*: Feelings and impressions that influence our actions.
- *Skills*: Cognitive and motoric skills. Cognitive skills are rules, steps and procedures about how to perform something, comparable with 'procedural knowledge' (Anderson, 1983).
- *Regulative knowledge*: Metacognitive knowledge: knowledge and control over our thinking and learning activities.

Bereiter (2002) argues that trainees have acquired competences when these components are integrated in so far that it is hardly possible to distinguish between them, or even stronger, that the whole of the person's competence appears to be greater than the sum of identifiable parts.

Learnability

In this study, we consider competences as the result of learning. To dispose of certain personal characteristics is a precondition to acquire specific competences; competences are inherently related to the task to be performed (Roe, 2002, 2005). The model in Figure 2.1 illustrates this view:

Figure 2.1. Model of competences (Roe, 2002, 2005)

People's abilities, personality and other special characteristics are relevant in employee selection, because they are required to learn the job. Next, skills, knowledge and attitudes are acquired in training. This results in the acquisition of the competences. We assume that competences are learnable to a certain extent, in accordance with Thijssen (1998), Hoekstra & Van Sluijs (2000) and Kessels (1999). These authors argue that competences are always learned and not innate, but some people are not able to learn them in a restricted period because of deficiencies in personal characteristics.

We should notice that this view deviates from the approach of Spencer and Spencer (1993) who do not consider competences as learnable but as personal characteristics. Their definition is often applied: "an underlying characteristic of an individual that is causally related to criterion-referenced effective and/or superior performance in a job or situation" (p. 9). Different perspectives become visible in the set-up of competence profiles and formulation of competences. Supporters of the last view define general competences that can be applied in different kinds of jobs (e.g., Spencer & Spencer, 1993), while others define task or job-related competences (e.g., Onstenk, 2000; Roe, 2002)

Implicit and explicit knowledge

It is widely accepted that part of competence belongs to the implicit knowledge of job performers. Implicit knowledge, also called *personal knowledge* or *tacit knowledge*, is knowledge that stays partly invisible during job performance; people are not aware that they have it (Polyani, 1967). Many (cognitive) processes occur unconsciously and require no attention anymore as a result of learning and job experience. This

involves the lower regulation levels of information processing according to Hacker (1986) and Rasmussen (1986), discussed in section 2.1.

Nowadays, the concept of implicit knowledge and its relationship with informal and formal learning is often used in the context of competences and knowledge management (see for an overview Oprins & Dankbaar, 2002). Knowledge management is assumed to make implicit knowledge explicit (Nonaka & Takeuchi, 1995). The same process plays a role in training situations. Coaches are continuously explicating their implicit knowledge, including skills and strategies, to their trainees by demonstration and explanation. For trainees the reverse process takes place: explicit knowledge becomes implicit by learning and practicing. Bereiter (2002) argues that implicit or personal knowledge contains all six kinds of knowledge as presented above. He assumes that almost all components can be made explicit in some way, except 'implicit understanding', which he explains as follows: "How do we change the lens through which people view some aspects of the world?". In contrast, declarative knowledge is considered always explicit (Anderson, 1995; Bereiter, 2002).

Competences and performance

Competences strongly relate to performance, but should also be differentiated from it. Task performance can be observed but competences cannot. They should be inferred from performance by taking into account other performance-shaping factors (Roe, 1999). A person can be competent but performance can still be insufficient, due to a temporary or long-term psycho-physiological state (see section 2.1). Figure 2.2 illustrates how competences are a precondition for performance but not vice versa:

Figure 2.2. Competences and performance (after Roe, 2002)

2.3 Competence acquisition

Views on learning

Since the beginning of the 20th century, several views on learning have been developed. The view of *behaviorism* emphasizes the influence of external 'stimuli' on learning processes, requiring specific 'responses' of the learner (Mayer, 1987). Outcome performance ('response') is supposed to be directly caused by instructional manipulations ('stimuli'). This theory is based on the 'law of effect' developed by Thorndike (1911). This law implies that if a behavior is followed by a positive result,

then there will be great chance that this behavior will occur again in the future under the same circumstances, and if behavior is followed by a negative result, it will occur less likely in the future. Skinner (1938) elaborated this idea with his reinforcement theory. This theory assumes that behavior is reinforced by the consequences of that behavior. Reward and punishment, as feedback on the learner's actions, are considered to be 'reinforcers'. Many educational and training systems are still based on these behaviorist ideas.

The cognitive approach to learning, *cognitivism*, considers the relation between instructional manipulation and performance to be indirect, because instructional manipulation (e.g., learning tasks, coach, environment) affects cognitive processes and learning. Cognitive psychologists emphasize internal processes within the learner himself, and focus on information processing. In contrast with behaviorism the learner is supposed to have an active role in his own learning process (Mayer, 1987). The main idea is that learning results in changes in knowledge structures in the brain of human beings (Anderson, 1995). Consequently, external influences on learning have moved to the background.

Learning theory has developed in the direction of *constructivism*. A revival of the influence of the external environment happens, although in a quite different way. Constructivism can be considered as a further development of cognitive psychological ideas about learning (see e.g., Duffy & Jonassen, 1992; Petraglia, 1998). However, constructivists argue that cognitive psychologists focus too much on internal processes within the learner himself without considering the context. Constructivists assume that people actively construct their knowledge in their own way, which differs from person to person, because everyone has his own prior knowledge, experiences, motivation, learning styles and learning purposes.

Proponents of 'situated cognition' assume that construction of knowledge takes place in a relevant context such as the social or working environment (Brown, Collins & Duguid, 1989; Lave & Wenger, 1991; Petraglia, 1998). The influence of Vygotsky's ideas is visible here: learning is determined by (social) interactions with the environment and experiences in the real world. These are the most fundamental principles of constructivism, although extreme and moderate constructivists have some different ideas (see, for instance, the discussion with Merrill in Duffy & Jonassen, 1992). Most important differences exist in the amount of guidance by instructors and prescription of instructional events, both totally rejected by the extremists (e.g., Petraglia, 1998).

The competence-based view appears to correspond best with the constructivist approach. The focus is on 'authentic learning environments'. Competences are considered as context-specific and being acquired in a realistic environment to stimulate transfer processes (Petraglia, 1998; Onstenk, 2000). Following our definition of competence (see section 2.2), we could say that knowledge, skills and attitudes can be acquired independently of the context in which they are applied, but competences cannot (Bartram & Roe, forthcoming). Learning builds on competences that have been acquired earlier, and also involves a reorganization of a person's repertoire. People will learn from their experiences if there are enough opportunities to learn. Competences are usually acquired in a process of learning-by-doing, often in an in-

formal way as is common in on-the-job training (Onstenk, 2000; Bartram & Roe, forthcoming). Individual differences in learning (e.g., learning styles) are supposed to have much influence on the learning process (e.g., Jonassen & Grabowski, 1993). This is also caused by many dispositional factors such as personality features and cognitive abilities (e.g., Bartram & Roe, forthcoming; see also Figure 2.1). Moreover, the competence-based view implies that competences consist of many aspects, including non-cognitive. With growing interest in competences and constructivist ideas, psychologists pay attention to other kinds of learning such as experiential or social learning (Boekaerts & Simons, 1993).

Complex skill acquisition

This study focuses on complex cognitive skills (e.g., Schneider, 1990; Van Merriënboer, 1997; Regian & Schneider, 1990; Taatgen & Lee, 2003). According to some psychologists, complex skills are considered rather similar to competences. This section summarizes cognitive-psychological learning theories that describe complex skill acquisition. However, we should realize that these theories are limited insofar that they are mainly restricted to cognitive skills, while competences cover other components as well.

Skill acquisition

The acquisition of complex skills is more typical for learning in practice than the acquisition of simple skills, but a uniform theory does not exist yet. Simple skill acquisition forms the basis for a theory of complex skill acquisition (Lee & Anderson, 2001; Taatgen & Lee, 2003).

A general assumption of skill acquisition is that performance increases by practicing. The relationship between practice and performance is originally presented as the 'power law of practice', functioning as a general learning curve model (Newell and Rosenbloom, 1981). This law shows the relationship between the time or speed to perform a task and the number of practice trials in a quantitative equation. It predicts that performance improves rapidly in initial trials, but that the rate of climb diminishes with more and more trials. The specific form of the learning curve is supposed to be dependent on factors such as task characteristics, individual differences and instructional methods. Individual differences lead to different initial performance levels, acquisition rates and final learning asymptotes (e.g., Schneider, 1990; Ackerman, 1989, 1990).

Schneider (1990) performed empirical studies that show heterogeneity of component improvement rates. He distinguishes consistent task components from varied task components, based on the earlier theory of Schneider & Shiffrin (1977) and Shiffrin & Schneider (1977) concerning consistency of tasks: "Consistent task components show large improvements with practice whereas varied components do not. A consistent component is defined as one in which the subject can make the same response to a particular stimulus situation every time the stimulus situation occurs (. . .). In contrast, a varied component is one in which the mapping between the stimulus and response varies across trials." (Schneider, 1990, p. 304). This distinction can be related to unstable (trainable) and stable (untrainable) skills (Regian & Schneider,

1990), and is also comparable with the distinction in respectively recurrent and non-recurrent skills (Van Merriënboer, 1997).

Cognitive psychologists have tried to find an explanation for the 'power law of practice' focusing on information processing. Their theories rely on the two modes of information processing: controlled and automated processing (Shiffrin & Schneider, 1977; Schneider & Shiffrin, 1977). The basis of these theories is that the learning process starts with controlled information processing; controlled processing changes into the direction of automated processing by practicing. Fitts (1965), Fitts and Posner (1967) and Anderson (1982, 1983, 1995) argue that the learning process can be divided into three phases, based on the distinction between these two modes of information processing. Anderson (1983) describes in more detail which mechanisms are responsible for complex skill acquisition. His theory is used in the majority of instructional design theories (Romiszowski, 1988; Merrill, 1994).

In the first phase, called the 'cognitive phase' by Fitts (1967), prerequisite declarative knowledge is learned (facts, theories, concepts) by forming complex semantic networks, which consist of sets of propositions and schemata. Information processing occurs controlled and consciously, performance is slow because of the limited capacity of working memory. The second stage follows and is called the 'associative phase' by Fitts and Posner (1967). Declarative knowledge changes into procedural knowledge; 'production rules' are learned that are more domain-specific. Anderson (1982) calls this process the 'knowledge compilation process' or 'proceduralization'. The speed of performance increases because of the development towards more automatic information processing. Declarative knowledge does not need to be retrieved from and held active in working memory anymore, because this knowledge is built in the domain-specific production rules. At the end of the third phase, called 'autonomous phase' (Fitts, 1967), the process of proceduralization (Anderson, 1982) has finished. The trainee has reached a skill level that results in very fast performance, which occurs unconsciously without any effort due to automated information processing. In this phase, more experience is built up by practicing the skills to give the production more strength.

The development in three learning stages relates to three levels of behavior that are distinguished by Rasmussen (1986) and Hacker (1998) (see also section 2.1). The 'knowledge-based' level corresponds with the 'cognitive stage'; behavior is conducted by controlled information processing. At the 'rule-based' level, which corresponds with the 'associative stage', a combination of controlled and automated information processing occurs. Behavior at the 'skill-based' level is conducted by automatic processing, which corresponds with the 'autonomous stage' in learning processes.

However, complex skill acquisition appears to be more complicated than described above (Van Merriënboer, 1997; Lee & Anderson, 2001; Taatgen & Lee, 2003). In many domains, automation of skills is not sufficient because a competent job performer has also to be able to deal with novel situations (cf. 'non-consistent tasks'; Schneider, 1990). This implies that a transition in three stages as described above does not need to exist, and that practice does not always improve performance (Schneider, 1990; Ackerman, 1989). This relates to problem solving for which specific theories have been developed.

Acquisition of problem solving

Newell and Simon (1972) are often cited in the context of problem solving. They consider problem solving as a search through a problem space that consists of several states or representations of the problem. The problem solver continuously tries to change one state into another to achieve the goal state finally. The solution of a problem is a sequence of operators. These operators refer to actions that transform one problem state into another. Anderson (1995) distinguishes two kinds of problem solving methods.

The first group of problem solving methods is required in really novel and unfamiliar situations. These methods are supposed to be applied at the knowledge-based level in terms of Rasmussen (1986). They are called *weak methods*, because they are domain-independent (Anderson, 1983). They depend on innate or personal capacities, such as intelligence (Anderson, 1995; Ackerman, 1988). According to Anderson (1982, 1995), all learning processes start with this kind of problem solving, because all situations are novel in the beginning. Some examples of methods for learning how to solve problems in unfamiliar situations are the difference-reduction method or hill-climbing, analogy and means-ends-analysis (Newell & Simon, 1972; Anderson, 1995).

The second group consists of domain-specific problem solving methods. The problems are not really new and unknown; thus, problem solving is mainly applied at the rule-based level (Rasmussen, 1986). Learning how to organize the solution for a problem ('strategies') is also called *strategic learning*. The problem solving strategies of experts are called 'stronger' than the aforementioned 'weak' methods (Anderson, 1995). Studies showed that these strategies are acquired differently in different domains; that's why researchers emphasize the role of domain knowledge in acquiring expertise. For instance, in the domain of physics and geography, a solution of a novice is characterized by reasoning backwards in contrary to the expert who applies forward reasoning (Larkin, McDermott, Simon & Simon, 1980). Chi, Feltovich & Glaser (1981) demonstrated that experts in physics rely on deeper principles referring more directly to the solution, while novices are more focused on surface features. Furthermore, experts seem to have an extended memory because they are able to recognize complex patterns in problem representations. Experts are assumed to have bigger and more 'chunks' (meaningful pieces of information) available than novices. This was first discovered in the chess domain (De Groot, 1965).

However, it has been observed that strategy learning is lacking in many training situations (Schaafstal, 1991; Schneider, 1990). This may be caused by the nature of strategies, which are situation-dependent and individual-dependent (Anderson, 1995; Van Bakel, 1995; Hettema, 1979, 1989) as argued in section 2.1. It is rather difficult to find out which general strategies experts use, because they belong to their implicit knowledge. Yet, explicitation of these strategies is needed to define specific (domain-dependent) procedures or heuristic rules to be acquired. It appears to be not quite clear how experts have acquired problem solving strategies and how they use them in novel or less novel situations.

This problem is further illustrated by Spiro, Feltovich, Jacobson & Coulson (1992). They make the distinction between well-structured and ill-structured problems. In well-structured problems, the given state is clearly defined, the goal state is

clearly recognized and there are clear rules for getting from one state to another (Newell & Simon, 1972). A well-structured problem has an unambiguous solution and can be presented in a systematic manner like arriving at checkmate in chess or a mathematically correct answer in math. An ill-structured problem does not have one single solution and is sensitive to the context. Many real-life problems can be defined as ill-structured, for instance, designing. Each problem is different and novel; thus, solving these problems cannot be learned as learning chess or physics.

Individual differences in learning

In cognitive psychological research, the emphasis on individual characteristics in learning has increased (see for an overview Jonassen & Grabowski, 1993). In training situations, individual differences are specifically manifested in learning styles. There are many classifications. The learning styles of Kolb (1984) are generally applied in vocational training or training in practical learning environments. The four styles are: active experimentation, reflective observation, abstract conceptualization, and concrete experience. According to Kolb, individuals differ in preference to start with one of these four styles.

Ackerman (1989, 1990) is often cited with regard to the relation between (complex) skill acquisition and individual differences in cognitive abilities, using the three-stage learning model of skill acquisition (see previous section). According to his theory, general intelligence must be a determinant of individual differences in phase one of learning processes described by Fitts, Anderson, Schneider and Shiffrin. Differences in perceptual speed ability are most relevant in phase two; psychomotor ability is especially demanded in phase three. Several empirical studies have been performed to prove this theory (Ackerman, 1989). They showed that as people gain experience, their performance is less predictable from measures of cognitive abilities than earlier in the training. In this way, the theory of Ackerman supports the observation of Schneider (1990) that initial performance does not necessarily predicts final performance. Several abilities are required in different learning phases.

Ackerman's theory also corresponds with the general idea that intelligence is a precondition to learn and with the idea that 'weak methods' in problem solving relate to intelligence. Ackerman (1989) argues that his theory is only true for highly 'consistent' tasks. For problem solving tasks that are really novel ('non-consistent'), there is no transition to stage three and perceptual and psychomotoric abilities are subsequent to general cognitive abilities. This idea is also supported by the observation that in 'consistent' tasks interindividual variability decreases with task practice, while in 'non-consistent tasks' the magnitude of individual differences stays constant or increases. Higher task complexity usually requires more demands on individual cognitive abilities (Ackerman, 1989).

Ackerman (1989) states that research on differences in learning curves in relation to individual differences and task characteristics may help with finding possible solutions to the problems of predicting performance at various stages of task proficiencies. It may provide diagnostic information about failures and what aspects of the training program prevent or facilitate phase transitions of skill acquisition. Studies of aptitude-treatment interaction (ATI) try to find out how instruction can be adapted to

individual characteristics (Snow, 1989). Aptitudes are defined as: "(. . .) properties of the initial states of persons that account for the end states they reach through attempting to learn in particular ongoing situations; treatment refers to any one such situation" (p. 14 Snow, 1989). However, others think that adaptation of training to individual needs is hardly possible, because there are too many individual differences in learning (Gagné, 1985; Petraglia, 1998; Bereiter, 2002).

2.4 Conclusions

We reviewed the literature about performance, competence and competence acquisition with the design of a competence-based assessment system in mind (see chapter 7).

In our view, competences are the result of learning and inherently related to the task: "the ability to effectively apply acquire knowledge, skills and attitudes while performing tasks in realistic settings". Competences are usually acquired in real work settings or in simulators; knowledge, skills and attitudes are merged in a process of learning-by-doing. Competences are assumed to cover everything including non-cognitive aspects. It should be noticed that competences are a precondition for performance, but various regulation processes have influence, for instance, stress may cause temporarily low performance while a person is competent. Three levels of regulation of information processing are distinguished: knowledge-based (controlled processing), rule-based (automatic and controlled processing), and skill-based (automatic processing). The highest (knowledge-based) level corresponds with novel task situations while the lowest (skill-based) level is needed in familiar or routine tasks. The middle (rule-based) level takes an intermediate position. We consider the application of strategies important in complex skills such as air traffic control. It relates to the difference between 'objective' and 'subjective' tasks. The subjective task determines which strategy is applied. When operators are confronted with increased task demands, they adapt their strategies into the direction of less effortful. It is argued that strategy learning should get more attention in training.

Learning implies a transition from controlled towards automatic information processing, but only for 'consistent' task components that improve by practicing in contrast with 'non-consistent' task components that are supposed to be less trainable. However, theories about complex skill acquisition, usually founded in simple skill acquisition, have to be further developed. Many individual differences in learning exist. This makes it difficult to predict how people learn and to design a training that optimally supports learning for each individual.

* Competence acquisition in ATC

In this chapter, we review the literature concerning competence acquisition in ATC. We examine the available literature concerning the nature of the air traffic control task (section 3.1), we pay attention to the acquisition of ATC competences (section 3.1), and we discuss relevant literature related to the desired properties of training for ATC (section 3.3).

3.1 The air traffic control task

The nature of the task

Air traffic control (ATC) is considered a highly complex and dynamic process control task (e.g., Wickens, 1992; Garland, Stein & Muller, 1999; Taatgen & Lee, 2003). Air traffic controllers are supposed to be part of the group of human operators of complex human-machine systems, working in safety-critical environments such as transport (aviation, railways, shipping), and chemical or nuclear plants (Niessen, 1999; Hockey, 2003). The processes typically consist of a large number of interrelated variables that must be controlled continuously (Wickens, 1992).

The primary aim of air traffic control is to expedite and maintain a safe and orderly flow of air traffic (ICAO, 2004) by issuing information and instructions (headings, altitudes, speeds) to pilots. The continuous flow of moving aircraft cannot be stopped; timely actions are needed to create safe and efficient traffic flows before possible collisions become critical. Decision making has to occur quickly due to the dynamic nature of the control process, and temporal co-ordination of actions is required to control the dynamically changing air traffic situation.

The ATC environment can be characterized by “a continuous sequence of ever-changing, transient information (e.g., series of aircraft being handled by an air traffic controller), which must be encoded, retained primarily for tactical use (3-5 min) and secondarily for strategic planning, and subsequently discarded” (Garland, Stein & Muller, 1999, p. 457). Two forms of planning are distinguished: *tactical* planning when controllers make decisions that are related to the current moment (e.g., separation between two aircraft) and *strategic* planning about longer periods for more aircraft (e.g., sequencing aircraft for approach to a common destination (Moertl, Canning, Gronlund, Dougherty, Johansson, & Mills, 2002)).

Air traffic controllers have to operate under strict requirements of safety. Responsibility of air traffic controllers is high since accidents may cost human's life and create considerable damage. Human factors research in air traffic control indicates that the high demands on cognitive load of controllers can result in human errors, because human performance has its limitations and constraints (Isaac & Ruitenberg, 1999; EATCHIP, 2002; Garland, Stein & Muller, 1999). Task complexity and, consequently, the chance of human errors have increased by the growing volume of air traffic. Traffic increase is also dependent on organizational factors such as the number of personnel, specialized air traffic controllers, division in sectors, or number of towers. Advanced technology is developed to help controllers to perform their jobs despite of the increasing traffic load, but high demands on human performance are not likely to diminish in the near future (Shebilske, Goettle & Garland, 2000).

Air traffic control is usually supposed to imply a high workload, although this workload differs over time. Controllers have to switch between low and high workload (cf. vigilance) dependent on the traffic load. Traffic load refers to the number of aircraft to be handled, traffic complexity and time pressure, but it is assumed to have a strong subjective component as well (Collet, Averty, Delhomme, Dittmar & Vernet-Maury, 2003; Averty, Collet, Dittmar, Athènes & Vernet-Maury, 2004). This relates to the distinction between 'objective task' and 'subjective task' according to Hacker (1998; see section 2.1). Mental workload can be conceptualized as the interaction between the structure of systems and tasks (type of traffic, instruments) and capabilities, motivation and psycho-physiological state of the controller.

Information processing

From the perspective of learning, air traffic control is called a complex cognitive skill (Schneider, 1990; Van Merriënboer, 1997; Seamster, Redding & Kaempff, 1997; Garland, Stein & Muller, 1999). This implies that cognitive processes are mainly responsible for performance and that the complex skill comprises several sub skills. Air traffic control is also defined as a high-performance skill (Schneider, 1990): more than 100 hours of training are required, substantial numbers of individuals fail to develop proficiency (25 – 70%), and the performance of an expert is quantitatively and qualitatively different from that of a novice.

A commonly made assumption is that operators in dynamic and complex tasks such as ATC create a mental representation of the changing environment. This makes it possible to keep an image of the relevant but transient information in working memory (Garland, Stein & Muller, 1999). Pattern recognition plays a central role; aircraft are grouped in a certain way to be able to memorize their positions. Controllers "see" the patterns that help them to create order in a seemingly chaos by streaming traffic flows. Davison and Hanson (2003) argue that controllers make *abstractions* of the traffic situation to be able to manage the large amounts of information within the constraints of cognitive abilities. They distinguish three key abstractions: standard flows, groupings and critical points. Standard flows are aircraft as members of the normal operating pattern within a sector. Groupings are sets of aircraft within a standard flow linked by common properties such as standard flight lev-

els. Critical points in the airspace are intersection points of several standard aircraft flows. Figure 3.1 illustrates this view.

Forming a 'mental picture' in ATC is usually referred to as *Situational Assessment*. This is defined in the following way (Endsley, 1995): "The perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future". *Situational Awareness* (SA) is considered the product of the process *Situation Assessment* that takes place at three levels: perception (SA1), interpretation (SA2) and anticipation (SA3). Attention management strategies are crucial to keep this continuously changing 'picture' up-to-date (Shebilske, Goettle & Garland, 2000).

Figure 3.1. Abstractions of a traffic situation (Davison and Hanson, 2003)

With regard to the regulation of information processing, air traffic control is supposed to occur at the three levels of control as explained in section 2.1 (Isaac & Ruitenbergh, 1999). First, there is evidence showing that experts in air traffic control apply automated skills. According to Isaac & Ruitenbergh (1999), skills such as equipment operation, strip- and label management, and application of radiotelephony are performed at the skill-based level. Problem solving activities that are needed in unfamiliar or novel situations (e.g., emergencies) are supposed to occur at the knowledge-based level, and the application of standard operating procedures (e.g., maintaining separation minima) at the intermediate rule-based level.

Emotional-motivational factors can influence the cognitive regulation of information processing (see section 2.1). Human errors may occur at all levels of control. Isaac & Ruitenbergh (1999) provide some examples. Errors at the skill-based level are associated with attentional or memory failure (slips and lapses); for example, omitting information from the strip or keep using the old runway number in radio transmissions after a runway change. Rule-based mistakes concern misapplication of good rules or application of bad rules. An example of the first one is forgetting that the airfield surface is icy in winter periods, and of the second one that a trainee insists on giving a flight level instruction before a heading direction, irrespective of the

situation. Mistakes at the knowledge-based level occur if controllers cannot produce a timely solution in an unfamiliar situation (e.g., an emergency).

Isaac & Ruitenbergh (1999) argue that stress changes the type of attention given to different sources of information qualitatively. Controllers who are under stress tend to concentrate on certain strips in the bay, some areas on the radar screen or areas of the air space from the tower. They concentrate on more frequently occurring signals and neglect sources of less probable information. This is also called a 'tunnel vision' (EATCHIP, 1999). Controllers become pressed by time and do not anticipate on situations anymore. They may lose the 'mental picture'.

Strategies

We assume that controllers are keeping safety (conflict detection), efficiency (traffic delay) and their own mental workload (cf. 'personal efficiency'; Zijlstra, 1993) in optimal balance by adapting their strategies continuously (Oprins & Burggraaff, 2006). Situational awareness is needed to get insight into the most safe and efficient solution. Controllers keep their own mental workload under control by adjusting their strategies towards less effortful if needed. Then they tend to revert to routine actions and standard procedures that require less attention and gain time. A possibility is to divide a problem into small pieces with several, more easy solutions. Examples are: building in safety buffers (choosing intermediate flight levels instead of a continuous descent) or choosing standard routes that do not require attention for some time and are less time-consuming. Other time-saving solutions are: early transfer of traffic to adjacent centers, combining instructions and calls to avoid redundant radiotelephony, and splitting the traffic to a colleague when working on a combined sector (Oprins & Burggraaff, 2006).

Several studies support these ideas. The study of Sperandio (1978) is often cited as representative example for the theory of adaptive strategies (Hockey, 1997, 2003). Sperandio found that controllers adjust their strategy when the number of aircraft increases beyond their comfortable workload level. Controllers adopt a more routinized work pattern by assigning standard routes to the aircraft instead of establishing the more efficient routes for each individual flight. The primary task goal of safety is achieved despite of the higher number of aircraft under control, while secondary task goals such as efficiency are compromised. Sauer, Hockey and Wastell (2000) found comparable effects on the operator's strategy in terms of compensatory control (Hockey, 1997), although this was not the primary goal of their study: operators adjusted their strategies by skipping less essential secondary tasks.

In addition, it is argued that the 'mental picture' (Situational Awareness) is strongly dependent on the controller's availability of strategies. According to Nunes and Mogford (1993), the selection of a particular strategy, stored in the mental model, depends on training, experience, preference, and the specific traffic situation. They mention four strategies for conflict detection: trajectory prediction, time comparison, distance comparison, and altitude comparison. Altitude comparison appears to involve lower cognitive costs (cf. effort) than time comparison that requires calculation of speed.

Studies of mental workload of air traffic controllers relate to these adaptive strategies. Averty et al. (2004) connect mental workload directly with the controllers' psychophysiological functional state in conformance with Hockey (2003). They tried to find a workload index based on both objective aspects (e.g., number of aircraft) and subjective aspects (e.g., interventions to be carried out by the controller). They argue that air traffic controllers can modulate their own workload by using strategies that adjust their workload to an optimum level. These strategies refer to the interval between conflict diagnosis and the time when a decision is made, called 'maturing time' (MT). Early decisions could be redundant or efficient, but extending maturing time increases attentional demands and workload. In the study of Collet et al. (2003) the workload of controllers was measured in real-life sessions physiologically. They distinguish objective estimation of workload (counting aircraft to be monitored) from subjective self-estimation comparable with Averty et al. (2004), and they relate these estimates to the physiological measures. They also conclude that these subjective aspects should be taken into account. Costa (1993) evaluated the workload of controllers based on physiological measures taken during three different work shifts in order to find out whether controllers' workload was acceptable during any hour of the day. They found that this overall workload was acceptable in comparison with other jobs, but that this was influenced by the flexible arrangements of work teams in relation to variations of air traffic.

Evidence for the application of problem solving strategies (see section 2.1) is provided by studies on differences between experts and novices. It is assumed that the quality of the controller's mental picture directly relates to expertise in ATC (Garland, Stein & Muller, 1999; Shebilske, Goettle & Garland, 2000). Garland, Stein and Muller (1999) mention that experts see patterns in combinations of aircraft. This refers to the principle of chunking: experienced controllers organize a potential conflict between three aircraft as a chunk (see also Figure 3.1). Another example concerns visual search strategies: experienced controllers search for information visually, while inexperienced controllers pay attention to all potentially relevant information. Lamoureux & Cox (1999) found comparable differences between instructors and students, although this was not the primary aim of their study. Instructors consider only essential information in contrast with students, and instructors are able to generate more potential solutions and alternatives than students because they understand the implications of their actions better. Niessen (1999) emphasizes the dynamic character of the air traffic control task. He states that the continuously changing situation can demand a new prioritization of simultaneously occurring events. Anticipation on future states of aircraft is supposed to be crucial. His study showed that experienced controllers have a larger time scale for anticipation than less experienced controllers.

Many psychologists emphasize the importance of strategies in air traffic control training (Schneider, 1990; Niessen, 1999; Nunes & Mogford, 2003; Garland, Stein & Muller, 1999; Oprins & Burggraaff, 2006). Nunes & Mogford (2003) argue that trainees should be made aware of these strategies, and they should be able to assess which ones work best for themselves. Incorporation of standardized strategies in training may even reduce the on-the-job training time according to them. Schneider (1990)

provides some guidelines how to train strategies that minimize workload and to train time-sharing skills for dealing with high-workload environments. We also believe that strategy learning can be a key to success in training.

3.2 ATC competence acquisition

The learning process

It is interesting to note that air traffic control is often used as an illustrative example of complex skill acquisition. Many experiments have been done with the so-called Kanfer-Ackerman task (Ackerman, 1988). This simplified, simulated ATC task can be performed by novices. Candidates have to make several aircraft landing on different runways, making use of several hold-level positions, and taking into account runway conditions and weather circumstances. Actions such as moving planes between hold levels are executed by clicking on the keyboard. Feedback is provided on errors and the rules of the task are displayed on the screen, such as wind restrictions for landing on specific runways and the availability of only one hold-level for each plane. Time plays a critical role in the task execution.

Lee & Anderson (2001) used the Kanfer-Ackerman task to examine whether theories of simple skill acquisition can be used to explain the acquisition of complex skills. Their assumption is that complex skills can be split up into several smaller parts, which are learned in conformance with basic learning theory. They made learning curves of the decombined ATC tasks. The results confirmed their hypothesis, called the 'reducibility hypothesis'. They found that learning at the different levels of decomposition is well fit by the power function (cf. Newell & Rosenbloom, 1981; see section 1.3).

Taatgen and Lee (2003) applied the Kanfer-Ackerman task to provide evidence for the transition from automated to controlled processing in three stages. For instance, declarative knowledge acquired in the first stage involves the instructions how to land a plane and the locations of the hold levels. Production compilation combines this task-specific declarative knowledge with general procedures into task-specific production rules. Experience improves performance further, for instance, which planes may land on which runways. Their model gives a prediction of the entire learning process, from the initial instruction to the final asymptotic performance. However, a restriction of this model mentioned by the authors themselves is that it does not account for differences and developments in strategy use.

John & Lallement (2003) investigated the use of strategies and how people switch between them by using the same Kanfer-Ackerman task. They identified two dimensions of strategies. The first dimension was the level to which aircraft are brought; they found seven possibilities. The second dimension involved three kinds of patterns: *stacked* (a series of aircraft one right after each other), *sequential* (attending to one aircraft at a time by making land one and then another), *opportunistic* (manipulating several aircraft at the same time). In their view, people shift between strategies, either abrupt or gradual, and this shifting can enhance performance.

Another study with the Kanfer-Ackerman task, related to strategies, is by Schunn & Reder (1998). They investigated individual differences in adaptation of strategies. They assume that while individuals may have the same set of strategies, they may differ in their ability to select the best strategy for a given situation. They found evidence for this assumption and suggest that abilities such as inductive reasoning and higher working memory capacity influences flexibility in strategy use.

The Kanfer-Ackerman task was also used to examine individual differences in skill acquisition. In the studies of Ackerman (1988, 1990), learning behavior on this ATC task was correlated with performance on simpler tasks. The differences between the three stages with respect to general abilities, perceptual abilities and psychomotoric abilities were also found for this ATC task. Taatgen (2001) tested this theory by manipulating the relevant factors general intelligence, speed of proceduralization, and psychomotoric speed. Both studies showed the same patterns and provided evidence for Ackerman's theory in the domain of ATC.

The results of these studies certainly shed light on the learning processes involved. However, we should realize that the Kanfer-Ackerman task may be quite different from 'real' ATC tasks that are much more complex. This problem will be further addressed in chapter 10.

Another view on learning processes emerges from research by Shebilske, Goettle & Garland (2001). They explain how Situational Awareness is acquired and emphasize the acquisition of attention management strategies. They distinguish four changes in Situational Awareness of operators of complex dynamic systems (e.g., pilots, air traffic controllers, drivers) that explain the learning process in three stages. First, the operator gets a richer database of possible events and corrective actions; execution of actions can reach automatic control (skill-based level). Second, the operator can estimate the probability that certain events will occur and can predict which events are likely to follow the current event; these events may be 'unexpected'. When the unexpected events are familiar enough to be associated with general rules, they invoke 'rule-based' behavior. When they are unfamiliar, automatic processes are interrupted and task operation must be taken over by controlled processing ('knowledge-based'). Attention management strategies are needed for switching to this higher level. Third, the set of truly novel events becomes smaller, but the remaining ones become more complex. Fourth, the mental model is enriched with a higher understanding of the system, task goals, the operator's own capabilities and environmental factors. This results in better diagnostic activities that are performed at the highest level ('knowledge-based').

In sum, various studies have been done to learning processes in ATC, and they confirm existing theories. They can be applied to get a better insight into these learning processes (see also chapter 10).

Learnability

The learnability of the ATC task is known to be limited. Therefore, the selection requirements in ATC organizations generally are very high. There is a reasonable amount of research on selection requirements for air traffic controllers. This research

provides an overview of the cognitive abilities, personality aspects and other personal characteristics required to become a competent air traffic controller (e.g., Ramos, Heil & Manning, 2001; Eissfeldt, Heil & Broad, 2002). Despite the strong selection procedures, a high number of trainees fail in training in many organizations. According to Schneider (1990), this is typical for a high performance skill: practice does not necessarily improve performance, and initial performance often provides a poor prediction of final performance. Eyring, Johnson and Francis (1993) applied Ackerman's theory (Ackerman, 1989, 1990) to air traffic control. They argue that not only innate differences in skill acquisition are important, but also motivational and experiential factors, which are less stable. Self-efficacy is supposed to be strongly related to task performance: successful task performance enhances self-efficacy, while failure lowers self-efficacy. The results confirm that self-efficacy and task familiarity influence skill acquisition as well.

We argued in section 2.2 that some task components are more learnable than others by making the distinction between 'consistent' and 'non-consistent' components (e.g., Schneider, 1990). In our view, it is important to distinguish between them because it may help to predict learning in future training. This will be elaborated in chapter 11.

3.3 Training design for ATC

This section reviews the literature concerning the design of a training system that optimally facilitates competence acquisition in ATC. In the review, we use a framework that illustrates the various aspects of training design and interactions (see Figure 3.2):

Figure 3.2. Framework of training design

We distinguish the following notions. The trainee undertakes *learning activities* that directly relate to the (learning) task. The coach is interaction with the trainee by ap-

plying *coaching interventions* and is responsible for *task selection* as far as the learning tasks are not a standard part of the curriculum. The final goal is to achieve the *training objectives*. Learning occurs in the *learning environment*: in ATC this usually involves simulator followed by on-the-job training.

Learning environment

Simulator training is particularly applied in domains in which safety plays a crucial role. It is assumed that trainees can be trained to the highest possible level of competence, while avoiding any damage or fatality in on-the-job training (OJT). In OJT, the trainee usually learns "how real traffic is prioritized, how workload is controlled and planned, how tasks are actually done and how controllers function as a team" (Hopkins, 1995, p. 162).

Simulator training

Literature on simulator training is available for air traffic control (e.g., Hopkins, 1995; Manning & Stein, 2005), process industry (Ridderbos, 1992), shipping (Kojevnikov, 2001), medicine (Helmreich & Davies, 1997; Morgan & Guest, 2001; Adrales et al, 2002; Pugh & Youngblood, 2002), aircrew (Beaubien, Baker & Salvaggio, 2004; Roessingh, 2005) and the military (Farmer et al, 1999).

Training simulators have to be used when situations cannot be trained in reality or for safety reasons, for instance, emergencies or insufficient opportunities to train on the real system due to costs, time or environmental factors (Ridderbos, 1992; Farmer et al., 1999). In addition, simulators are supposed to have many didactical advantages above the on-the-job training. Farmer et al. (1999) summarize them as follows: 1) control of the type and timing of training events, and in this way the learning experiences offered to the trainee; 2) adapting the training task to individual performance; 3) providing cues and feedback extrinsic to the (training) task; 4) registering and diagnosing trainee performance; 5) improving efficiency by automating training and instruction. Making mistakes and experiencing the consequences of mistakes without causing any danger, are also assumed to support learning processes (Ridderbos, 1992; Farmer et al., 1999). It makes (guided) discovery learning and experimental learning possible (De Jong & Van Joolingen, 1996). The technical possibilities to freeze the simulator, replay and playback functions, and logging functions add to the didactic advantages of simulators. For instance, a discussion with the trainee about alternatives ('what-if' scenarios) can be held. Others argue that *authenticity* is the surplus value (Laurillard, 1993; Petraglia, 1998). Some speak about the strengths of learning in Virtual Reality environments because of the possibility of 'immersion' (Kommers & Zhiming, 1999). Learning in these environments is supposed to enhance transfer to the work environment; experimental research on transfer of training in virtual environments to reality has provided some evidence for this (Rose, 2000).

A considerable amount of literature focuses on *fidelity* or physical similarity with reality. Fidelity is sometimes seen as the primary aim of simulation, but we agree with others that the didactic value of simulator training is dependent on more factors such as scenario development and the quality of coaching (Farmer et al., 1999).

High fidelity simulators are usually found in safety-related domains, such as aviation and the military, but it appears from several studies that in these domains low-fidelity simulation can be very useful as well. One of them is a study to the transfer of manual flying skills from PC-based simulation to actual flight (Roessingh, 2005). Roessingh found some evidence for a positive transfer of flying skills that have been learned in this low-fidelity simulation.

However, guidelines on simulator training design are lacking or formulated only in general terms. The literature is mainly restricted to simple skills, while simulator training is especially applied to train complex skills (Farmer et al, 1999; De Jong & Van Joolingen, 1996; Manning & Stein, 2005).

On-the-job training

Jacobs and Jones (1995, p. 12) define on-the-job training (OJT) as "the process in which one employee, most often the supervisor or lead person of a work area, passes job knowledge and skills to another employee. OJT occurs at the location in which the work is done or at least as near to the work as possible, and it is often thought of as involving both learning and doing at the same time". Its main advantage is supposed to be transfer of training. Transfer increases when there is a close match between the training setting and job setting (Jacobs & Jones, 1995). Proponents of 'situated learning' support this idea strongly (Brown, Collins & Duguid, 1989; Lave & Wenger, 1991; Petraglia, 1998).

Learning at the workplace, or on-the-job training, generally implies *informal learning* (Bolhuis & Simons, 1999; Onstenk, 1997). Informal learning can be characterized as an absence of structure and guidance of learning by the environment as opposed to *formal learning* in educational settings or organized courses (Onstenk, 1997). Informal OJT occurs without an advanced planning or involvement by the management (Jacobs & Jones, 1995). Learning-by-doing is an important characteristic of informal learning. Informal learning is usually accompanied by *incidental learning*. Incidental learning is supposed to be non-intentional and implicit. The learner does not have control over his own learning process, but learns unconsciously during work or other activities (Onstenk, 1997; Bolhuis & Simons, 1999).

A related distinction is *unplanned / unstructured* versus *planned / structured* on-the-job training (Rothwell & Kanazas, 1994; Jacobs & Jones, 1995). Unplanned OJT is only driven by work demands and crises. Trainees acquire job knowledge and skills from some explanation or demonstrations by others, through trial-and-error, self-motivated learning or questioning, and by imitating the behavior of others (Jacobs & Jones, 1995). Research has shown that unstructured OJT leads to increased error rates, lower productivity and training efficiency. The main problem seems that there is no standardization of level of expertise, training content and training methods (Jacobs & Jones, 1995; Rothwell & Kazanas, 1994).

Jacobs and Jones (1995, p. 22) define structured on-the-job training as: "the planned process of developing task-level expertise by having an experienced employee train a novice employee at or near the actual work setting". Like other structured training approaches, structured OJT requires a substantial investment of time and effort before it can be used. The interaction between the individuals focuses on passing along expertise about specific areas or tasks. The information is put there intention-

ally and coaching occurs in a structured way (Jacobs & Jones, 1995). Although OJT has been applied since the Middle Ages, it has been largely ignored in past decades. With the new economy, the importance of training in organizations has increased and interest in structured OJT has returned (Jacobs & Jones, 1995).

Air traffic control training can be characterized as *formal learning* and *intentional learning*. The learning process is usually organized in terms of a fixed number of hours to be trained. Certified OJT- instructors train the trainees at the workplace (ESARR5, 2000), and examinations are held. However, in our opinion, the OJT in ATC training is usually not structured in the way as explained above.

Training objectives

The formulation of training objectives often is the first step in any training design process (Kessels & Smit, 1999; Romiszowski, 1988; Merrill, 1994). Task selection, coaching interventions and learning activities depend on training objectives. In ATC, formulating training objectives is required by European regulations (ESARR5, 2000). There is some discussion as to how training objectives (also called performance objectives, instructional objectives, behavioral objectives, just objectives and so on) have to be formulated, especially in the instructional design literature (Kessels & Smit, 1999). Mager (1962) was one of the first who gave a prescription: objectives have to describe the behavior that the trainee is expected to perform at the end of learning, including the conditions imposed on the learner and including the criterion level of minimal acceptable performance. Mager's definition is generally accepted and further elaborated by instructional designers like Gagné, Romiszowski and Merrill.

Training objectives are generally presented as a list of tasks (Farmer et al., 1999). However, we agree with Farmer et al. that these kinds of objectives do not fit well in a competence-based training design. They do not indicate the competences to be acquired for performing those tasks. Training objectives are often formulated at a very detailed level and relations between objectives are not directly visible. This makes it difficult to determine the order of stages a novice should pass to become an expert in a particular domain.

Task selection

Various factors are assumed to play a role in task selection for simulator training. Schneider (1990) argues that the task sequence (cf. set of simulator exercises) should offer a gradual increase in complexity. The number of new learning tasks has to be limited to avoid cognitive overload such as temporary memory. This is needed for the development of automatic components. Automation of skills calls for repetition (Schneider, 1990; Mulder, 1986). Shebilske, Goettle & Garland (2000) add that it is important to build in events, novel situations or unexpected situations. These situations or interruptions, as a kind of stressor, aim at bringing trainees out of their routine level of working and force them to switch from automated to controlled information processing (Shebilske, Goettle & Garland, 2000).

It appears to be that a well-balanced mix between novel situations (events) and repetition of well-known tasks is required to train complex skills. Moreover, a well-known distinction in task selection is *part-task training* versus *whole-task training* (Van Merriënboer, 1997; Van Merriënboer & Kirschner, 2007; Schneider, 1990; Farmer et al, 1999). Part-task-training is needed for automation of recurrent component skills, while whole-task training is required for the acquisition of non-recurrent skills. An optimal balance between part-task and whole-task training is supposed to result in optimal learning processes.

Another issue involves the *adaptivity* of task selection (Laurillard, 1993). The most optimal situation is assumed to be one in which task characteristics are maximally adapted to the trainee's progression, learning styles, prior knowledge, acquired competences and experience, emotional states and so on. This implies that flexibility in task selection is required. It has a relationship with ATI (aptitude-treatment interaction, Snow (1989); see section 2.3). Various studies were done on *dynamic task selection* in the ATC domain (Salden, Paas, Broers & Van Merriënboer, 2004; Salden, 2005; Salden, Paas, Van der Pal & Van Merriënboer, 2006; Van Merriënboer & Kirschner, 2007). Salden et al. (2004) believe that ATC training is usually not efficient enough due to the non-dynamical instructional methods. Dynamic task selection implies that training is adjusted to the cognitive state of the learner, and that the cognitive load (cf. mental effort) being imposed is controlled. The researchers assume that training conditions will be most efficient if they involve high performance in combination with low mental effort. Salden et al. (2004) measured cognitive load and performance level in initial exercises, and they adjusted the next exercises to the learner based on these measures. Their study supports the hypothesis that dynamic task selection improves training. We believe that dynamic task selection certainly will improve the quality of simulator training. However, we should realize that these studies were done with novices in a simplified ATC task in which the variables can be controlled much easier than in practice.

Unfortunately, there is hardly any literature about task selection in OJT. In OJT, the possibilities to pre-structure the tasks to be executed are limited. In many working environments, it is possible to build up task complexity by making trainees begin with easy tasks (Jacobs & Jones, 1995). Experienced job performers, like the coach, may take over tasks that are more complex. However, this is difficult to realize in ATC, because it is not possible to predict in advance which tasks are available for training.

Learning activities

Learning activities are defined as the (cognitive) activities undertaken by the learner in order to influence the learning process (Lowijck & Verloop, 1995). They are called *training activities* by Farmer et al (1999): those activities that the trainee has to perform to achieve the training objectives. While training objectives specify the standards of performance, learning activities specify how these standards are to be achieved. Learning activities can be different from operational activities in order to achieve intermediate skill levels. The gap with operational activities is usually larger in initial training phases than later on (Farmer et al, 1999). For instance, in ATC

standard procedures (e.g., flying the basic circuit) are usually acquired in initial exercises, but deviations from these standards (e.g., straight-ins) are not allowed yet while they are more common in real operational activities. This is supposed to be valuable for the automation of routines. It relates to strategy learning (see section 3.1). Standard procedures must be automated because controllers must be able to revert to them in hectic, complex or stressful situations. They need less attention and are less time-consuming (Oprins & Burggraaff, 2006). On the other hand, it is argued that too much emphasis on automation during training, or 'overlearning', should be avoided because this may reduce the effective operation of controlled processing in novel and unexpected situations (Schneider, 1990; Shebilske, Goettle & Garland, 2001; Garland, Stein & Muller, 1999).

It should be noticed that these kinds of learning activities are only possible in simulator training but not in the on-the-job training because of the consequences for safety and efficiency (costs).

Coaching interventions

In simulator training and on-the-job training, the coach usually looks over-the-shoulder of the trainee while providing guidance or feedback on the trainee's activities. This feedback is called *extrinsic feedback* as opposite to *intrinsic feedback*: the trainee does not experience the consequence of his actions directly, but the coach is talking about his/her actions or the result of these actions (Laurillard, 1993).

General guidelines have been defined for coaching in simulator training and on-the-job training in ATC. A coach is supposed to provide a *briefing* before, *tutoring* during and *debriefing* after scenario execution or on-the-job training session (Hopkins, 1995; EATCHIP, 2000). In the briefing prior knowledge is supposed to be activated and the trainee is prepared on the learning activities to make the learning process as effective as possible. The coach may refer to training objectives and perhaps some individual-determined objectives. The aim of the debriefing is to evaluate the trainee's performance. Thinking-aloud protocols and asking questions during and after task execution are stimulated to get insight in trainees' cognitive processes (EATCHIP, 2000).

Moreover, the theory of *cognitive apprenticeship* (Collins, Brown & Newman, 1989) has been widely spread, especially at the workplace or OJT (Onstenk, 1997; Bolhuis & Simons, 1999). According to this theory, coaches have the function of experts who explicate their implicit knowledge to their trainee. The coach starts with accompanying the trainee intensively ('coaching'), followed by just providing tips ('scaffolding') and diminishing guidance later on ('fading'). Finally, the trainee is able to work independently. However, educational psychologists differ in their opinion how much guidance is needed. De Jong and Joolingen (1996) argue for *guided discovery learning* in simulator environments as explained before: an optimal mix between discovery learning and guidance. Others refuse guidance almost completely (e.g., Petraglia, 1998). In our view, a certain amount of guidance is needed, but this should be combined with many possibilities in which trainees can experiment, especially in simulators.

3.4 Conclusions

The literature review in this section has shown that the air traffic control task is highly complex and dynamic, which puts constraints on learnability. Complex cognitive processes are involved and strategies play a central role. In addition, air traffic controllers are working in a safety-critical environment. Safety is the main reason why training usually starts with simulator training before trainees enter the real working environment in the on-the-job training (OJT). Simulator training makes it possible to discover and experiment in a safe environment and in a structured way, while in OJT the traffic is ongoing and learning tasks are difficult to plan in advance. Literature on learning processes in ATC, for instance, by using the Kanfer-Ackerman task, has taught us that ATC competence acquisition is rather complex.

Training design aims at facilitating these learning processes in an optimal way. The literature provides some guidelines on the design of ATC training described in this study (see chapter 7). We believe that in competence-based training, training objectives have to be defined in terms of competences. They indicate what has to be learned, for both simulator training and OJT. We learned from the literature that various issues in task selection (cf. simulator exercises) should be taken into account in simulator training design: a gradual increase in complexity, a well-balanced mix between novel situations and repetition of well-known tasks, and an optimal balance between part-task and whole-task training. Task selection has to be adaptive to the trainee's needs as far as possible. Learning activities have to be chosen in such a way that they are optimal for the trainee's learning process. This implies that it may be useful to automate standard procedures in initial training phases, although the operational activities being executed in the real-life environment may be more complex. Strategy learning has to be emphasized by letting trainees experiment and discover own preferences in various task situations. Coaches should provide a briefing and a debriefing in a training session. Usually, guidance diminishes during training over time in conformance with the principle of 'cognitive apprenticeship'. Coaches should get insight into the cognitive processes of trainees by asking questions and by applying think-aloud protocols.

This chapter gives a review of the literature on assessment and assessment design as far as relevant for the development of an ATC assessment system. We make a distinction between assessment design embedded in educational and training design on the one hand, which is usually covered by instructional design theories (e.g., Romiszowski, 1988; Gagné, 1985; Merrill, 1994), and assessment design as a self-standing activity on the other hand, which is common in performance appraisal (e.g., Murphy & Cleveland, 1995; Roe & Daniels, 1994). As the educational perspective and the organizational perspective on assessment systems are both relevant for assessment in the context of competence-based training, we review both perspectives and try to integrate the findings they yield.

First, we address assessment in relation to learning processes and discuss several new forms of assessment that result from recent developments in learning theory (section 4.1). Second, we pay explicit attention to *performance assessment* as applied in organizations (section 4.2). Third, existing assessment systems for training in ATC-related domains are reviewed (section 4.3).

4.1 Assessment and learning

Views on assessment

The purpose of assessment in training is to evaluate the level of competence obtained. Therefore, there exist a strong relationship between learning and assessment. As theories of learning have developed from behaviorism towards cognitivism and constructivism, views of assessment have changed as well.

Behaviorism

Assessment plays a central role in behaviorist ideas about learning: rewards and punishments are considered reinforcers that are responsible for learning (Mayer, 1987). This implies that behaviorists assume that learning is just the result of feedback on the learner's actions (stimulus-response). The role of the learners themselves and their underlying cognitive processes are entirely ignored. Behaviorists pay much attention to the design of assessment systems because assessment is supposed to be mainly responsible for the learning process. These assessment systems measure only the outcomes (cf. the right answers) instead of the cognitive

processes that lead to these outcomes. This makes it possible to concentrate on objective measurement using sophisticated psychometric models. Instruction and assessment are considered separate activities. The teacher or trainer is responsible for instruction, the measurement expert for assessment. This has been referred to as a 'testing culture' (Wolf, 1993).

Although many assessment systems are still based on behaviorism, educational psychologists criticize them for several reasons. Among the criticisms is the behavioristic assumption that complex skills can be broken down in small, isolated components that can be assessed separately (Resnick & Resnick, 1992; Jonassen, 1992; Pellegrino, Chudowski & Glaser, 2001; Gipps, 1994; Birenbaum, 1996).

Cognitivism

In the cognitive view of learning, one of the most important roles of assessment is to provide effective and informative feedback during training or instruction. Training would improve if instructors have more insight in thinking patterns or strategies trainees use. Cognitivists emphasize the assessment of the cognitive *process* besides assessment of the *product* or outcomes (Jonassen, 1992; Gipps, 1994; Pellegrino, Chudowski & Glaser, 2001). The main rationale for the process-oriented approach is that assessment improves learning; assessment and learning are interdependent. This is called an 'assessment culture', because a continuous dialogue is going on between learner and trainer, assessment being embedded in instruction (Wolf, 1993; Birenbaum, 1996).

Measuring how many items were right or wrong, as in the behavioristic approach, cannot assess the underlying cognitive processes. Pellegrino, Chudowski & Glaser (2001) argue that "a model of cognition and learning should serve as the cornerstone of the assessment design process. This model should be based on the best available understanding of how students represent knowledge and develop competence in the domain" (p. 3). In addition, progression is supposed to be an important indication for learning and should be assessed too. Consequently, assessment should not be restricted to particular points in time but take place continuously (Pellegrino, Chudowski & Glaser, 2001).

Finally, it is assumed that assessment of the cognitive process results in metacognitive awareness of learning, which enhances the learning process itself (Jonassen, 1992; Gipps, 1992; Pellegrino, Chudowski & Glaser, 2001). Metacognition is defined as 'thinking about thinking'. Learning is improved because learners are aware of and in control of one's own knowledge and thinking, while they use particular strategies to reflect on the meaning of what they are learning (Gipps, 1994).

Constructivism

From the constructivist perspective, the assessment of the cognitive process is considered insufficient, as the context has to be taken into account as well. Learning and assessment have to take place in concrete and authentic situations and are often combined. Testing in abstract situations cannot predict how learners will perform in real environments (Pellegrino, Chudowski & Glaser, 2001). Evaluation should occur in contexts that are as rich and complex as those used during instruction or training (Jonassen, 1992; Cronin, 1997). Assessments should involve the direct observation

of performance on tasks that are valued in their own right as contrasted with tests that are represented as mere indicators of valued 'real world' performances (Linn, Baker & Dunbar, 1991; Linn & Baker, 1996; Wiggins, 1990; Resnick & Resnick, 1992).

According to Jonassen (1992), assessment should be more 'goal-free' because the learner constructs his knowledge. The assumption is that a single reality does not exist. Jonassen argues that the learning process and assessment will be biased if the specific goals are known before the learning process starts. A multi-perspective approach is needed. A domain of possible outcomes has to be applied instead of one single type, and more perspectives from different assessors should be used instead of only one (Jonassen, 1992; Cronin, 1997). This view implies that assessments require open-ended tasks to assess higher order skills or problem solving strategies (Gipps, 1994; Pellegrino, Chudowski & Glaser, 2001).

As explained in section 2.3, the constructivist approach relates most to the competence-based view on training and assessment. We agree that assessment of competences should take place in realistic settings. Because competences imply the merging of knowledge, skills and attitudes, it is useless to assess smaller pieces. This does not provide enough information about the trainee's overall performance when executing tasks in realistic settings. Assessment of competences allows adapting training to the individual's needs, based on his/her strengths and weaknesses. Assessment of cognitive processes and higher order skills should be emphasized. Information from multiple assessors is needed to get a complete picture.

Forms of assessment

Several forms of assessment exist that were developed in educational settings, although the majority can be applied in training in organizations too.

Formative versus summative assessment

The purpose of formative assessment is to assist learning, while the purpose of summative assessment is to assess individual achievement (Pellegrino, Chudowski & Glaser, 2001). Formative assessment takes place during teaching and the information is used to adapt teaching and learning to meet student needs (Mayer, 1987; Gipps, 1994; Boston, 2002). Summative assessment takes place at the end of a learning period and is used to provide information about how much students have learned. External standards are usually formulated against which students are assessed. Formative assessment measures progress in learning over a certain period, whereas summative assessment takes place at one moment in time (Boston, 2002).

Formative assessment has become more popular because the purpose of assisting learning has received more emphasis. Black and William (1998) conclude that formative assessment can improve learning if it is appropriately used. Summative assessment is also needed to objectively assess what students have learned at certain moments of time, for instance, to decide about further education. In our view, these

arguments ask for a combination of formative and summative assessments in training.

Peer- and self-assessment

Peer assessment means that other students in educational settings or colleagues in working environments assess their peers, and self-assessment implies that people assess themselves (e.g., Klarus, 2000; Roe & Daniels, 1994). A specific method is *360° feedback*: a person does not only assess himself, but is also assessed by a group of other interested people. Self-assessment is supposed to support self-reflection on learning processes and performance. The reliability seems to be often high because people have usually a good image of their own performance (Roe & Daniels, 1994).

Norm-referenced versus criterion-referenced testing

Norm-referenced testing means that students are compared with each other, while criterion-referenced testing implies that it is measured whether the objectives or external standards are achieved (Mayer, 1987). Glaser (1963, p. 520) has defined criterion-referenced testing as follows: "measures which assess student achievement in terms of criterion standard thus provide information as to the degree of competence attained by a particular student which is independent of reference to the performance of others". Norm-referenced and criterion-referenced measurements are often combined (Gipps, 1994). Norm-referenced testing can be useful to determine whether a student is progressing at the same rate as his peers (Pellegrino, Chudowski & Glaser, 2001), but it does not show what a student can or cannot. Therefore, criterion-referenced testing dominates in education and training (Glaser, 1963).

Dynamic assessment

Dynamic assessment aims at the assessment of learning ability or cognitive functioning (Schneider Lidz, 1987). In contrast with static testing, the common feature of dynamic assessment is that the psychological processes involved in task performance are measured (Campione & Brown, 1990). The process of learning itself is evaluated. Dynamic assessment is used to measure learning potential otherwise than by measurement of intelligence. It can also be used to identify strengths and weaknesses to develop remedial programs or to guide instruction (Campione & Brown, 1990). The specific learning problems are distilled which makes it possible to apply individualized instruction adaptive to the learner's needs.

This method is strongly influenced by the ideas of Vygotsky (Campione & Brown, 1990; Gipps, 1994), in particular by his notion of 'zone of proximal development'. This notion refers to the distance between the level of performance that a child can reach unaided and the level of performance that a child can reach by guidance by a more knowledgeable participant such as a teacher. The zone of proximal development implies the amount of aid needed to show progression in the learning curve, which differs for each learner. Vygotsky emphasizes that learning is mediated by social interactions and propagates collaborative learning environments in accordance with the constructivist view on learning. By observing learning in the zone of

proximal development, information can be gained about the efficiency with which new cognitive resources are acquired (Campione & Brown, 1990).

As in formative assessments, the teacher is also assessor and assessment takes place in a learning situation; assessment and instruction are interdependent (Campione & Brown, 1990). Learners get a task one step beyond their competence. During the learning process, the degree of aid needed to learn new principles and to apply them is measured. The idea is that the amount of aid required is a good indication of student's further learning. Learners only get help when they need it, so that they can show their own competences (Gipps, 1994). In this way, dynamic assessment can be used to predict how learners deal with novel problems and to separate good from poor learners, while static measurements may underestimate the learner's capabilities (Campione & Brown, 1990). Although dynamic assessment originally has children as a target group, it has been applied in other assessment situations too, for instance, with college students (Campione & Brown, 1990).

Diagnostic assessment

Regian and Schneider (1990) apply so-called diagnostic assessment to predict skill acquisition in complex tasks. They argue that diagnostic assessment focuses on the assessment of both stable and unstable skills. It is called diagnostic because it aims at finding causes for insufficient performance. Performance of stable skills does not increase with practice, for instance, due to deficiencies on short-memory capacity or working memory capacity, whereas performance of unstable skills does (see also section 2.3). In diagnostic assessment the student's speed, accuracy and workload for a given component of a complex skill should all be measured according to Regian and Schneider. They call for a closer relationship between training, prediction and diagnostic assessment, keeping in mind trainable and less trainable skills.

Seamster, Redding and Kaempff (1997), referring to aviation, state that diagnostic assessment is essential when specific remediation is needed or when specialized training must be designed. They support 'cognitive assessment' which is needed to get diagnostic information. In their view, cognitive assessment aims at measuring the stages in trainees' skill development, their representational skills, their domain relevant strategies and changes in their structural knowledge. Cognitive task analysis is required to identify what have to be assessed and which cognitive processes are underlying performance in a specific task execution (Seamster, Redding and Kaempff, 1997).

These new forms of assessment appear to fit well in competence-based assessment systems. Dynamic and diagnostic assessment focus on underlying cognitive processes and on the amount of learning. Formative assessment also emphasizes the importance of assessing progression. In our view, this is important for assisting learning. Criterion-referenced testing dominates in the type of system that is subject of this study, since trainees are assessed against standards. This seems also to be the reason why peer- and self-assessment are not applied very often yet, although these forms may stimulate self-reflection.

4.2 Performance assessment

As mentioned before, performance assessment is usually applied in training systems that are discussed in this study. Therefore, this section addresses this subject in more detail.

Definition

In educational settings, performance assessment belongs to the new forms of assessment, although it has been applied in organizations for many decades. Most literature available is about personnel appraisal and selection in organizations. Berk (1986, p. 9) defines performance assessment as follows: "Performance assessment is the process of gathering data by systematic observation for decisions about an individual". Others, especially from the perspective of education, emphasize the authentic character of performance assessment. They characterize performance assessment to be assessment by direct observation of the learner's behavior, carried out in authentic tasks (Pellegrino, Chudowski & Glaser, 2001; Linn & Baker, 1996; Birenbaum, 1996; Gipps, 1994; Klarus, 2000; Tillema, 1996). Therefore, performance assessment is strongly related to 'authentic assessment' (Wiggins, 1990; Resnick & Resnick, 1992), but it is not identical. Authentic assessment is performance assessment carried out in an authentic context. This implies that not all performance assessments are authentic, but authentic assessments that are not performance assessments do not exist (Gipps, 1994). Similar terms of performance assessment are 'performance-based assessment', 'performance measurement' or 'performance testing'.

A specific form of performance assessment is a *performance test*. A test is considered to be a "systematic procedure for evaluating the knowledge, skills or other abilities of a person in some area of endeavor" (Aiken, 1996, p. 19). More specifically, a performance test is supposed to be a test in which performance is demonstrated through directly observable behavior. Frequently mentioned examples are work samples or exams in the simulator or at the workplace (Berk, 1986; Guion, 1998).

Performance assessment has been gaining renewed attention since the 1990s, but the term has existed for more than fifty years (Birenbaum, 1996; Pellegrino, Chudowski & Glaser, 2001). Performance assessment is often the only possible method to assess people's job performance at the workplace (Berk, 1986), while performance assessment methods in educational contexts represents one of several approaches on assessment. In both settings, performance assessment is assumed to strongly relate to assessment of competences (Wigdor & Green, 1991; Klarus, 2000). The emphasis is on higher order skills, including non-cognitive skills, blended together in the execution of tasks in realistic contexts.

Applications

Performance assessment has become more popular in educational settings, because it fits well in modern views on learning and assessment. More practical as-

assessment methods replace the traditional paper-and-pencil tests. Performance assessment is even considered to be any form of assessment different from multiple choice (Gipps, 1994). Often mentioned methods are experiments, simulations, essays, hands-on science problems, and portfolios (Linn & Baker, 1996). Linn & Baker (1996) describe the most important characteristics of performance assessment in educational settings: open-ended tasks; assessment of higher order, complex skills; evaluation over extended periods of time for performance; group performance; choice of task by trainees themselves; human judgmental scoring (rating scales).

Performance assessment has a longer history in work settings and has found several applications. The design and use of the assessment system is dependent on the specific purpose of application (Berk, 1986; Gitomer, 1993; Landy & Farr, 1983; Wigdor & Green, 1991). First, performance assessment is used to make personnel decisions, for instance, about promotions, compensation or administrative actions (Gitomer, 1993; Berk, 1986). This type of assessment is called *performance appraisal* and has received much attention in the literature (Landy & Farr, 1983; Murphy & Cleveland, 1995; Berk, 1986; Landy, Zedeck & Cleveland 1983; Roe & Daniels, 1994). Cascio (1986) considers performance appraisal to be the systematic description of the job-relevant strengths and weaknesses within and between employees. Therefore, performance standards have to be defined based on a job analysis. Murphy and Cleveland (1995) focus on performance appraisal as a social and communication process, taking the social and organizational context into account (see also Roe & Daniels, 1994).

Second, performance assessment is used in selection systems (Gitomer, 1993; Guion, 1998; Roe, 1983). Performance assessments are used as criterion measures for validation of selection, for instance, by means of performance tests at the workplace in which proficiency in some aspect of job performance is assessed (Guion, 1998). Performance tests are also applied as work samples in selection in the consideration that they are samples of a job content domain taken under standard conditions (Guion, 1998). These performance tests are often simulations (Ridderbos, 1992; Eissfeldt, 2002, 2003; Ramos, Heil & Manning, 2001).

Third, performance assessment is used in simulator and on-the-job training. Unfortunately, only a minimal amount of research on assessment in OJT exists, probably because structured OJT has not received much attention. An exception is the assessment of aircrew (e.g., Beaubien, Baker & Salvaggio, 2004). Performance assessment is applied extensively in simulator training in many domains. This offers the possibility for automated measurement (Vreuls & Obermayer, 1985; Sanders, 1991; see section 4.3).

Measurement techniques

Performance can be measured in many ways, dependent on the nature of the task and purpose of assessment. Much has been published about rating, rating scales and raters in performance appraisal systems (Berk, 1986; Landy & Farr, 1983; Murphy & Cleveland, 1995). A basic distinction is between objective (nonjudgmental) measurement and subjective (judgmental) measurement (Landy & Farr, 1983; Murphy & Cleveland, 1995). Examples of objective measurements in appraisal systems

are: time, error and frequency measures or production outcomes. However, these aspects are supposed to be too restricted to adequately cover performance in the majority of jobs (Murphy & Cleveland, 1995). In subjective measurement, one or more assessors give their judgment about the individual's behavior on a rating scale pertaining to a specific criterion (Aiken, 1996). We describe the most common techniques below.

Graphic rating scales

Graphic rating scales are most common of all rating methods. They are mostly used to rate different dimensions of performance (Guion, 1998). There exist many variations. The most basic form is a (unbroken) line, but a set of adjectives is often assigned to scale points (e.g., excellent, very good, good, average, poor, very poor). A more elaborated variation is to add a description of behavior to every anchor in order to make assessments more objective (Aiken, 1994). This may solve the problem of the lack of clarity and definitions of many graphical rating scales (Murphy & Cleveland, 1995). The number of scale points strongly differs, dependent on the discrimination that is possible and required. The five, seven and nine point scales are most frequently used (Guion, 1998). When more dimensions are used to produce a final judgment over performance, a specific weight can be assigned to each dimension to indicate how important a dimension is considered to be in comparison with others (Aiken, 1996).

Behaviorally anchored rating scales (BARS) and variations

Smith and Kendall (1963) have tried to enhance and objectify the graphic rating scale by developing a new form of ratings. The specific form gained attention and resulted in the *behaviorally anchored rating scales* (BARS) method (Guion, 1998; Borman, 1986). The basic idea is that a group of subject matter experts (SME's) achieve consensus on a series of behaviorally descriptive statements on several dimensions. These dimensions are used for subsequent appraisal. The construction of BARS requires a specific method. First, critical incidents analysis is needed to identify the most critical situations in a specific job. Next, the critical incidents are assigned to different dimensions, and SME's link the incidents to specific scale positions (e.g., good or slightly good performance). A rather objective and reliable rating scale can be constructed after several reordering and retranslation cycles within the group of SME's. The BARS technique seems to be psychometrically superior to others (Aiken, 1996; Roe & Daniels, 1994), although studies have shown that most common errors of rating scales remain to occur.

Some alternatives to BARS were proposed that are less time-consuming to develop (Murphy & Cleveland, 1995; Guion, 1998; Aiken, 1996). The first is the *behavioral expectation scale* (BES). It differs from the BARS technique insofar that behaviors are rated in terms of expectations instead of actual behavior. A second alternative is the *behavioral observation scale* (BOS). This technique is identical with BARS as for the use of critical incidents on several performance dimensions (Aiken, 1996; Guion, 1998), but BOS ratings are based on the frequency of behavior that is observed in a specific period (e.g. never, seldom, sometimes, generally, always). BOS is assumed to have advantages above BARS because few complex inferences

are required of the rater (Borman, 1986). A third technique concerns the *behavioral criteria*, which is often used in assessment centers (Roe & Daniels, 1994). Criteria are supported by a description of specific, observable behavior, but these behavioral descriptions are not linked to specific scale positions. Reasons to choose for this technique instead of BARS is that critical incidents are too specific and occur not frequently, and that its construction is very time-consuming.

Checklists

Ratings can also be based on checklists with behavioral items that are checked by the rater. Guion (1998) describes two variants: the method of equal-appearing intervals and the method of summated ratings. The items of the checklist are previously scaled but their weights are not visible to the rater (Roe & Daniels, 1994). The final score on a dimension is calculated based on the underlying weights of items.

Performance distribution assessment

Kane (1986) proposes a measurement method based on variability in performance over time. It is called *performance distribution assessment* (PDA), and this method can be seen as a variant of BOS (Murphy & Cleveland, 1995). In PDA, raters must indicate the frequency of different outcomes (behaviors, results) that indicate different levels of performance on a given dimension. For instance, the scale describes the most effective and the least effective outcome and some intermediate outcomes. The rater has to estimate the frequency of each outcome level for each ratee. The distribution of variability in performance is measured as well as the average level. As PDA involves very complex scoring rules, it has not been widely used in practice (Guion, 1998). Moreover, it appears to be difficult for raters to objectively indicate the frequency of specific behaviors.

Ranking

Ranking methods are frequently used to compare people with each other. Ratees are ordered from the best to the poorest on an ordinal or an interval scale (Guion, 1998). Ratees can be compared with each other in pairs on a specified dimension. The one chosen most frequently is on the top of the rank order (Guion, 1998). However, this method is only applicable with small groups of ratees (Roe & Daniels, 1994).

Rating process and rating errors

Assessors usually observe a candidate's performance and give their judgment, for instance, by using a rating scale. Assessors interpret whether a candidate possesses the competences required, and this process inevitably implies a certain extent of subjectivity. The rating process is a complex cognitive process that is influenced by many factors (Murphy and Cleveland, 1995): raters mentally encode the information from the observations, they store it in their memory and retrieve it at a later time, and they integrate information from several different sources. Therefore, the quality of ratings heavily depends on the quality of the observations. Impres-

sions, expectations and experiences of the raters may have a strong effect (Guion, 1998).

Therefore, errors may occur which influence the psychometric quality of assessments. Some classical psychometric errors are always mentioned in the literature (e.g., Guion, 1998; Roe & Daniels, 1994; Murphy & Cleveland, 1995). First, *central tendency* implies that raters cluster all ratings around a central scale point, avoiding extreme judgments. Second, *leniency* or *severity* refers to raters who tend to respectively higher or lower judgment in a systematic way. Third, *halo* is present when judgments on different performance dimensions are too highly intercorrelated, caused by the tendency to judge someone as good or inferior in general (see chapter 9 for an elaboration on rating errors). A possible solution is supposed to be assessor training (Guion, 1998; Murphy & Cleveland, 1995). Frame-of-reference (FOR) training (Bernardin & Buckley, 1981) aims at matching behavior to performance dimensions. This helps to achieve common understanding of the dimensions by different raters.

Psychometric quality

The psychometric quality is assumed a crucial requirement of any assessment system. This involves *reliability* and *validity* (Guion, 1998; Roe & Daniels, 1994). Reliability is considered a precondition for validity: the assessment system cannot be valid if it is unreliable. Chapter 9 addresses this subject in more detail; we only describe some main aspects here.

Reliability

The reliability of assessment or testing implies that repeated measures lead to similar results (Drenth & Sijtsma, 1990). General test theory knows various types. First, *test-retest reliability* refers to repeated administration of the same test at a later moment. Second, *alternative form reliability* is examined by comparison of two versions of a test or a comparable test. Third, *split-half reliability* implies splitting the test into two parts that should be interchangeable. Fourth, *internal consistency* refers to the extent in which the items measure the same attribute, usually represented by Cronbach's alpha (Wigdor & Green, 1991). Finally, *interrater reliability* is relevant in assessment systems in which multiple assessors (raters) are involved (Murphy & Cleveland, 1995; Guion, 1998). Guion (1998) makes a distinction between *interrater reliability* and *interrater agreement*: assessors agree if they make the same ratings, but they are reliable if they put ratees in a comparable relative order.

Validity

Validity of tests refers to the extent in which a test achieves its goal (Drenth & Sijtsma, 1990). Generally, four main types of validity are distinguished. First, *predictive validity* refers to the degree in which a test predicts future performance by determining the relationship between criterion measures and predictors. Second, a special type of predictive validity is *concurrent validity*, which only differs in time: criterion measures are not in the future but in the same time period, for instance, in another test. Concurrent validity is usually examined if predictive validity cannot be estab-

lished. Third, *content validity* involves the extent in which the test is representative for the content in training (Messick, 1989; Shavelson, Gao & Baxter, 1996). Content validity is related to *face validity*: how representative is the test in relation to the job? The fourth one is the most difficult one to measure: *construct validity*. Traditionally, construct validity refers to the psychological constructs that are measured in a test (Drenth & Sijsma, 1990). Messick (1993, 1994) views construct validity as a unified concept that can be split up into content, substantive, structural, external, generalizability and consequential aspects. From this perspective, content validity is assumed an aspect of construct validity (Birenbaum, 1996).

4.3 Assessment in ATC-related learning environments

Performance assessment

In ATC training, performance assessment is applied to assess the competences that trainees have acquired. The most widely used method of performance assessment, over-the-shoulder observation (OTS) (Wickens et al, 1997), is applied in simulator training and OJT. Several checklists and rating scales are used.

Common in many ATC training organizations is *continuous assessment*. This means that assessment takes place not only once, but at various moments during training (Hopkins, 1995). We could say that continuous assessment builds on the aforementioned new perspectives on assessment (see section 4.2). Continuous assessment has similarities with 'dynamic assessment', because assessment is partly based on the amount of instruction a trainee needs. It also relates to 'diagnostic assessment', because the tutoring of the coach depends on the diagnostics about the trainee's performance. However, trainee's performance cannot objectively be measured during training, because the coach is influencing the trainee's performance by providing feedback. Therefore, continuous assessment is often combined with performance tests.

Although the number of studies on ATC assessment is very limited, a substantial amount of literature is available about performance assessment systems for aircrew. One of these assessment systems is NOTECHS (Van Avermaete & Kruijsen, 1998; O'Connor et al., 2002; O'Connor, 2004; Flin et al., 2003; Flin, 2004). NOTECHS is the result of a European project. It aimed at the design of "a feasible and efficient method for assessing pilots' non-technical skills during flight and simulator checks" (Van Avermaete & Kruijsen, 1998; p. 15) to be used by various airline industries. In the NOTECHS project, behavioral markers were designed based on a comparison between evaluation systems of airlines all over the world (Flin & Martin, 2001). KLM designed one of these systems originally developed for Crew Resource Management (CRM) training. The system comprises five main categories to assess non-technical skills. The categories are designated with the acronym WILSC: Work attitude, Information management, Leadership, Stress management and Cooperation (Antersijn & Verhoef, 1995). Each category contains some sub-categories with a definition and several behavioral markers. Later on, KLM and Transavia adopted

SHAPE (Self, Human interaction, Aircraft, Procedures, Environment and situation) that is derived from WILSC. The SHAPE system resembles the aforementioned BARS technique (Smith & Kendall, 1963) and was an important source for NOTECHS. The NOTECHS system has four rating categories, anchored by 'behavioral markers': 1) Cooperation, 2) Leadership and managerial skills, 3) Situation awareness; and 4) Decision making. The system was evaluated in the JARTEL project (O'Connor et al., 2002, O'Connor, 2004). The results showed that NOTECHS provides a usable and reliable assessment method for flight checks in a controlled condition (O'Connor et al, 2002).

Apart from the NOTECHS project, other efforts have been made to develop assessment instruments for aircrew. Proficiency of individual pilots is measured by methods such as 'line checks', 'line operational evaluations' (LOE) (Goldsmith & Johnson, 2002), or 'line operations safety audit' (LOSA) (Thomas, Sexton & Helmreich, 2004). In these systems, experts observe the pilot's performance in the cockpit during normal flights, using a system of behavioral markers to rate pilot's behavior. Simulators are also used to train and assess aircrew. They are referred to as 'simulator checks' or 'line operational simulators' (LOS) (Beaubien, Baker & Salvaggio, 2004). The assessors are expert-observers who are commonly trained in an assessor training. Calibration sessions during training aim at raising agreement between assessors about required performance standards (Goldsmith & Johnson, 2002).

Automated measurement

Automated assessment in the simulator implies that assessment is automated, for instance, by using logged data. Automated measurement has been used for training purposes, selection systems and proficiency checks in various settings (Guion, 1998; Sanders, 1991; Vreuls & Obermayer, 1985). Automated measurement in simulator training has also been applied in ATC, but scientific research is rare.

The first proposals for ATC were already made many years ago (Boone, 1976; Buckley & Rood, 1977). Buckley and Rood (1977) identified four important and independent performance categories: confliction (number and duration of conflicts), occupancy (time and distance flown under control), communication (number and duration of ground-to-air communication), and delay (number and duration of delays). Boone (1976) suggested that the simulator should calculate how well aircrafts are separated, potentials for conflicts, delay time, number of aircraft handled, number and duration of communication transmissions and so on. However, Manning and Stein (2005) argue that it is a problem to understand what the measures mean. Numbers can be computed that describe what is happening in an ATC traffic situation, but it is difficult to determine which numbers are meaningful. Another problem that they mention is that simulator scenarios develop differently because there are various ways to handle the air traffic. This causes a lack of consistency: work sample testing allows the same initial situation, but the outcome measures will depend on the working methods of the candidates.

Other examples can be found in the domain of medicine. Studies were performed in training in a physical examination (pelvic) simulator (Pugh & Youngblood, 2002), and a laparoscopic simulator (Adrales et al, 2003). Both studies aimed at the

validation of the automated measurement method. In these simulators, measurement of trainee's competences was performed with the aid of computer-generated data, which are captured and converted into scores, using variables such as time to complete the task (Pugh & Youngblood, 2002). The researchers argue that while traditional learning in simulators is highly dependent on the quality of the coach and subjective evaluation, automated assessment provides objective and corrective feedback that helps trainees learn from their mistakes.

However, automated performance measurement in training simulators has not been widely used in general. Vreuls and Obermayer (1985) argued that most existing systems are so poorly designed that they are rather useless. They mentioned four fundamental problems. First, the hidden part of knowledge and the embedded nature of performance cannot be measured, while complex processes such as decision-making, system control and teamwork are usually involved in the types of tasks to be trained in a simulator. Second, there is a lack of general theories of human performance that can predict actions over a wide range of circumstances. Third, the validity of performance measurement has hardly been tested. The fact that this is costly and time-consuming poses an obstacle to the advancement of automated measurement. Fourth, there is a lack of quantitative criteria for performance for many tasks; criteria such as the amount of time on a job or exposure to various situations are not adequate for training purposes.

Sanders (1991) recognizes this and suggests as an underlying issue that simulator designers focus more on physical fidelity than on human performance. He distinguishes three problems with performance measurement. First, cognitive skills are more difficult to measure than perceptual-motor skills. Second, the adequacy of responses that are time-locked to outside stimuli is difficult to measure if relations between environmental stimuli and human responses cannot be established. Third, optimal performance is difficult to define since clear and univocal criteria are lacking. Sanders (1991) finds verbal reports ("thinking aloud") useful for the measurement of cognitive skills. He suggests that subjective evaluations, providing pass-fail criteria in many training programs (car driving, air traffic control), will remain the only possibility for performance measurement in the majority of complex skill training. Farmer et al. (1999) and Hopkins (1995) also suggest that objective performance measures are not sufficient to fulfill all purposes of assessment, even if it were technically possible. In their opinion, data gathered from a computer can represent only low-level phenomena rather than complex processes like decision making and perception. Subjective judgments are needed to assess these cognitive processes.

In sum, there appear to be many arguments to assume that automated measurement is insufficient to adequately assess trainees in simulators, despite of its objectivity. Additional subjective judgments are needed to get a complete picture of trainees' performance and to support learning in an optimal way.

ATC work samples for selection

Work samples for selection in ATC, especially those using simulations or simulators, may be useful for assessment in training for several reasons. These work samples are comparable with simulator tests to some extent, because they assess trainability

or skill acquisition in typical ATC tasks (Eissfeldt, 2002, 2003). Much effort is put into reliable and valid assessment methods of these job samples, because they aim to make valid predictions about candidate's future performance. Performance criteria are usually comparable with those applied in training since typical ATC competences are assessed. Various studies on work samples for ATC selection exist. Examples are: the Multiple Controller Aptitude Test (MCAT) developed by the FAA, the Dynamic Air Traffic Control Test (DAC) by the DLR/DFS, and the Strip Display Management test (SDM) also by DLR (Eissfeldt, 2003).

The DAC (Eissfeldt, 2002, 2003) is of special interest in relation to training, because it is a so-called 'learnability test': the test tries to predict learning potential of candidates by means of a structure that involves pretest, feedback and instruction by an instructor, and posttest. The DAC is a work sample in which the candidate has to perform a simplified ATC task. The airspace is presented on a computer screen, the aircraft are displayed as blips with labels attached. The candidate must give commands such as change of altitude, speed or heading by using a special keyboard in order to keep the aircraft separated and to maintain an orderly flow of traffic. Additionally, the candidate has to process several kinds of acoustic information. Some variables are automatically scored by the system: violation, conflicts, errors, efficiency, altitude, precision, and questions. The differences between the pretest and the posttest are calculated to serve as an index for learning potential. The instructors rate some additional aspects of the candidate's performance such as stress resistance, combined with a written description, and they provide a rating that reflects the amount of learning during the test (Eissfeldt, 2002, 2003).

Another work sample is found in the AT-SAT selection battery (Ramos, Heil & Manning, 2001; King & Dattel, 2005). The battery is the result of a profound and extensive research. The FAA (Ramos, Heil & Manning, 2001) has performed a study to develop criterion measures of air traffic controllers' performance. These criteria were needed to validate a new selection battery AT-SAT. They developed a situational judgment test (Computer Based Performance Measure (CBPM)). Controllers were involved in defining the most important performance examples and categories. Behavior-based rating scales (7 dimensions) were used, with behavioral statements anchoring different effectiveness levels on each dimension. The design method followed by Heil et al. (2001) is basically that of BARS. A total of 73 controllers participated in workshops in which they generate 708 performance examples. Then 24 controllers defined a set of categories and definitions in an iterative design process in five workshops. The next step was to assign the performance examples to one performance category, and to rate the level of effectiveness of each performance example. These examples were compared, classified and selected until a set of unambiguous examples belonging to each dimension and effectiveness level was achieved. A final check was made, and the final set of performance dimensions was established.

4.4 Conclusions

This chapter reviewed the literature on assessment as far as relevant for the design of our assessment system. We argued that developments in learning have led to new perspectives on assessment, such as 'dynamic assessment' and 'diagnostic assessment'. In our view, both forms are needed in assessment of competences in ATC training. It is important to assess the trainee's learning potential to predict future learning, and to find the causes for insufficient performance so that training can be adapted to the trainee's needs. Furthermore, formative assessment should be combined with summative assessment, and criterion-referenced testing is needed to measure the degree of competence attained by a trainee.

We consider assessment in ATC training to be 'performance assessment', referring to the observation of the trainees' behaviors carried out in realistic environments. We consider competence assessment to be a specific type of performance assessment in which competences are assessed. Competence acquisition implies that knowledge, skills and attitudes are merged (see section 2.2), and therefore assessment of smaller pieces is not useful in competence-based assessment. They should be assessed in a realistic context, all aspects of competences should be covered, and assessment of cognitive processes should be emphasized. Various measurement techniques and rating scales are applied in performance assessments. The BARS method appears to be very useful, because the behavioral examples, based on achieved consensus between SME's, make the ratings rather objective. We also discussed the surplus value of automated measurement in simulators, but we concluded that this should be combined with subjective judgments for assisting learning. The psychometric quality, which comprises reliability and validity, is considered a basic requirement of any assessment system. The properties of the aforementioned existing assessment systems in ATC-related environments, as well as applied design methods, are used to design the assessment system in this study (see chapter 7).

Air ATC competence analysis

The first step in a competence-based training design process is a competence analysis. Since competences are considered the acquired capacities to perform tasks (see section 2.1), competence analysis is based on task analysis. While many cognitive task analyses are available for air traffic control, competence analyses have rarely been performed. At Air Traffic Control the Netherlands (LVNL), we developed a model of ATC performance (Oprins & Schuiver-van Blanken, 2003; Oprins, Burggraaff & Van Weerdenburg, 2006a) in order to provide training designers and coaches with insight into the processes underlying ATC performance. This model is based on recently conducted cognitive task analyses and a competence analysis. In this chapter, we pay attention to the competence analysis (section 5.1), we present the model itself and explain its background (section 5.2), and we describe the components of the model in more detail (section 5.3).

5.1 Task and competence analysis

The competence analysis comprised literature research, looking for published cognitive task analyses, and two workshops with air traffic controllers from LVNL.

Cognitive task analyses

In recent decades, cognitive task analyses have become more popular in many domains, including air traffic control (EATCHIP, 1997, 1999, 2002; Hadley, Guttman & Stringer, 1999; Redding & Seamster, 1994; Lamoureux & Cox, 1992). Tasks and subtasks are analyzed in an hierarchically structured manner: tasks are divided into subtasks that are further divided into operations and so on. The result of a cognitive task analysis is usually a model that represents cognitive structures (e.g., mental models and memory organization) and cognitive processes (e.g., problem solving and decision making) that are supposed to underlie expert performance (Redding & Seamster, 1994). Information processing models (e.g., Wickens, 1992; Card, Moral & Newell, 1986) generally form the basis. Cognitive task analysis is considered to be useful in training design, because it provides insight into the cognitive processes underlying learning processes (Redding & Seamster, 1994).

There are many published analyses of the ATC task and they show much agreement. However, it is important to realize that notions such as 'task', 'cognitive process', '(human) performance' and 'competence' are not differentiated sharply in

these analyses. The specific notions used and the filling-in of the models are dependent on the specific purposes of the analyses. For instance, some models are mainly focused on the relations between the cognitive processes involved, while others accentuate 'human factors' that influence performance. We restrict ourselves to the most recent and extensive analyses that were used to construct the ATC Performance Model.

Eurocontrol has summarized the results of different task analyses for en-route control (EATCHIP, 1999). There are five core tasks and one additional task, and each task can be further divided into sub tasks:

- *Maintain situational awareness*: build and maintain a mental 'traffic picture', which has to be continuously projected into the future and checked with the actual traffic situation.
- *Develop and receive sector control plan*: a 'conflict avoidance task'; integrate flight progress strip information and traffic forecast in a future plan, allowing for a safe and efficient traffic flow
- *Make decision for control actions*: closely related to the control plan; make decisions actively, sometimes triggered by pilot's requests
- *Solve conflicts*: provide separations and solve conflicts between aircraft
- *Provide tactical ATM*: accept an aircraft, carry out handovers, provide pilots with relevant information, and provide assistance in abnormal situations (and some other specific activities).
- *Complementary tasks*: briefing and updating of the working knowledge of en-route controllers.

Included in the general information processing models, described in section 2.1 (e.g., Wickens, 1992; Card, Moran & Newell, 1986), are three basic cognitive processes: (1) *situational awareness*, (2) *planning* (2), and (3) *decision making*. Others recognize these three processes for air traffic control as well (e.g., Schneider, 1990; Garland, Stein & Muller, 1999).

However, such a list of distinct tasks and sub tasks does not show the relations with underlying (cognitive) processes. Eurocontrol (EATCHIP, 1997, 1999) made these relations visible with the aid of flowcharts. An overall flowchart represents the following processes: a control process (switching attention), several task processes (taking over position/building up MP, monitoring, managing routine traffic, managing requests/assisting pilots, solving conflicts), and sub task processes (confirming/ updating MP – maintaining Situational Awareness, checking, searching conflicts/checking safety). Each process is further specified and presented in more detailed flowcharts at a lower level, for instance, all specific aspects of the task process 'monitoring' are visualized and related to each other in this flowchart. These flowcharts represent the result of an extensive cognitive task analysis (EATCHIP, 1999).

Another approach is human performance analysis. The HERA project (EATCHIP, 2002) aims at analyzing human errors in ATC, considering the contribution of human error to incidents to be 90% or more. Based on a literature review, the HERA project developed a human information processing model, specified for ATC, in accordance with the model of Wickens (1992). This model shows many similarities

with the findings of EATCHIP (1997, 1999), although it has a slightly different purpose.

Hadley, Guttman and Stringer (1999) developed a performance model with the purpose of designing a 'performance measurement database'. This is a compilation of performance measures and measurement techniques used by researchers in different kinds of human factors studies. The database is useful for our study, since performance measures are also needed in assessment in training situations. The so-called 'ATCS Functional Performance Model' of Hadley, Guttman and Stringer (1999) comprises four general processes: *situation assessment*, *planning*, *decision making* and *implementation*; we recognize the cognitive processes mentioned by EATCHIP (1999) described above. Each part of the model is specified for the domain of ATC. For instance, different tasks belonging to *situation assessment* and the related behaviors are described with reference to ATC. The model is presented in Figure 5.1:

Figure 5.1. ATCS Functional Performance Model (Hadley, Guttman & Stringer, 1999)

We conclude that the many models that have resulted from (cognitive) task analyses show many similarities. This is not surprising, because most models are based on general information processing models (e.g., Wickens, 1992; Card, Moran & Newell, 1986). Although these models are useful to get insight in the tasks and cognitive processes involved in ATC, they are not directly applicable to training. They are not specifically geared to training, they are often too complicated and detailed, and the psychological notions used are not fully understood by the target group of controllers. For training design purposes, additional methods are required with a focus on competences.

Competence analysis

A competence analysis was conducted by LVNL with the aim to identify what makes someone a competent controller. Controllers have become competent by learning, a process in which explicit knowledge has become implicit. Competence analysis is generally considered to make implicit knowledge explicit (e.g., Tillema, 1996); thus, it could be considered the reverse process of learning. The involvement of controllers in the workshop, being subject matter experts (SME's), was considered essential

because their implicit knowledge had to be explicated. Therefore, they were involved in workshops dedicated to competence analysis. The primary aim of the analysis was improvement of selection, but it was obvious that its results could be used in training design as well since there exists a strong relationship between them. Selection has to predict whether a candidate is able to acquire the ATC competences; the same competences are also relevant for training (Roe, 2005).

In the competence workshops (April, 2001), the collaboration tool "Meeting-works" was used. This makes it possible to brainstorm, discuss, structure and prioritize electronically. We consider this tool very useful in the process of competence analysis, because the discussion is well structured, everyone has equal input and people can stay anonymous. With aid of this tool, a group of twelve air traffic controllers (TWR/APP and ACC) formulated thirteen competences in two sessions. They ordered these competences from most to least relevant. The result is presented in Table 5.1 (translated into English):

Table 5.1. *Set of competences defined in LVNL workshops (2001)*

LVNL set of competences
Situational Awareness
Decisiveness
Acting in unexpected situations
Workload management
Conflict solving
Multitasking
Prioritizing
Coordination and communication
Planning (with flexibility)
Leadership
Teamwork ability
Perseverance
Critical attitude

Next, each competence was operationalized by a set of 8 to 12 behavioral markers. Table 5.2 presents an example for the competence *planning* (translated into English):

Table 5.2. *Example of the competence planning and its behavioral markers*

Planning
Is able to make a plan, execute the plan and adapt the plan to (changed) circumstances
Is able to make a collective plan in collaboration with a colleague, while making compromises
Is able to revert to standard procedures when necessary
Is able to adapt his/her own plan to requirements and wishes of others
Is able to work concentrated on his routine, but is able to interrupt the process at once, following another plan
Is able to propose alternatives
Is able to order a chaos
Is able to deal with procedures in a flexible way
Is able to present the traffic situation schematically in order to realize an orderly flow of traffic

The first requirement is usability. The competences should be recognizable by the user group for their effective use within the organization. The user group consists of controllers who have also the role of coach or assessor. We may conclude that this requirement was completely met due to the method applied. In our view, the main value of this set of competences is that they were formulated in the jargon of the controllers themselves, since they were the sole participants of the workshops. The method enabled controllers to come to agreement about the comprehensiveness of the set of competences and about the interpretation of each competence. They were forced to think about their own work performance at a more abstract level. It seems that they have really explicated their implicit knowledge and verbalized what makes someone a competent controller.

The second requirement relates to the view on competences discussed in section 2.1. We argued that competences should be formulated at the higher level of integration of knowledge, skills and attitudes. We emphasized the context-dependency and learnability of competences (cf. Figure 2.1). It appears that not all of the competences formulated in the workshops agreed with this view. Among them were *perseverance* and *leadership*. These were in fact personality features relevant in any profession. They are not directly task-related and belong more to selection requirements than to competences.

The third requirement involves the validity of the competence analysis. The set of competences has to cover the job of air traffic controller as performed at Schiphol Airport, and it should be true, complete and unambiguous. Although the set of competences appeared to be rather complete, some important characteristics of the air traffic control job were missing. The competences were formulated at a highly abstract level. As a result, the most crucial and task-related aspects of air traffic control, to attain *safety* and *efficiency* were not included as separate competences, even though they were present as behavioral markers to a certain extent. *Strip and label management* were missing completely. A related problem is that the competences were formulated at varying levels of abstraction. This also applies to the behavioral markers. For instance, some behavioral markers were used for others such as the 9th aspect of *planning*, which refers to the competence *situational awareness*, while the 2nd aspect is strongly related to *teamwork ability*. This lack of differentiation between behavioral markers produces a considerable interdependency and ambiguity of the competences. In this form the competences and their behavioral markers could not be used directly, for instance, as performance criteria in assessment. Performance criteria should be complete and unambiguous, and specific behavior may not be ascribed to more criteria.

The fourth requirement is that insight into the relations between the cognitive processes should be provided, because this is relevant in training. This requirement was not completely met, because these relations cannot be made visible with just a list of thirteen competences. For instance, there exist relations between *situational awareness*, *strip management* and possible consequences for a safe solution of conflicts, which should be addressed in training. In addition, different aspects of performance require different training and assessment methods. We refer to the distinction between cognitive learning, psychomotoric learning and affective learning (Romiszowski, 1988; Lowijck & Verloop, 1995).

Operationalization of competences

Apparently, the competence analysis performed by LVNL has its limitations when it comes to direct use in practical training. The input of controllers only could lead to incompleteness or mistakes, because air traffic control in the Netherlands is not unique. Therefore, additional research was needed to operationalize the competences. This resulted in the ATC Performance Model.

In the process of operationalization we compared the list of competences with the different published cognitive task analyses looking for additional aspects of ATC performance that might have been forgotten in the workshops. We were also interested in the relations between these aspects when categorizing the 13 competences. In addition, we restructured the competences and their behavioral markers to reduce their ambiguity. On the basis of this additional research we developed the ATC Performance Model (Oprins & Schuver-van Blanken, 2003; Oprins, Burggraaff & Van Weerdenburg, 2006a). This model represents the several aspects of performance and their interrelations, and is meant to serve as a framework for training and assessment design at LVNL (this will be further explained in chapter 7).

5.2 The ATC Performance Model

In section 2.1 we described the several aspects of performance and influences on performance in relation to learning and competences. We present a model that forms a basis for the ATC Performance Model (see Figure 5.2) and summarize its background as discussed in section 2.1:

Figure 5.2. Performance and the factors involved

The definition of performance (Roe, 1999) states that it is divided into *process* and *outcome*. In Figure 5.2, the *process* is sub divided into *actions* and *information processing* under the assumption that information processing guides our actions. Performance is supposed to imply some goal-directed activity, which relates to the *task* (Hacker, 1998). We should notice that a distinction exists between the 'objective task' (task-as-given) and 'subjective task' (task-as-taken), although this is not pre-

sent in Figure 5.2. We argued that several kinds of regulation processes are responsible for *influencing factors* on both the process and outcome of performance (e.g., Hockey, 1997). The ATC Performance Model is an elaboration of the components of performance (*actions, information processing, outcome*) with regard to ATC, completed with some *influencing factors*. We specified the components of the model on the basis of existing cognitive task analyses and performance models, combined with the LVNL competence analysis. The specific formulation of competences in the controllers' jargon was kept as much as possible. The ATC Performance Model is presented in Figure 5.3.

The ATC Performance Model (Oprins & Schuwer-van Blanken, 2003; Oprins, Burggraaff & Van Weerdenburg, 2006a) shows the dominant role of information processing in ATC work. All components of the model are specified in terms of specific competences. We recognize the majority of the competences as defined in the workshops (see Table 5.1) in the dark gray parts, others in the white parts. Some were revised and others added on the basis of the literature.

However, it should be noticed that the *task*, often present in a performance model (Roe, 1999), is not included in Figure 5.3. The task refers to the specific situation to be handled by the controller, such as a combination of inbound and outbound traffic, various conflict situations, or an emergency situation. The reason for the absence of these situational factors lies in the specific application of the ATC Performance Model: the formulation of performance criteria. As we will explain in chapter 7, the performance criteria to be rated by the assessor are independent of these task variables. Characteristics of the assessment task are included in the performance standards, for instance, a certain degree of traffic intensity.

In the model, information processing comprises *situation assessment, planning* and *decision making*. This distinction is also found in other ATC performance models (e.g., Hadley, Guttman & Stringer, 1999; EATCHIP, 1999). These cognitive processes are not necessarily ATC-specific in contrast with the actions and outcome, because they are applied in any task execution. They were derived from general information processing models. Situation assessment is supposed to be the process that results in the product *situational awareness* (SA). Endsley (1995) considers SA to be the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future. Based on this theory, situation assessment in our model is subdivided into *perception, attention management* and *interpretation* (mental picture). The actions consist of *communication, co-ordination, label and strip management, and equipment operation*. The outcome comprises *safety* and *efficiency*. Finally, *workload management* and *teamwork ability* mainly represent the influencing factors.

Figure 5.3. The ATC Performance Model

The ATC Performance Model helps to differentiate between components that can be automated ('consistent', 'stable' or 'recurrent' skills) and higher-order skills that are applied at higher levels of control ('non-consistent', 'unstable', 'non-recurrent' skills) as discussed in section 2.3. The *actions* and the *outcome* (underlying procedures) are components that can be automated in contrast with the components *information processing* and *influencing factors* that are more dependent on personal characteristics (abilities, personality). We could say that the latter components are possible causes for insufficient *actions* and *outcome*, being often the reasons for failing. Furthermore, the *outcome* and *actions* can be measured objectively since they are observable, while *information processing* and *influencing factors*, being not directly observable, should be derived from behaviors.

5.3 Description of the components of the ATC Performance Model

This section explains the meaning of each component of the ATC Performance Model. It is based on the findings of the literature review presented in chapters 2 and 3, but this section only describes how it is used in practice at LVNL. The model is applied for each air traffic controller function available at LVNL by specifying the components for each function: Ground Control (GND), Arrival and Approach Control (ARR, APP), Tower Control (TWR) and Area Control (ACC)⁵.

Outcome (safety & efficiency)

We stated in section 3.1 that the main purpose of air traffic control is to expedite and maintain a safe and orderly flow of air traffic. In our model, *safety* implies that collisions are prevented, for example, by maintaining the separation minima. Specific regulations and procedures differ for each ATC position (GND, ARR, APP, TWR and ACC). Additionally, safety is maintained by dealing with emergencies correctly. *Efficiency* means that flights do not have unnecessary delay by handling as much traffic as possible within a certain time period. Possibilities are to select a shorter (alternative) route, to avoid double separation and to select exact headings, speeds or levels.

There exist a third outcome of air traffic control. This is care for the 'environment' according to the concept 'VEM', the Dutch abbreviation for safety, efficiency and environment. This concept was developed within LVNL for designing the ATM system (procedures, regulations) at the organization level. We did not explicitly include 'environment' in the outcome, next to *safety* and *efficiency*, because the procedures involved are incorporated in general regulations for air traffic management. It does not play a distinct role in ATC training and assessment in contrast with the others.

Actions

An important part of the controller's job is *communication*. Controllers communicate to provide pilots with instructions (e.g., changing level, speed or heading) or traffic information (e.g., runway, gate, weather). They must use correct radiotelephony (RT), the spoken message must be clear and the pilot's messages must be heard and interpreted correctly (e.g., requests, readbacks). The communication with adjacent centers and colleague-controllers is called *co-ordination*, for instance, by making appointments about traffic flows or by handing over the traffic. Additionally, a controller is constantly updating the flight information on strips or labels, which is called *label and strip management*. Finally, *equipment operation* is needed to perform all these actions.

⁵ A more elaborated description of the organization at LVNL is presented in Appendix A.

Information processing

As mentioned in section 5.2, we divided information processing into *situation assessment*, *planning*, and *decision making*.

Situation assessment

The first process that belongs to situation assessment is *perception*. Controllers use effective monitoring and scanning techniques to obtain the required traffic information, such as positions of aircraft but also what happens in the direct environment (e.g., colleague controllers). They register the information that they see and hear in their memory. Information is available on radar screens (labels, aircraft positions), additional information screens, or flight strips. Controllers at the tower are looking outside. Controllers apply identification procedures to verify the strip or label information with the specific aircraft.

Second, continuous *attention management* is needed to get the information from the different visual and auditive sources. Therefore, multitasking is required. This implies that controllers perform many activities simultaneously, such as listening to pilot's requests or readbacks, being alert on remarks from colleagues, issuing instructions by using the correct RT, monitoring the air traffic situation on the radar screen or outside, and updating strips or labels. They continuously have to keep the several pieces of information in their memory.

Third, controllers do not just register what they see and hear, but an *interpretation* of the information is needed. They combine the various pieces of information into a three-dimensional mental picture of the air traffic situation. Controllers anticipate on the future traffic situation by making predictions, based on consequences of their instructions and their experience with aircraft behavior (e.g., speeds of climbing and descending of specific aircraft in combination with weather circumstances). Finally, controllers are continuously checking whether the traffic situation is developing as expected.

Planning

Based on their mental picture, controllers make a *planning* how to manage air traffic in a systematic way. Anticipation is a precondition to make an appropriate planning. This implies that 'ad hoc' working should be avoided, because this could result in getting behind. Flexibility is needed because situations may develop differently from what is expected. If this happens, controllers should be able to deviate from their original plan and to switch to an alternative plan immediately. Also controllers should be sufficiently creative in inventing solutions (problem solving) and in having several alternatives available. Controllers have standard solutions for specific situations, but they should be able to invent real novel solutions for situations with which they have never coped as well. It also depends on the controller's own workload, and this refers to the notion of strategies (see sections 2.1 and 3.1). We should notice that we did not include strategies in the model explicitly, because it involves various cognitive processes such as *situation assessment*, *workload management*, and *planning*.

Decision making

Decision making forms the direct step to undertake concrete actions. A high tempo in decision making is needed, due to the dynamically changing traffic situations. Controllers are not allowed to hesitate much, because this will make them to respond too late. They must take initiative and act very pro-actively. If several actions must be taken, controllers must prioritize them and indicate which action should be executed first, for instance, issuing specific instructions to different aircraft. Strict temporality is required. This implies that the right action has to be performed at the right moment: neither too late, not too soon.

Influencing factors

The category of *influencing factors* contains everything that might influence the controllers' information processing, his actions or the outcome of his actions.

Workload management relates to the controller's psychophysiological state that might affect the regulation of information processing and activity; it includes stress, fatigue and self-efficacy (see sections 2.1 and 3.1). Consequences of an unfavorable state are that controllers cannot think logically anymore, that they have a narrowed field of attention ('tunnelvision'), or that they switch to automated control of information processing (routine or basic actions). Here, work tempo plays an essential role.

Teamwork ability is relevant because controllers work in teams. Team resource management (TRM) is a central notion in this context. Controllers must collaborate continuously with colleague-controllers and adjacent centers. A keyword in the controller's behavior is *leadership*: a controller has to keep control over the situation as a leader, for instance, by only accepting instructions that are sufficiently convenient for himself.

5.4 Conclusions

This chapter described the development of the ATC Performance Model as a general framework for the design of the training and assessment system at LVNL. We have seen that the model is largely based on a competence analysis internally performed at LVNL in combination with a literature review (i.e. ATC cognitive task analyses and performance models). Controllers were involved as subject matter experts to explicate their implicit knowledge in order to get insight into the competences to be acquired in ATC training. The main purpose of this method was to achieve recognizability and common understandings of the set of competences that are formulated in the controllers' own jargon, needed for practical use in training. A set of behavioral markers, expressed in observable behaviors, represents each competence.

6 Analysis of the design problem: Program of Requirements

The first step in the design cycle of Roozenburg & Eekels (1991) is 'analysis' of the problem to be solved with a Program of Requirements (PoR) as the final result. Chapter 1 described the general aim at the organizational level of LVNL: the redesign of the assessment system had to contribute to the improvement of efficiency and effectiveness of training. This chapter discusses the analysis of the problem emerging from the previous assessment system. We analyzed the purposes, the requirements and constraints of the assessment system to be designed. We considered the input from potential users to be critical. They belong to the group of 'stakeholders' who have interest in the redesign project. Section 6.1 discusses the methods applied in the process of analysis, section 6.2 characterizes the design problem, and section 6.3 presents the final PoR.

6.1 Methods

As a researcher and designer working at the training department of LVNL, we were assigned to redesign the assessment system from an educational-psychological point-of-view. Since this system should be implemented in practice, a close collaboration with operational personnel was needed. Being at work in the organization itself resulted in a deep involvement in the process. This relates to the method of 'participant observation' (Van Dijk, De Goede, Hart & Teunissen, 1999). In the role of participant within the training department, problems were recognized which could be ascribed to deficiencies of the previous assessment system. Participation is very useful to get a profound insight into the interests and needs of the users (i.e., coaches, trainees, training managers), but brings along informal contacts and unstructured research methods. At certain moments, some distance is needed to make the design process more objective and transparent. The development of a Program of Requirements can help to achieve this purpose.

Three main methods were applied: observations, interviews and document analysis. They were used in a rather unstructured way without a clear and chronological order of steps, due to the practical setting of the design process. We should notice that the PoR pertains to Unit Training as a whole, not only area control (ACC), although this is the main case in this study (see chapter 1).

Observations

Structured observations in simulator training of Unit Training were conducted for several purposes: (1) to get insight into the practical setting of air traffic control (ATC) training at LVNL; (2) to familiarize with the content of ATC training, and (3) to identify possible problems of training and assessment more specifically. Two coaches train two trainees at a time (one coach per trainee). Besides training sessions, we observed simulator tests of trainees in various courses of Unit Training. During the observations we made notes. We wrote down the comments of trainees and coaches as literally as possible. After the observation sessions we categorized them according to the competences defined by LVNL (see chapter 5) in order to clarify the meaning of the competences in the context of different courses and to get a feeling of the importance of each competence. At a later stage these comments were also used to define the exemplary behaviors in the renewed assessment system (see chapter 7). Finally, we summarized the main findings in internal reports. These reports also contain recommendations for improvement of training and assessment design. They were discussed with the controllers involved, especially with the course and training managers who are responsible for the courses in Unit Training.

In addition, there were observations in the simulator that were not structurally planned or summarized in a report. Among them were observations of controllers at work in the real working environment. Observations of OJT sessions (coaching) were not allowed because of safety reasons.

Interviews

Trainees can give important information about their expectations, needs and interests in the improvement of the assessment system. Therefore, each course is concluded with a general evaluation by means of a questionnaire and an individual interview with the trainee (see also chapter 12), not only for this research but also as a general evaluation of Unit Training. Assessment is one topic in this evaluation. The interviews are semi-structured by using a checklist as a guideline. The most relevant issues concerning assessment are: the trainee's own assessment results, objectivity of assessors, usability and clarity of assessment reports, clearness of performance criteria and performance standards, and representativeness of assessment tasks. The findings are summarized in anonymous evaluation reports twice a year.

The assessment problem was also discussed in several formal meetings with course managers, training managers and coaches. Additionally, we held semi-structured individual interviews and informal conversations with some stakeholders. We considered these informal moments as very important in the design process as they allow explicating expectations, problems or wishes, without any pressure of formal settings and procedures. The strong involvement of the user group (controllers) can contribute to a higher chance of successful implementation of the new assessment system.

Document analysis

A crucial step in the process was the analysis of assessment reports filled in by different assessors during the training. The purposes were: (1) analyzing the properties of the actual assessment system (i.e., procedures, performance criteria, assessment reports); (2) analyzing behaviors that are often mentioned and that are crucial for ATC performance, and (3) analyzing the main reasons for failure. For the last two purposes we analyzed the assessment reports from 15 trainees who failed in several courses (both simulator training and OJT) in more detail. We classified the comments made by the assessors according to the set of competences in a way comparable with the observations described above, and used them to define the exemplary behaviors as well (see chapter 7). We continued this qualitative analysis when the redesign project was already finished, because it keeps providing useful information about the main reasons for failing (see chapter 11). Unfortunately, we were not able to analyze the assessment reports quantitatively, because the training results of the previous assessment system were not digitally available.

6.2 The redesign problem

Since redesign of the assessment system was the general aim of the project, we analyzed the main shortcomings of the previous system, which resulted in requirements on the new system.

Assessment methods

Continuous assessment in combination with performance tests was applied, and there was no reason to change this. However, the system was not consistent over all courses (GND, ARR, APP, TWR, ACC) and training phases (preOJT, OJT). We consider this to be undesirable from the point-of-view of transparency of procedures and comparability of assessment results. Course managers had designed their own assessment system for their courses. An overall view on assessment from an educational-scientific perspective did not exist. Validation and reliability analyses were never done. Validation of the selection system, based on the training results, was difficult. Training results were not digitally available but only stored in paper files. Overviews and summaries of trainees' performance could not be made easily. This also caused problems for providing adequate feedback.

Performance criteria

Rephrasing the performance criteria was desirable for several reasons. First, the performance criteria were not expressed in observable behavior, but only defined as general notions. This resulted in vagueness about the meaning of the performance criteria and in low agreement between assessors.

Second, the performance criteria and their categorization were not consistent over the training on a whole since course designers had designed their own as-

assessment systems. It was unclear why some criteria were mentioned at an assessment report in one course and were lacking from another course. Also, the performance criteria were not always classified in a logical way. For example, three categories were very generic and their composition was unclear: the category *general* included *tempo* and *attention management* while the second category, *traffic handling*, contained almost all other relevant performance criteria such as *efficiency* and *conflict solving*. The third category, *safety*, was a separate category, but it was unclear why *conflict solving* did not belong to this category. This vagueness hampered assessors in making univocal and reliable judgments.

Third, the set of performance criteria was insufficient to indicate trainee's strengths and weaknesses in an appropriate way. Assessors found it extremely difficult to judge objectively why a trainee did perform (in)sufficiently. They mainly relied on their 'gut feeling' and were not able to provide concrete evidence on the basis of the existing set of performance criteria. This gut feeling did make their assessments rather subjective. For the same reason, assessors were not able to predict future performance adequately, while this is needed for pass-fail decisions. Assessors wanted to be certain that trainees would work safely, also in situations they have never coped with. When they did not have this feeling, they could not easily express their doubts. Consequently, many trainees without prospects had been held too long in the OJT (false positives), making the number of failures in final OJT phases too high. On the other hand, it was quite likely that trainees failed undeservedly (false negatives). However, we could not prove this assumption because we would never know if they had passed finally after a longer training period. In addition, the reasons for failing could not clearly be analyzed while they are relevant for profound argumentation of pass/fail decisions and for the validation of selection.

Finally, the performance criteria did not directly relate to training objectives, or training objectives were not defined at all. Performance standards required in intermediate training phases, especially in OJT, did not exist yet. The standards in simulator training were mainly determined by task complexity implied in simulator exercises. Training objectives were formulated at the lower task level (e.g., the amount of inbound traffic to be handled), but they did not have a clear relationship with the performance criteria, usually formulated at a higher level (e.g., planning). The OJT was not divided into phases at all. This resulted in disagreement between assessors about the requirements on trainee's performance during training. The assessors, who usually prefer different working styles and strategies, could only rely on their personal experience and ideas. For trainees it was not clear what was expected from them in a specific training phase, and which competences they had to develop further.

Measurement techniques

Many different measurement techniques were applied in the various courses. In the majority of simulator courses only plain text was used (no ratings), while in the OJT a four points rating scale was applied. It was not clear why these differences existed. It would be preferable to use the same rating scale within one course, from the start of simulator training till the end of OJT, in order to be able to compare assessment re-

sults obtained over a training period and to be better able to follow trainee's progress over time.

Furthermore, the ratings in the OJT were not sufficiently reliable. Assessors must fill in an assessment report everyday while the situations occurring may not be representative enough. Assessors did not always get a complete picture about trainee's performance based on a single day. In contrast, the plain text used in simulator training, without ratings, did not provide enough information about the trainee's performance. Assessors tended to avoid too much negative feedback in their written comments. Consequently, trainees who failed did not always know that they performed below the required learning curve. An additional problem was that assessors were hardly trained in their role as an assessor. They followed a coach course to become a coach, but refreshing training was scarce.

In simulator tests detailed scoring techniques were used. The final score was calculated in a percentage with a norm of 70%. Weights were assigned to different performance criteria by assigning a different maximum score to each criterion. Assessors could fill in any score as long as this maximum was not exceeded. For instance, the maximum score of planning was 25 points; assessors could assign 13, 17, 21 points or whatever they wanted with a maximum of 25. The maximum score of safety was 100 points because of its higher weight, and the assessor could assign the score 0 – 100 (100 possible options in sum). However, in our opinion, this scoring technique affects the reliability of the test, because there exist too many options for assessors to assign. It is impossible that an assessor can distinguish, for example, 13 from 14 points for planning, only by observing a trainee's overt behavior. Reliability analysis was not executed in the design process, for instance, for assigning weights to performance criteria.

Summary of shortcomings

The deficiencies of the previous assessment system can be reduced to two main shortcomings. First, the assessment system did not sufficiently contribute to learning processes and to profound pass-fail decisions, while these are important requirements for the effectiveness and efficiency of training as we mentioned before. Assessors mainly relied on their 'gut feeling' in their judgment of (in)sufficient performance, and trainees did not receive clear feedback about their strengths and weaknesses during training. This problem was probably caused by the lack of clear performance standards for different training phases, vagueness in the classification and meaning of performance criteria, and unreliable measurement techniques. Consequently, agreement between assessors was low in general.

Second, each course had its own procedures, performance criteria and measurement methods. This inconsistency made it difficult to follow trainee's progression over subsequent courses and training phases, and to discover trends in trainee performance such as differences in learning curves and main reasons for failing. The fact that training results were not digitally available added to this problem. Overviews of trainee results could not easily be made, and reliability and validity analyses were not possible. Additionally, inconsistency has disadvantages for practical use because

assessors and trainees should be familiarized with the system again in each course or training phase.

The idea was that a more systematic and comprehensive approach in the re-design process, from an educational-psychological perspective, would result in a more adequate assessment system with greater reliability and validity. The construction of the PoR was a first step into this direction.

6.3 Program of Requirements

The Program of Requirements (PoR) was based on the outcomes of the analysis described in section 6.2, and on general requirements for assessment found in the literature (see chapter 4). The first version of the PoR was made by ourselves. Next, it was commented upon by a group of stakeholders. This group consisted of two training managers who are operational controllers with the responsibility for Unit Training (TWR/APP and ACC; see Appendix A), the general manager of the training department, and his management team (5 persons; not controllers). Next, we adjusted the PoR based on their comments. The same group of stakeholders approved the final version.

A primary function of the PoR is to give direction in which solutions for the assessment design must be sought. A second function is to provide specific guidance to the design process. Some requirements and constraints were literally taken from the PoR drafted for the selection system to be redesigned at LVNL (see for the format and examples: Roe, 2005). It should be kept in mind that the selection system was redesigned simultaneously, and that the main purposes of both redesign projects were the same: a reduction of the failure rate in training. However, the projects were independent of each other and could hardly have influenced each other. Table 6.1 presents the final PoR. Weights were not assigned to the different requirements (R) and constraints (C) in the PoR; the order was chosen arbitrarily.

A few aspects from the Program of Requirements deserve to be emphasized. The final purpose of the design of the assessment system is presented by R1, R2 and R18: increasing the outcome of qualified (competent) air traffic controllers from training. Many requirements relate to learning processes and pass-fail decisions in training (R4, R5, R6, R9, R14, R16). We argued that the redesigned assessment system should improve the feedback towards trainees and the argumentation of (in)sufficient performance, because these were essential shortcomings of the previous assessment system. A special, new requirement in this context is R3: the assessment system should be competence-based. This implies an important change of the existing assessment system. R17 indicates that selection and training should be one integrated system, which was also new for LVNL. Furthermore, more consistency and clarity are needed (R7, R8, R15) which was a problem in the previous assessment system as well. In addition, psychometric requirements and constraints are included in the PoR (R10, R12, C2, C3, C10). We discussed that reliability and validity could not be investigated with the previous assessment system; these analyses will be an essential part of the design process of the new assessment system (see

chapter 9). Finally, many constraints relate to the practical usability and implementation of the assessment system for operational controllers (R13, C1, C4, C5, C6, C7, C8). In this respect, R11 indicates that the method of continuous assessment should be maintained in order to minimize redundant changes of the existing assessment system.

Table 6.1. *Program of Requirements for the LVNL assessment system to be designed*

Requirements

R1 Quality of controllers

- a. The assessment system has to contribute to the deliverance of sufficiently qualified trainees from preOJT to OJT.
- b. The assessment system has to contribute to the deliverance of sufficiently qualified trainees from OJT to OPS (operational air traffic controllers).

R2 Pass rate

- a. The assessment system in the preOJT aims at the highest possible pass rate in the OJT (90%)
- b. The assessment system in the first modules aims at the highest possible pass rate in the following modules.

R3 Competences

- a. The assessment system has to contribute to sufficiently competent controllers as outcome from training.
- b. The assessment system has to measure to which extent trainees have acquired the competences required in a certain training phase.
- c. The assessment system has to be based on the competences that are formulated by controllers in the competence workshops (April, 2001); competences need to be recognized by them.

R4 Learning processes

- a. The assessment system has to assess what has acquired at the end of training and in intermediate phases during training (represented by training objectives).
- b. The assessment system has to support the learning process by contributing to possible training interventions that may follow from the assessments.
- c. The assessment system has to indicate the trainee's position at the optimal learning curve in order to make clear how much the trainee still has to learn before the end of the training period.
- d. The assessment system should assess progression besides the required level of performance in order to determine whether a trainee is still learning or whether he has reached a 'learning plateau'.

R5 Pass/fail decisions

- a. Pass/fail decisions have to be built in early training phases (e.g., preOJT instead of OJT), although the primary aim of the assessment system is not selection. These decisions are needed to avoid useless effort and costs in further training of prospectless trainees.
- b. Possibilities and constraints for failed trainees have to be clear, including the procedures to be applied in pass/fail decisions.

R6 Following trainee's progress

- a. The assessment system has to facilitate in following trainee's progress from selection and Initial Training towards Unit Training. This makes it possible to adapt coaching and training design to individual needs.
- b. Assessment results have to be stored in a database to be able to follow trainee's progress.

R7 Clarity

The assessment system has to contribute to clearness in assessments. This implies that the progress and performance of trainees can be described more clearly than before (for trainees, managers, coaches).

R8 Consistency

The assessment system has to be designed in a consistent way for all parts in ATC training, so that users (trainees, assessors) become easily used to a comparable system (e.g., performance criteria, procedures, forms).

R9 Level of performance (standards)

- a. The required levels of performance (standards) in each training phase have to be described as uniformly as possible, so that it is clear which level of performance has to be achieved.
- b. Besides the required level of performance, the required amount of progression in a training phase has to be determined. This is necessary to analyze to which extent the job of air traffic controller is 'learnable' for someone.

R10 Objectivity

- a. The assessment system has to measure trainee's performance as objectively as possible.
- b. The subjective 'gut feeling' may never be neglected, however, it may not have as much impact as in the previous assessment system in final judgments about trainee's performance.
- c. When subjectivity is inevitable, the subjective 'gut feeling' has to be well founded with meaningful arguments and evidence for the trainee and others involved.
- d. Multiple assessors have to be involved in assessments.

R11 Continuous assessment

The assessment system has to fit in the principles of 'continuous assessment', which has to be maintained within LVNL.

R12 External influences to be eliminated

- a. The assessment system has to contribute to the elimination of personal preferences in working styles (strategies) of assessors / coaches.
- b. The assessment system has to contribute to the elimination of external influences on trainee's performance during assessments (e.g., private circumstances, working environment).

R13 Culture

The assessment system has to avoid assessment of assimilation to the (changing) culture of operational departments.

R14 Training redesign

- a. If training is redesigned, the assessment system can be related to it (e.g., new training objectives, phases).
- b. Flexibility in changing (parts of) the assessment system is required with regards to redesign of training.

R15 Optimal construction

The various parts of the assessment system have to be in optimal harmony with each other with respect to purposes of each part (what has to be measured), assessment methods, role of assessor/coach, weights in pass/fail decisions etc.

R16 Role of assessor / coach

- a. The role of assessor / coach has to be clearly defined for various training and assessment situations, including the combined role of assessor and coach (e.g., during training situations).
- b. Assessors have to be trained in getting familiarized with the new assessment system and general assessment principles.

R17 Integration of the selection and training system

- a. The selection system has to be integrated with the training system, which includes the assessment system. The total success rate of training has to be determined based on selection, Initial Training and Unit Training.
- b. Continuous feedback from training to selection and in reverse is required in order to be able to improve the selection and training system (e.g., reasons for failing, qualities of 'aces').

R18 Contribution to personnel shortage

- a. The assessment system has to contribute to deliver a sufficient number of qualified controllers to OPS.
- b. The assessment system has to contribute to optimal efficiency (lower costs) in personnel management of OPS

Constraints**C1 Consequences of redesign**

- a. The assessment system may not be changed too much, because then it will not be used in practice.
- b. The redesign of the assessment system has to be implemented gradually in order to increase the chance on success.
- c. Intensive communication with controllers is required about changes in the assessment system.

C2 Reliability

The assessment system has to be reliable (e.g., interrater reliability, internal consistency)

C3 Validity

The assessment system has to be valid (e.g., predictive validity, construct validity).

C4 Practical usability

The products of the assessment system (e.g., reports) have to be practically usable for assessors, trainees and managers.

C5 Transparency

The assessment system and its various parts have to be transparent to trainees and assessors. Purposes and methods of each part have to be clear.

C6 Acceptance

The new assessment system has to be acceptable to assessors and trainees.

C7 Confidentiality

The assessment system has to guarantee the confidentiality of personal data.

C8 Input of controllers

Sufficient input of controllers is required in the design of the assessment system.

The assessment system has to be developed by intensive collaboration with controllers.

C9 Benchmarking

Benchmarking on assessment systems is required. Methods and techniques have to fit in actual scientific and technological knowledge (e.g., related domains, other ATC organizations).

C10 Continuous evaluation

The assessment system has to be evaluated continuously.

C11 Quality assurance

Quality assurance has to be maintained by (educational) psychologists and OPS (operational air traffic controllers).

6.4 Conclusions

The first step in the redesign process was to construct a Program of Requirements with the most essential requirements and constraints of the assessment system to be designed. Therefore, we analyzed the main shortcomings with the previous assessment system in collaboration with the user group and stakeholders, using three general methods: *observations*, *interviews* and *document analysis*. The two main shortcomings of the previous system were: (1) inadequate contribution to learning processes and profound pass-fail decisions, which are required to improve the effectiveness and efficiency of training; and (2) inconsistency in the assessment system as a whole and the lack of digitally available training results, which are also needed for reliability and validity analyses. These issues were included as requirements and constraints of the new assessment system in the final PoR. Additional constraints that are essential emphasize the practical usability for operational controllers. The PoR was approved by a group of stakeholders. It has guided the most important decisions in the design process, discussed in chapter 7.

Assistant, ground and tower controllers at work in the tower at Schiphol Airport

Tower controller at work in the tower at Schiphol Airport.

PART VI DESIGN & IMPLEMENTATION

The phase of *design* refers to the activities that are needed to produce the final version of the assessment system designed for LVNL in the redesign project. It was alternated by *implementation* of parts of the system. Chapter 7 discusses how we designed this system and it describes its main properties. Chapter 8 explains how we implemented the system.

7 Redesign of the assessment system

This chapter discusses the redesign of the assessment system in terms of the design process and product properties⁶. The emphasis lies on Area Control (ACC) as main case to exemplify the assessment design for Unit Training. We put the symbols and numbers of the various requirements (R) and constraints (C) in the Program of Requirements (as presented in Table 6.1) in the right margin in order to explain why we designed the assessment system in this way, for instance:

First, we describe the general characteristics of the actual design process in section 7.1. Next, we explain the main principles of competence-based assessment in section 7.2, and its relationship with training in section 7.3. A more elaborated description of the properties of the assessment system is given in section 7.4. In this way, the description of the assessment system goes from 'architectural design' to 'detailed design'. Finally, section 7.5 describes the design process and properties of a web-based assessment tool as part of the assessment design.

7.1 The design process

The 'synthesis' phase in the design cycle is considered to be the most unstructured and creative step (Roe, 1989; Roozenburg & Eekels, 1991). During this step, the designer invents and proposes certain solutions to the design problem. When designing a complex system that comprises a number of subsystems or components, this may happen a number of times or simultaneously. This was also the case for the design process at LVNL. For instance, findings from one course were incorporated in another course by improving first versions and so on. There did not exist a logical chronological order of steps and intermediate products. We emphasized interaction with the user group in accordance with training design methodology (Moonen, 2000).

The design process, especially this step, has started in April 2001 for Area Control (ACC) but it is still an ongoing process. Improvements have been implemented in the assessment system until now (end of 2007). By September 2003 the whole Unit Training (preOJT and OJT) was redesigned. Design decisions were

⁶ Parts of this chapter are copied from Oprins, Burggraaff & Van Weerdenburg (2005, 2006a).

guided by the results of the step of 'analysis' in the design cycle, that is, the Program of Requirements (chapter 6) in combination with literature research (chapters 2, 3 and 4).

The design process implied the use of several methods, being typical for practical design and research within organizations. Air traffic controllers were strongly involved in the process since they form the end user group (managers, coaches, trainees). We found this very important to be able to make the system practically usable and acceptable. As for the construction of this PoR, we organized meetings, observations and interviews. These were other activities than described for the construction of the PoR (chapter 6) because the PoR was finished earlier.

C4

C6

C8

For each course a design team was formed that consisted of a few coaches, a course manager and one or two training experts. Brainstorming and discussion about ideas and design documents, prepared by the training experts, were the main activities in the meetings. Also observations in simulator training and (informal) conversations with coaches and trainees contributed to the redesign process. The majority of these observations and conversations were not well-structured but were part of regular work activities executed by the training department. Practical constraints and possibilities were investigated and were an important input for the aforementioned meetings. Furthermore, we used the findings from the literature on assessment systems in related domains such as aircrew, and on work samples as part of ATC selection systems (see section 4.3). We also examined existing assessment systems in use by other ANSP's and training academies (DFS, NATS, FAA, SATSA). Other design methods that are specifically related to certain parts of the assessment system are elaborated upon in the next sections.

C9

7.2 Competence-based assessment

We argued that the two main purposes of assessment in training are: (1) to support learning processes by adequate feedback, and (2) to make well-founded and valid pass-fail decisions. This should make training more effective and efficient and should result in a higher output from training. A fundamental choice in the design was to make the assessment system *competence-based* in accordance with developments in views on learning and assessment as described in chapters 2 and 4. We describe our view on competence-based assessment and we explain how this has guided our decisions for the specific properties of the system.

R1a

R1b

Focus on competences

We consider competence-based assessment to be a specific type of performance assessment (Berk, 1986; see section 4.2). The emphasis is put on the assessment of competences, as derived from the competence analysis that resulted in the ATC Performance Model (see chapter 5). Following our definition of competences presented in section 2.2, this implies that they involve the integration of knowledge, skills and attitudes during task execution. Competences optimally reflect the individual capacities to perform effectively while controlling for situational factors that may

R3c

influence performance. In our view, this is important for learning processes: assessment of competences is a base for adequate feedback, needed to improve individual performance.

Due to the generic character of competences, assessment takes place at the higher-order level (Pellegrino, Chudowski & Glaser, 2001). We chose for this, because assessment of small, isolated (sub) skills does not provide enough information about trainee's overall performance and his/her strengths and weaknesses in various task situations. Knowledge, skills and attitudes are merged in realistic learning environments such as simulators and OJT. The complexity of task situations is very high. This makes it difficult and useless to assess smaller subskills. We explained that underlying personal characteristics are needed to acquire competences (see Figure 2.1; Roe, 2002), and that many individual differences in learning exist. Assessment of competences, at the higher-order level, allows for different learning curves: skills and small pieces of knowledge may be learned in a different order or tempo, as long as the required competences are attained after a certain training period. Training can be adapted to trainee's needs, based on his/her assessment results. In our opinion, individualized training will result in more effectiveness and efficiency.

R3b

R4b

We emphasize assessment of cognitive processes in order to obtain diagnostic information on performance and especially performance shortcomings. Diagnostic assessment has been assumed to be useful for determining remediation or specialized training in the aviation domain (Seamster, Redding and Kaempff, 1997; see section 4.1), and for predicting future performance in training (Regian & Schneider, 1990). Causes for (in)sufficient performance are usually found in the underlying cognitive processes. This calls for an inference from observable behaviour and interaction with the trainee, for instance, by asking questions during coaching. Feedback will be more effective if coaches have insight into trainee's thinking patterns and strategies. The importance of assessment of cognitive processes is one reason why automated measurement in training simulators has been hardly applied (Sanders, 1991; Vreuls & Obermayer, 1985; see section 4.3). In addition, it seems to be difficult to automate assessment in complex tasks such as ATC because too many variables are involved, although some information can be logged. In our view, automated measurement, as far as possible, should always be combined with other assessment methods such as over-the-shoulder (OTS) observation to be able to facilitate learning processes.

Our competence-based view implies that not only the cognitive and technical aspects, but also the emotional and social aspects are to be assessed, since they all belong to competence (Bereiter, 2002). In this way, a complete picture about trainee's competence can be obtained. In assessment of aircrew these latter parts are referred to as non-technical skills; they are assumed to be indispensable for the maintenance of flight safety (e.g., O' Connor et al., 2002).

Finally, we should notice that competence-based assessment may indirectly contribute to better training results in ATC by using trainees' assessments for validation research of selection. Performance measures obtained in training must be reliable and valid to make this research valuable. Competences of qualified controllers can be applied as criteria in selection, because they optimally reflect the personal basis of job performance (Roe, 2005). Furthermore, the competences can be applied

R17a

R17b

in the design of work samples that are part of selection systems, since the tasks to be executed are strongly job-related (see section 4.3). The competences can be assessed by using the same performance criteria as in training, serving as predictors for future performance. In this way, we used the competences in the redesign of the selection system at LVNL. This system is called the 'Dutch air traffic controller selection system' (DATCOSS; Oprins, Geven, Veldhuijzen & Roe, 2006, 2007). However, selection is not the subject of this study and is not discussed further.

Principles of competence-based assessment

Our view on competence-based assessment made us adopt certain principles of assessment. Competences are not directly visible (see section 2.2) and have to be inferred from observable performance. Therefore, the assessment system uses a set of *performance criteria* by which a competence is made measurable. They are formulated in observable behaviors, also called 'behavioral markers', 'behavioral criteria' or 'performance dimensions' (e.g., Berk, 1986; Flin & Martin, 2001; O'Connor et al., 2002; see chapter 4). As in many performance assessment systems, the assessor rates these performance criteria on a rating scale.

Furthermore, we decided to use the same competences for assessment during training, from initial training until final job performance (see chapter 5 for the competences used). The generic character and higher-order formulation of competences makes this possible. For instance, the competence *planning* is relevant in each training phase and in each task execution whether simple or complex. It implies that the assessor rates the same performance criteria that belong to a competence over time, for instance, the criterion "Is flexible in adjusting plans" that belongs to *planning*. In this way, assessment of the trainee's competences is based on a substantial amount of different task situations and circumstances. This makes it easier to identify the trainee's strengths and weaknesses and to define appropriate training interventions at an early stage. In addition, progression on each competence over a certain training period can be measured. Progression provides an important indicator of whether a trainee is still learning. We consider this essential in pass-fail decisions: if a trainee has not reached a final learning plateau yet, it may be useful to continue training. In this respect, the extent of trainability of ATC competences must be taken into account (Regian & Schneider, 1990; Schneider, 1990; see section 2.3): consistent components (e.g., *radio telephony*, *strip management*) that are not mastered yet can be improved by practice in contrast with non-consistent components (e.g., *planning*, *attention management*) that are reasons for failing often.

We defined *performance standards* for successive training phases as part of the assessment design. In our view, while the competences to be assessed can stay the same, trainees should be assessed against increasing performance standards since trainees will not directly perform at the final competence level. We made the phases flexible in length in order to take into account different learning curves. Trainees will transfer to the next phase if they have attained the required standards. These standards refer to the requirements on the assessment tasks to be executed, such as task complexity, in combination with specific norms for handling these tasks, such as the amount of mistakes allowed and the degree of efficiency required. For

R6a

R4b

R4d

R9a

R4a

R4b

instance, in an early phase a trainee must be able to *plan* standard traffic adequately in such a way that the situation stays safe. We chose for expressing the performance standards in descriptions of exemplary behaviors. They can be considered as a variant on *behaviorally anchored scales* (BARS; Smith & Kendall, 1963; see section 4.2). There are not scale positions, but behaviors to be shown are spelled out. The examples also specify the meaning of each competence and help assessors to assign behaviors to the various competences. We believe that setting these standards will improve learning because trainees know what they must learn and how they perform. Agreement among coaches and assessors will be higher. The coaches know what they have to explain and emphasize in a certain phase, and which performance level they can expect from their trainees. Then they can adapt training to the trainee's needs in an optimal way, and this will increase the chance for trainees to succeed in training.

Assessment tasks in simulator training are mainly dependent on the construction of simulator scenarios. This can be a sequence of exercises or a test scenario specifically designed for assessments (Farmer et al., 1999). Many task variables are involved in ATC. Examples are: amount of in- and outbound traffic, flight destinations, diversity in aircraft performances, departure and arrival runways in use, and weather circumstances. Also specific events may occur such as conflict situations, runway changes, or emergencies. However, real-life traffic situations in on-the-job training (OJT) cannot be predicted. Nevertheless, we consider it important to structure OJT in a certain way because this will increase the quality of training (Jacobs & Jones, 1995; Rothwell & Kazanas, 1994). In our system, we provided some structure in OJT by describing the tasks that a trainee must be able to handle independently (without help of a coach) in a certain training phase.

Figure 7.1 illustrates how trainees are assessed against increasing performance standards in successive training phases in our assessment system:

Figure 7.1. Assessment against increasing performance standards

Figure 7.1 presents the performance standards of two subsequent phases and three learning curves. The trainees attain the required standards in phase 1 at the same

time. The trainee with the green curve obtains the standards of phase 2 earlier than the trainee with the blue curve, and the trainee with the yellow curve probably fails because he/she has reached a final learning plateau. Figure 7.1 shows that the learning curve is not linear and always ends in an asymptote of learning (see section 2.3; Newell & Rosenbloom, 1981): each trainee ultimately achieves a certain 'learning plateau' (e.g., Schneider, 1990) that can be below or above standards. Figure 7.1 also exemplifies that many variations in learning exist.

In this way, we are able to measure the amount of progression during training. We consider this important for determining whether the trainee is still learning or whether he/she has reached a learning plateau. We argued that learning involves a transition in regulation levels, from controlled to automated information processing, but the way in which this happens strongly depends on the individual's cognitive abilities (Ackerman, 1988, 1989) and other features (see section 2.2). A temporary learning plateau during training can occur, for instance, when an individual needs more time for automation of skills so that he does not have (extra) resources available to acquire new skills at that moment. Another reason can be that learning styles do not match with the type of instruction or coaching styles, for instance, the trainee needs better chunking methods (dividing information into meaningful pieces) or different strategies to solve problems. Motivational or emotional influences can also affect the trainee's susceptibility for learning. These temporary learning plateaus make it difficult to predict future performance. Chapter 10 addresses learning curves in more detail as part of the evaluation of the assessment system. We explain in section 7.4 how we designed the specific parts of the assessment system by giving concrete examples, based on the principles of competence-based assessment as discussed in this section.

R9b

R4c

R4d

R5a

7.3 Training design

The *training objectives* express what has to be learned within a certain training period (see section 3.3). Assessment should ascertain to which extent these objectives have been achieved. This implies that training and assessment are intimately connected with each other. We should notice that the assessment system has been redesigned for an existing training system. This section only describes the properties of this training system as far as relevant for the assessment redesign. Improvements in training were made simultaneously, but this is not subject of the present study.

Training in relation to assessment

The main redesign activities in training that relate to the redesign of the assessment system were the reformulation of training objectives in terms of competences, derived from the ATC Performance Model (see chapter 5), and restructuring training into successive phases. Mager (1962) gave three requirements concerning the formulation of training objectives: (1) they have to be described in observable behavior; (2) they have to define the conditions under which this observable behavior is manifested; (3) they have to indicate the standards for assessment (see also Kessels &

R14a

Smit, 1989). These requirements directly show the relationship with the performance criteria and standards as applied in assessments.

We chose to define training objectives at the higher-order competence level in addition to existing objectives at the lower task level. Such training objectives did not exist yet in the previous system (see chapter 6). We argued that only having objectives as a list of tasks (Farmer et al., 1999) does not fit well in a competence-based training design (see section 3.3). Competence acquisition implies that knowledge, skills and attitudes are merged in realistic learning environments as simulators and OJT. Training objectives formulated at the higher-order level allows for individual differences. We did not only define these training objectives for the final performance level, but also for intermediate training phases (see also section 7.2). In simulator training, the training objectives remained subdivided into objectives at the knowledge, skill and attitude level for *sub phases*. These objectives at the lower or micro-design level (e.g., Merrill, 1994) directly relate to theory lessons or a set of simulator exercises (e.g., handling a certain number of inbound traffic with similar aircraft performances in conformance with the standard rules). It is important to notice that the performance standards for assessments are formulated for phases and not for sub phases to be able to assess at the higher-order (competence-based) level instead of assessment of smaller pieces as we explained in section 7.2. Training objectives in OJT are never split up in lower level knowledge or sub skill objectives, because knowledge, skills and attitudes are merged even more when trainees enter OJT. We could say that sub competences are acquired in simulator training, while competences are learned in OJT. In addition, in OJT sub skills cannot be practiced separately due to the ongoing traffic situations.

Figure 7.2 illustrates the relationship between training and assessment in a competence-based design for simulator training (preOJT) and on-the-job training (OJT). It shows that the competences, operationalized into the ATC Performance Model (see chapter 5), are the basis of the *training objectives*. We did not only define training objectives for the end of training, but also for successive phases in preOJT (cf. sub competences) and OJT (cf. competences). Sub phases in preOJT have also training objectives at knowledge, skill and attitude level. Training objectives comprise a task description in combination with required behavior in accordance with the description of Mager (1962). In the assessment system the task directly relates to the *assessment task* ('what' is executed; e.g., a test scenario), and required behavior is expressed in the *performance criteria* ('how' the task is executed). Each phase has *performance standards* (cf. 'norms') as explained in section 7.2.

R9a

R3b

R4a

R4c

Figure 7.2. Relation between training and assessment design

Training structure

This section describes the training structure of the course Area Control (ACC), as main case, for which the assessment system was designed (see Appendix A for other courses in Unit Training).

PreOJT

ACC preOJT takes 18 weeks and is divided into two phases of 9 weeks. Theory lessons are planned in-between simulator exercises. The division is based on two main variables (see Table 7.1):

- *Traffic handling*: the requirements move from *safety* towards *efficiency* during training.
- *Traffic complexity*: traffic complexity, built in simulator scenarios, increases during training.

Table 7.1. Training objectives for the two phases in ACC preOJT (in Unit Training Plan)

<p>Phase 1 (9 weeks)</p> <p>An independently working trainee has to be able to handle inbound and outbound traffic as well as overflights in the Amsterdam ACC control area <i>safely</i> in conformance with valid rules and regulations for Planning/Coordination (PLC) and for Executive Radar Control (EC) in the following simulated task situations:</p> <ul style="list-style-type: none"> • Flights to and from Schiphol, all TMA fields and all regional airports • Standard amount of traffic (no inbound or outbound peak) • Types of flights: mix of in- and outbounds with overflights, various aircraft types and performances. • Standard weather and sight circumstances <p>The execution of speed control and vectoring does not have to be perfect yet, but should be conflict-free and in accordance with the standard model. The sequence does not need to be optimal (5 – 7 NM) yet.</p>
<p>Phase 2 (9 weeks)</p> <p>An independently working trainee has to be able to handle inbound and outbound traffic as well as overflights in the Amsterdam ACC control area <i>safely and efficiently</i> in conformance with valid rules and regulations for Planning/Coordination (PLC) and for Executive Radar Control (EC) in the following simulated task situations:</p> <ul style="list-style-type: none"> • Flights to and from Schiphol, all TMA fields and all regional airports • Standard and high amount of traffic (including inbound and outbound peaks) • All types of flights: mix of in- and outbounds with overflights, various aircraft types and performances, special flights (military flights, para-jumping, photo flights etc.) • All weather and sight circumstances (special sight conditions, turbulence, showers, wind etc.) • All special events (emergencies, radio problems, communication failures, runway changes etc.)

Due to the high fidelity of the radar simulator, a very high performance level at the end of preOJT can be achieved. The final level of preOJT with respect to traffic handling and traffic complexity is almost the same as the final level of OJT but in a simulated environment. Each phase is sub divided into a few sub phases that contain a set of exercises. Subordinate training objectives are formulated for a sub phase. These objectives are mainly task-oriented. This implies that they reflect the most relevant characteristics of the simulator scenarios, for instance, type of flights, number of flights, weather circumstances or specific events (e.g., runway changes) that occur in the exercises. They are included in training material to serve as information for trainees and coaches.

Trainees do the same set of simulator exercises that gradually increase in complexity alternated by periods of consolidation. It is possible, till a certain extent, to choose for specific or additional exercises for individual trainees. We stimulate this to make training more adaptive to the trainee's needs (see also section 3.3). Four trainees are trained at the same time, accompanied by circa two or three coaches; this depends on the exercises. The coaches rotate each week. The involvement of multiple coaches serves two purposes: minimizing influences on the learning process as a result of differences in coaching styles, and giving trainees the possibility to familiarize with different working methods and strategies of various coaches, which increases their flexibility.

OJT

We structured OJT differently from preOJT, based on three main variables:

- *Traffic handling*: the requirements move from *safety* towards *efficiency* during training.
- *Traffic complexity*: the requirements on traffic complexity increase during training.
- *Aid of the coach*: the requirements on working independently increase during training.

ACC OJT comprises four phases that are flexible in length as we argued; see Table 7.2:

Table 7.2. *Training objectives for the four phases in ACC OJT*

Familiarization phase (1-2 weeks)
A trainee has to get familiarized with the live-surroundings concerning radiotelephony, co-ordination and working methods in conformance with valid regulations and procedures. The coach follows the trainees' actions closely and provides much feedback and corrections. The coach may take over traffic handling in all situations.
Learning phase 1 (8-14 weeks)
A trainee has to be able to handle <i>standard</i> traffic both <i>safely</i> and <i>efficiently</i> , and <i>complex</i> traffic only <i>safely</i> without aid of the coach. The coach monitors the trainees' actions closely but the amount of feedback and corrections decrease during this phase. The coach may take over traffic handling in complex situations.
Learning phase 2 (8-14 weeks)
A trainee has to be able to handle <i>standard</i> and <i>complex</i> traffic both <i>safely</i> and <i>efficiently</i> . Examples of complex traffic are: threatening overload of traffic (above capacity of 70%), communication problems, emergencies, special sight conditions, runway changes with delay, and much military traffic. The role of the coach moves from guidance towards providing only some feedback and corrections. As an exception, the coach may take over traffic handling only in extremely complex situations if necessary.
Consolidation phase (1 week)
A trainee has to get more experienced and to become more comfortable with working <i>completely independently</i> . The coach is in the background, more passively and reservedly. The coach does not provide instruction anymore. It may not be necessary that the coach takes over traffic handling.

In OJT the trainee has to handle the traffic situations that occur in practice. Only sometimes it is possible to choose for specific sectors for arranging special learning moments. Characteristic examples were defined to serve as guidelines how to interpret the difference between *standard* and *complex* traffic. It is important to realize that *safety* and a certain degree of *efficiency* is always required in real-traffic situations. The coach helps to achieve these requirements, independently of the trainee's performance. This is different from simulator training in which trainees can experiment and discover their own solutions. In OJT, coaches need to intervene more often when a situation may become dangerous, because they are finally responsible for the traffic handling. We recognize the principles of 'cognitive apprenticeship' in the description (Collins, Brown & Newman, 1989; see section 3.3). However, since coaching strongly depends on the learning curve of the individual trainee and human interaction processes, it is not possible to pre-structure it completely. In OJT each

trainee has his own 'coaching team' that consists of four or five coaches. Multiple coaches are involved for the same reasons as mentioned for preOJT. The descriptions in Table 7.1 and 7.2 are elaborated upon by means of the performance standards (see section 7.4).

7.4 Properties of the assessment system

This section describes how we elaborated the general principles of competence-based assessment when designing the different parts of the system (see section 7.2). The Program of Requirements (PoR; see Table 6.1) has guided the design. Illustrations come from the assessment system designed for ACC training.

Continuous assessment versus performance tests

The new assessment system is based on the principle of *continuous assessment* among the majority of the assessment systems in ATC training (see e.g., Hopkins, 1995). It is not necessarily part of competence-based systems (and therefore not discussed yet in section 7.2). Coaches, who are also assessor, continuously measure trainee's achievement during training based on over-the-shoulder observation (OTS). We maintained continuous assessment for several reasons.

First, a main advantage of continuous assessment is that cognitive processes can be better assessed because of the continuous interaction between coaches and trainees. Coaches can make trainees to verbalize their thoughts (Farmer et al, 1999), for instance, by asking questions. This enables them to obtain diagnostic information that is required for adequate feedback as argued in section 7.2. Second, progression during a training period can be observed and taken as an indication for learning (Pellegrino, Chudowski & Glaser, 2001). We explained that this helps to make valid pass-fail decisions. Third, a more complete picture about trainee's performance is achieved, based on various task situations. Representativeness of task situations is guaranteed, because assessment is not restricted to a particular moment. In the simulator specific tasks could be selected but this is not possible in the OJT as we argued. Fourth, the users were used to the system of continuous assessment and its procedures. We tried to avoid unnecessary changes as they might affect user acceptance and, consequently, the implementation of the new assessment system.

It has been decided to use the continuous assessment method as objectively as possible. Multiple assessors are expected to assess the same trainee during a certain training period in order to increase reliability of assessment. Procedures for reaching pass-fail decisions were made as explicit as possible. Furthermore, we tried to diminish the pressure on trainees resulting from being continuously assessed by distinguishing different methods that vary in 'status': (1) *debriefing* only aims at providing feedback after a coaching session; (2) *progression measurement* is 'real' assessment over a certain period (one or two weeks) by rating a set of performance criteria. Their main purposes are respectively 'assisting learning' and 'assessment of achievement' (Pellegrino, Chudowski & Glaser, 2002; see section 4.1). The time in-

R11

R9b

R4b

R4d

C1a

C6

R10d

R5b

Intervals of progression measurements for various trainees are allowed to vary because of operational schedules of coaches. Two different reports are used, respectively *debriefing reports* (only in preOJT) and *progression reports*. The coach has a double role as an assessor and a coach; this role differs for debriefing (coach) and progression measurement (assessor).

Continuous assessment has some restrictions. The main disadvantage is that coaches are constantly influencing trainees' task performance by their guidance and interventions. This makes it difficult to assess whether a trainee is able to work independently. Assessments of more trainees cannot easily be compared with each other when there is a continuous interaction with a coach. This problem is even higher in the OJT when the assessments are based on varying task situations that cannot be similar for all trainees. This considerably affects the reliability of assessments.

For this reason it was decided to combine continuous assessment with performance tests (Berk, 1986). Performance tests measure the trainee's performance at the end of each phase in preOJT and at the end of OJT, without interference in the process. Its main purpose is to measure the trainee's level of competence against the performance standards, or 'assessment of achievement' (Pellegrino, Chudowski & Glaser, 2002). We emphasize the objective character of the test as a counterpart of continuous assessment. Simulators offer more possibilities for reliable testing. Automated measurement (see section 4.3) has not been developed yet, but this can certainly be a possibility in the future, complemented with human judgments as argued. The main reasons were that this would be technically too complicated for this project, and that the controllers did not have enough trust in automated measures yet. We chose for the involvement of two or three assessors; at least one of them is not a coach during training to be able to assess objectively. They have to summarize their findings on a test report, applying detailed and objective scoring techniques. Table 7.3 presents an overview of the assessment methods in the design.

We tried to make the assessment system in preOJT and OJT as consistent and transparent as possible for the users. Below, we describe the system in more detail. Our description is limited to assessment in ACC, but the majority also applies to other courses of Unit Training as the design has aimed at consistency and congruence between the several parts of the assessment system.

Table 7.3. General assessment methods

Method	Report	Purpose	Time span	Role
Debriefing	Debriefing report	Assist learning (coaching)	PreOJT: one exercise OJT: one day	Coach
Progression measurement	Progression report	Assist learning Assessment of achievement	One or two weeks	Assessor
Performance test	Test report	Assessment of achievement	PreOJT: one exercise OJT: one day	Two assessors (Examiners)

Competences and performance criteria

As explained in section 7.2, we derived performance criteria that represent the competences from the ATC Performance Model. In their operationalization, the typical (Dutch) jargon of the controllers was maintained in order to maximize recognition and comprehension by the users. In the ATC Performance Model, (see Figure 5.3) the competences mainly correspond with the gray parts and are marked in *italics*. The white parts were elaborated in terms of performance criteria. Despite of some small variations across courses, the same competences are assessed for all ATC functions (APP, ARR, GND, TWR, ACC), from the start of Initial Training till final job performance, as explained in section 7.2. Table 7.4 presents the competences assessed in ACC. These competences are rated in progression measurements and in performance tests, except for *attitude* and *teamwork ability*, which cannot be assessed in performance tests.

R3c

Table 7.4. Competences to be assessed in ACC

ACC competences
Safety
Efficiency
Verbal expression
Listening
Co-ordination
Equipment operation
Label management
Mental picture
Attention management
Planning
Decisiveness
Workload management
Attitude
Teamwork ability

In order to formulate the performance criteria, we made spreadsheets with the criteria used in the previous assessment system, criteria used in other courses, and the behavioral markers formulated in the competence workshops (see the example in Table 5.2). The completeness of the set of performance criteria and ambiguity between them became directly visible. Sets of performance criteria were extracted from the spreadsheet and discussed with controllers in the meetings of the design teams (see section 7.1). Finally, the performance criteria were selected and they were included in the progression and test reports (see Appendix B). They are not present at the debriefing reports because criteria are not rated here. Table 7.5 presents examples of the performance criteria that represent the competences *attention management* and *workload management* (translated into English).

Table 7.5. Examples of performance criteria of attention management and workload management

Attention management
Divides his/her attention over different situations
Executes several actions simultaneously (multitasking)
Holds information in his/her memory without forgetting it
Workload management
Adapts his/her work tempo to traffic load
Minimizes his/her workload as much as possible

Not only the competences but also the performance criteria stay the same during training (preOJT, OJT) within a course (GND, TWR, APP, ARR, ACC), and they are identically formulated for each ATC function (see also section 7.2). However, there are two exceptions, *safety* and *efficiency*, because they really have different meanings. For instance, *safety* for ground controllers implies that collisions on the ground have to be avoided, while *safety* for radar controllers means that radar separation must be maintained. *Efficiency* refers to avoidance of redundant delay for flights in all ATC functions, but its realization differs between ground and radar control, due to respectively traffic handling on the ground (two-dimensional) and in the air (three-dimensional). This difference is illustrated for *safety* in Table 7.6 (translated into English).

Table 7.6. Examples of performance criteria that belong to safety in two different ATC functions.

Safety (GND)
Prevents runway intrusions
Arranges conflicts and right-of-way situations in time
Checks the correctness of clearances on strips and EDD
Safety (ACC)
Maintains separation minima correctly
Builds in sufficient safety buffers
Switches from monitoring to vectoring in time
Co-ordinates with adjacent centers when handing-over conflicts

The ATC Performance Model provides indications on how the performance criteria can be assessed (see also section 5.2). First, the model separates objectively measurable criteria (outcome, actions) from criteria that can only be assessed subjectively (information processing). This distinction is useful for assessors who have to be aware of their own restrictions when they depend on subjective measures; they should provide evidence and meaningful arguments for their findings. Second, the model provides some information about trainability of competences. The actions are more trainable (or to automate) because they always improve with practice in contrast with the majority of the cognitive processes. The latter express the 'gut feeling' of assessors. They help them to argue why trainees performs (in)sufficiently. A conflict, for instance, might be solved inefficiently (outcome) due to a lack of anticipation on future air traffic (information processing). Insight into trainability of competences can help to take valid pass-fail decisions. If less trainable competences are a prob-

R10b

R10c

lem, it will be less useful to continue training (this will be further addressed in chapter 11).

In addition to the ratings on competences, we consider it important that assessors provide two additional ratings on (1) progression, which gives an indication about the amount of progress during a period of training, and (2) general impression in which assessors can express their 'gut feeling' about trainee's performance. The combination of progression and general impression gives information about the learning curve of the trainee. However, these criteria are not very reliable because they highly depend on the assessor's subjective feelings.

R10b

Performance standards

Trainee's competence is assessed against increasing performance standards defined for successive phases as argued in section 7.2. We explained that we chose for exemplary behaviors as a variant on behaviorally anchored scales (BARS). However, we considered alternatives before we made this choice. The 'real' BARS technique (Smith & Kendall, 1963; see section 4.2), did not seem suitable. The users were not able to define detailed behavior that could be related to more scale positions (e.g., good, slightly good). The importance of cognitive processes probably made this too difficult. Other alternatives were BOS or its variation PDA (Kane, 1986), based on critical incidents on several performance dimensions. At LVNL a serious attempt to define these critical incidents was done a few years before the start of this design project. This method had failed because it appeared to be impossible to define critical incidents. A reason could be that too many variables are involved in ATC, which results in a very broad spectrum of possible solutions.

R4c

R4d

The method used to generate the exemplary behaviors mainly consisted of an intensive analysis of assessor's comments documented in assessment reports of trainees (see also section 6.1: we applied the same method there). We assigned the comments to the different competences. Also the aforementioned observations and interviews (see section 7.1) added to the design process. We considered it very important that the controllers themselves, as subject matter experts (SME's), achieved consensus about the exemplary behaviors. This should contribute to a higher inter-rater agreement and a better recognition of behaviors. A first version was made by training experts in order to suggest the appropriate 'level of abstraction'. We selected most commonly used and representative examples. This first version was adjusted and extended with more examples by the controllers. Several versions were discussed in the meetings before the final version was accepted.

R7

C8b

In ACC preOJT the performance standards are a direct elaboration of the training objectives presented in Table 7.1. Table 7.7 shows the performance standards for efficiency (translated into English) for the two phases.

R9a

Table 7.7. Performance standards in two phases in ACC preOJT for efficiency.

Efficiency	Phase 1	Phase 2
<ul style="list-style-type: none"> ▪ Applies speed control correctly ▪ Applies vector technique correctly ▪ Takes into account aircraft performances ▪ Takes into account different flight levels ▪ Builds a sequence of climbing and descending air traffic 	<p>Speed control and vector technique need not to be optimal, but must be conflict free and conform standard routes and transfer, while taking into account differences in aircraft performances and turning to own or published navigation or speed in time. Some delay may still occur; sequences may be not efficiently enough yet. Application of level separation and assessment of intermediate levels during sequencing need not to be always optimal, for instance, 2 flights in different STARS are cleared to the same flight level.</p>	<p>Sequences consist of 5 to 7 NM intervals, with minimal speed differences by optimal speed control of inbounds, and with parallel handling over of outbounds by optimal vectoring. There is a striving for a continuous climb and descend, while taking into account differences in aircraft performances and turning to own or published navigation or speed in time. Delay has been avoided whenever this is possible. Level separation and assessment of intermediate flight levels during sequencing are applied optimally.</p>

The examples in Table 7.7 show that requirements on *efficiency* become higher during preOJT. The standards also clarify the meaning of *efficiency*. The standards for other competences (e.g., *planning*) have the same format. We should realize that the task complexity is predefined in a sequence of simulator exercises during the phases. The training objectives that are formulated for a set of exercises (see section 7.3) indicate the task characteristics. Assessors do not assess against the final level of the phase, but they derive the required level during a phase (e.g., in the 5th week in phase 1) from the task characteristics and difficulty in the exercises, and the standards as defined for the phase level as illustrated in Table 7.7. This implies, for instance, that trainees in the 5th week in phase 1 have to be able to handle the traffic in the exercises safely, but not quite efficiently yet. The simulator tests are at the end of phase 1 and phase 2. Their norms correspond directly with the standards that have been defined for the end of the two phases. We realize that it would have been better for the objectivity to define these kinds of standards for each progression measurement that has to be made during preOJT. However, we decided not to do this because it was too difficult to make so many illustrative examples. In addition, the system should be flexible enough to take into account individual differences in learning (see also section 7.2). Finally, the training objectives defined for sets of exercises provide some guidance to the assessors already.

R10a

In ACC OJT the performance standards are less clear than in preOJT because the task situations cannot be arranged beforehand. Therefore, the complexity of the task situations to be handled without aid of the coach is included in the exemplary behaviors in OJT as we argued in section 7.2. In contrast with preOJT, assessors assess against the final level of the phase during the whole OJT in order to maximize objectivity. This means that, for instance, a trainee in the 3th week of learning phase 1

probably receives a certain number of insufficient ratings, because he/she cannot have reached the final level (cf. performance standards) of learning phase 1 yet. This has consequences for the interpretation of assessment results; this will be further elaborated in chapter 10.

Table 7.8 presents an example for mental picture in ACC OJT (translated into English). It should be noticed that the text in the middle columns refers to the competence as a whole; the paragraphs in the text don't necessarily correspond with the performance criteria in the left column.

Table 7.8. Example of performance standards for mental picture in ACC OJT.

Mental picture	Familiarisation Phase	Learning Phase 1	Learning Phase 2	Consolidation Phase
Keeps a clear overview of the traffic situation by scanning regularly	As the trainee is still getting used to live surroundings he is inclined to focus too much on certain conflicts or flights instead of scanning the entire sector regularly for other traffic requesting attention. Therefore it may happen that certain flight is not cleared further. He may also be surprised by the call of an aircraft.	During normal traffic the trainee has a good picture of the traffic situation as he scans regularly. He is no longer surprised by calls of any aircraft as he doesn't only focus on specific conflicts or flights, but scans the entire sector regularly. Only during complex traffic may he still be inclined to loose his overview.	During normal and complex situations the trainee has a good overview of the traffic because he is scanning regularly.	The trainee works independently as stated in Learning phase 2 and has acquired sufficient experience at the end of this phase to take the practical exam.
Looks observes and takes action if necessary	At this stage the trainee finds it hard to react at everything he sees and hears. There is more visual and audible information than in the simulator. Therefore it is difficult to act adequately when traffic gets heavier.	By watching changes in information closely, listening to calls by aircraft processing information from the sector(s) and taking action where necessary the trainee is in control during normal traffic situations. He is able to concentrate during a longer period of time and is pro-active.	By watching changes in information closely, listening to calls by aircraft processing information from the sector(s) and taking action where necessary the trainee is also in control during complex traffic situations. He is able to concentrate during a longer period of time and is pro-active.	
Coach's available information is to be correct				
Guard is the identification	The trainee finds it hard to estimate intermediate FL's, the sequence of aircraft (e.g. diverging/converging tracks, speeds, and influence of the wind).	The trainee is monitoring all the time checking that the traffic situation develops as expected, e.g. after giving certain instructions, checking the wind influence and the pilot's reaction to the instructions.	The trainee has no problems estimating intermediate FL's, the sequence of aircraft (e.g. diverging/converging tracks, speeds, and influence of the wind).	
Anticipates future and variable traffic situations	At this stage the trainee finds it hard to divide his attention well and have a complete image of the traffic at all times, as he still has to get used to the live surroundings. Therefore it may happen he doesn't discover a flight without a label in time. And when he fits the strip into the sequence he doesn't realise the SSR code has to be adapted. Sometimes the coach has to remind him to change the SSR code or have it changed.	Only in complex situations the trainee finds it hard to estimate intermediate FL's, the sequence of aircraft (e.g. diverging/converging tracks, speeds, influence of the wind). During normal traffic situations the trainee can create a good mental picture of all traffic, also of flights that have not been identified by the system. He links the flightplan to the radarposition in time and correctly and sees to it that every flight has the correct SSR code.	During any traffic situation (also complex) the trainee can create a good mental picture of all traffic. Also of flights that have not been identified by the system. He links the flightplan to the radar position in time and correctly and sees to it that every flight has the correct SSR code.	
	=====	=====	=====	=====
	The coach is closely involved in the handling of traffic and gives tips regularly for a safe and efficient traffic flow. The coach asks many questions like "what will you do with that information, what does this information mean", etc. During heavier traffic the coach can take over control completely for a while.	The coach still gives tips to increase efficiency and only has to intervene in complex situations when safety is at stake.	The coach gives tips only occasionally for efficiency reasons and intervenes only rarely during complex situations and if safety requires.	The coach is only present as the person who is formally responsible for the safety.

Measurement techniques

Achievement of consistency in measurement techniques was one of the purposes of the redesign of the assessment system. We tried to design the measurement techniques in such a way that the reliability of the measures would be maximized.

We chose for a 6-points rating scale for progression measurements in ACC preOJT and OJT. Alternatives would be a 2- or 4- points rating scales as used in some other courses of Initial and Unit Training. However, we consider a 6-points rating scale better because it differentiates more. Each of these three possibilities strictly separates sufficient and insufficient achievements. Purposes are to distinguish the trainee's strengths and weaknesses clearly and to avoid the 'central tendency' error by not offering a scale position in the middle (Murphy & Cleveland, 1995). Insufficient ratings have to be accompanied by a comment (plain text) with further explanation; comments on sufficient judgments are not obligatory. These comments are considered to be important to elaborate the assessor's judgments.

We designed detailed scoring techniques for simulator tests in preOJT. Reliability of these tests is higher since they are based on well-balanced test scenarios. We asked the assessors to make their judgments as follows. After the trainee has executed the test, the three assessors start to fill in an individual test report independently of each other, based on their own notes during the test execution. Next, they discuss the differences between them and finally they fill in a final test report together. In this way, each assessor has an equal input. Furthermore, the interrater agreement across assessors can be calculated easily. This is needed for evaluation purposes (see chapter 9). We should note that in contrast with performance reports, the competences are rated instead of the performance criteria on test reports (see Appendix B). The competences are rated on a six points rating scale except for *safety*. This competence is measured more objectively by counting the number of safety violations. We should notice that these aspects could be automated by logging the data; this could be a further development in the future as we mentioned before. We made guidelines as to how many points should be lowered for each safety violation that belongs to a specific criterion, for instance, 20 points for missed conflicts. Table 7.9 represents the way in which *safety* is scored (translated into English). The total test report is included in appendix B:

Table 7.9. *Scoring of safety at the ACC test report*

Safety	(max 150)
V1 Conflicts solved adhoc but according to the correct separation minima (- 5 a time)	(max 30)
V2 Conflicts not solved according to the correct separation minima (- 10 a time)	(max 40)
V3 Conflicts not detected (- 20 a time)	(max 60)
V4 Traffic not transferred to adjacent ATC unit correctly and/or not in time (- 5 a time)	(max 20)

We decided to assign weights to the competences. The weighting relates to the ATC Performance Model (see Figure 5.3): the sum of the scores belonging to *information processing* (50%) has the same weight as the sum of the scores belonging to *actions*

and *outcome* together (also 50%), because information processing refers to the causes for (in)sufficient actions and outcome. This forces the assessors to think about possible causes for safety and efficiency problems and mistakes in actions. The assessors are told that the total score on *actions* and *outcomes* should be comparable with the total score on *information processing* because it cannot be, for instance, that a trainee works sufficiently safely and efficiently while the cognitive processes are insufficient. This system also helps assessors not to attribute mistakes in performance to too many different competences. We developed the weighting scheme in meetings with the controllers of the design team with the aid of several cases: how many safety violations, and what kinds of violations, are acceptable to pass the test? We calculated figures after several reports for fictive trainees had been filled out.

R4b

The final test score is the sum of weighted scores assigned to each performance criterion, presented in a 'percentage' (maximum value of 100). For ACC the maximum scores of competences (weighting scheme) are presented in Table 7.10. It shows that the final test score (in a percentage) is calculated by dividing the total score by 6, since the maximum total is 600. The norm for sufficient performance is set to 70% in accordance with past practice.

Table 7.10. *Weighting of competences in the ACC test report*

Outcome	(max 215)
Safety	(max 150)
Efficiency	(max 65)
Actions	(max 85)
Verbal expression	(max 20)
Listening	(max 20)
Co-ordination	(max 30)
Equipment operation	(max 5)
Label management	(max 10)
Information Processing	(max 300)
Mental picture	(max 80)
Attention management	(max 40)
Planning	(max 70)
Decision making	(max 60)
Workload management	(max 50)
Final test score (in %)	(max 600/6)

However, we should realize that this score is not a real percentage: the scale is based on ratings at an ordinal 6-points scale (except for *safety*). We decided to maintain this 'percentage' in the tests because this scale is easy to interpret for assessors who have been used to it for many years. This makes it easier for them to compare their 'gut feeling', expressed in *general impression*, with the weighted sum of ratings that results in the final test score. A comparison between the final test score and *general impression*, also expressed in the same percentage scale, was made to examine whether the weighting was chosen correctly. For ACC, the absolute deviation between these two variables was only 3.3% based on 36 tests. We considered this deviation very low. We also calculated the deviations (in %) between

R10b

the total score on *actions* and *outcomes*, and the total score on *information processing* because these two should be comparable as well in conformance with the aforementioned philosophy of the weighting. The absolute deviation for ACC tests ($N = 36$) was only 6.7%. We concluded that the weighting was well chosen and we kept it in the final version of the design.

Pass-fail decisions

At the end of each phase in preOJT and OJT (see section 5.2), a pass-fail decision is made in order to avoid useless effort and costs in further training of prospectless trainees. These are fixed moments of time in preOJT but not in OJT, because OJT phases have a flexible length as argued in section 7.3. In OJT, a certain maximum of hours has not been clearly defined, but a trainee normally fails if he/she does not show any progression after being in the training phase for quite a long period of time. Training managers have the responsibility to make these pass-fail and other decisions (e.g., remedial teaching, retraining) in both preOJT and OJT. They are advised by coaches and training experts. Decisions are based on progression measurements and performance tests because they all offer essential information. In preOJT a trainee may have a certain number of insufficient ratings on some competences and go on to the OJT, but in OJT all competences must be rated as sufficiently to go on to the next phase.

We decided not to have a fixed quantitative cutoff because the reliability of the ratings is not sufficient yet, and qualitative information is also essential. Training experts do an extensive quantitative and qualitative analysis to get insight into the trainee's deficiencies. This serves as the basis for a well-founded prediction of performance in further training. Additionally, possible external influences (e.g., private circumstances, working environment) are taken into account.

Assessors

Multiple assessors are always involved. In preOJT, we assigned a fixed group of eight controllers to assess the trainees. The other coaches are only allowed to coach, not to assess. At least one out of three assessors who are involved in simulator tests has not been the coach of the trainee in the period before in order to achieve more objectivity. In OJT, coaching teams of four or five coaches train one trainee as explained in section 7.3. The same coaches also assess their trainees in OJT.

Assessors experience some difficulties. The assessors have a double role as a coach and an assessor in the continuous assessment system, especially in OJT. The main advantages of this double role are that the assessor gets a complete picture about trainee's competences due to the continuous interaction with the trainee during training, and that progression over a certain training period is measured. But assessors find it difficult to separate the two roles. Experiences in earlier training, a general image of the trainee and opinions of colleagues may influence their judgments. This problem relates to the presence of rating errors such as the *halo-error* or

R5a

R5b

R12b

R10d

leniency and *severity errors* (see section 4.2). The occurrence of rating errors is evaluated in chapter 9.

Sometimes *self-assessment* (see section 4.1) is applied, especially in OJT. Self-assessment is not obligatory or formal, but in some cases, the coaching team finds it useful to let trainees assess themselves. It enhances self-reflection and often improves the trainee's self-confidence. Our experience is that trainees tend to assess themselves less high than their assessors do. Self-assessment can certainly be extended and made more formal in the future, but therefore more research on reliability of self-assessments is needed. *Peer-assessment* is not applied, because trainees are trained rather individually and they do not see their peers working often (see also section 3.3).

7.5 Design of a web-based assessment tool

The final step in the design process was the creation of a web-based assessment tool, which was the subject of a separate LVNL project. We followed the steps in the design process described in chapter 1. In fact, it serves as a small *design cycle* within the total cycle for the assessment design. The first step, 'analysis', consists of the formulation of the purposes, requirements and constraints, and has resulted in a Program of Requirements (PoR). The second step, 'design & implementation', involves the development of a functional design, followed by programming the system, which was implemented gradually. The third step, 'evaluation', concerns an evaluation that has led to improvements of the first version of the design.

Construction of the PoR

The first step in the phase of 'analysis' was to define the purposes and the main properties of the web-based assessment tool, and to draw up a Program of Requirements (PoR). A project team was formed to design the system, and also a group of 'stakeholders' was identified. The group of stakeholders consisted of a delegation of managers. The project team constructed the first version of the PoR, and the group of stakeholders added their comments. Next, a general meeting with this group and the project team was organized to discuss the required properties of the system.

The system has been defined as a web-based tool that has three main functions: to fill in assessment reports digitally (input), to store the training results directly in a database (content), and to extract several reports and overviews of trainee performance (output). The different ways of access to the system must be arranged with authorizations. This resulted in a global picture of the system, consisting of three main components, presented in Figure 7.3:

Figure 7.3. Overview of the components of the web-based assessment tool

A web-based assessment tool is supported to have several advantages above (previously applied) written assessment reports. First, the system forces assessors to fill in assessment reports accurately and univocally. This may increase reliability and allows for more control. Authorized persons can check from the office whether reports are filled in correctly. Mistakes with dates, names, and practical use of the rating scales cannot occur anymore. Second, storage of assessment results in a database becomes feasible. This is more efficient and secure than storing reports on paper, and the confidentiality of personal data can better guaranteed. Third, in contrast with reports on paper that have to be read one after each other, various summaries and overviews can be made, for instance, average scores on the different competences for a single trainee or the number of (in)sufficient reports obtained in a specific course. It would even be possible to generate graphs that represent the learning curves of trainees (see chapter 10). These various summaries and overviews provide more insight into individual learning processes and trainees' strengths and weaknesses. History can be built up about trainee performance, which can be used on a foundation of pass-fail decisions. Fourth, authorization can be arranged for security reasons and authorized persons can have access to the system from several places. The different users should get access with a login code and password. In this way, interested persons who have access from several places can better follow trainees' progression. From the office they may get an overview of training results obtained in the OJT that takes place at the tower. Finally, assessment results can be used for reliability studies and validation research for both training and selection purposes, exporting the results from the database to a statistical program (SPSS). The system strongly resembles a *competence management* system, because it allows to keep track of trainees' competence development from selection till operational controller.

The requirements in the PoR were sub divided into the components *input*, *content*, *output* and *system* as presented in Figure 7.3. The majority of requirements are very specific in order to be as explicit as possible before the start of the Functional Design, the next step in the design process. In this respect, this PoR is different from the general PoR for the assessment system as a whole (see chapter 6), due to its focus on the development of a computer system. The majority of general constraints for the design of the assessment system are the same for the web-based assessment tool, such as practical usability (C4), transparency (C5), acceptance (C6), confidentiality (C7), input of controllers (C8), benchmarking (C9), continuous evaluation (C10), and quality assurance (C11). The complete PoR for the web-based assess-

C2

C7

R4b

R4c

R4d

R6a

R6b

ment tool is presented in Table 7.11. The PoR highlights a few important properties of the web-based assessment tool. The tool is only used for training purposes: filling in assessment reports, storing the assessment results in a database, and generating several overviews of trainee performance. Access from several places, via the intranet, is arranged by authorizations.

Table 7.11. *Program of Requirements (PoR) of the web-based assessment tool*

Input

R1 Boundaries of input of assessment results

- a. The input of assessment results starts from Initial Training and ends when trainees have finished Unit Training. The web-based assessment tool is only used for training purposes.
- b. The tool is not mentioned for the input of selection results, however, it should be easily possible to combine selection results and assessment results (e.g., in another database) for validation research.
- c. The assessment results that are obtained earlier (typed over in Excel) should be easily combined with the assessments that are filled in digitally by the assessors in the tool itself.
- d. The tool is mentioned for training of air traffic controllers at Schiphol Airport (ACC, TWR/APP) and regional airports and related functions (VLA ACC, SUC, FIC), but not for other functions (e.g., at the office).

R2 Filling-in assessment reports digitally

- a. The functionalities to fill in assessment reports should be very easy for assessors in order to maximize acceptability by the user group.
- b. The layout of assessment reports in the tool should be identical to the reports on paper as far as possible.
- c. It should be possible to fill in both qualitative and quantitative information, for instance, comments that belong to a specific competence to be rated.

Content

R3 Personal information

- a. The tool should contain personal information of trainees (e.g., name, address)
- b. The tool should be connected to the database in use by the HRM-department.

R4 Assessment results

- a. The tool should store the final results of theory tests in Initial Training and in Unit Training.
- b. The tool should store the results of practical training (simulator, OJT), but only progression measurements and performance tests (no debriefing).
- c. The ratings on each competence and performance criteria should be stored in the database.
- d. The tool should store general information about the assessment: names of assessors, dates, number of hours in OJT, and possible special information (e.g., specific training positions or runway combinations)

R5 Management information

The tool should contain general management information that directly follows from assessments such as training trajectory, (order of) courses, pass-fail decisions, and dates of entrance and outcome.

Output

R6 Reports and overviews

- a. The tool should make it possible to generate an overview of training results for one trainee with personal information, training trajectory, average scores on each competences in several courses, pass-fail and so on.
- b. The tool should make it possible to generate an overview of training results of several trainees in a specific course with average scores on each competences, numbers of hours in OJT, numbers of pass-fail and so on.
- c. It should be possible to print the various overviews and summaries.
- d. It should be possible to print the assessment reports filled in by the assessor for the signatures of assessors and trainees and for keeping them in a paper dossier.

R7 Export and import

It should be possible to export assessment results to another database (Excel, SPSS) for statistical calculations, and to import assessment results that are obtained earlier to the database of the tool

R8 Searching

A searching functionality should be available to search for trainee, course, date, assessor, etc.

System

R9 Authorization

- a. Access to the tool should be arranged by authorizations for different types of users (e.g., filling in reports, getting out reports, administration) with personal accounts and passwords for security reasons.
- b. An administrator should be responsible for the arrangement of authorizations.

R10 Administration

- a. Assessment reports (e.g., layout, content) should be easily to make and adjust by an administrator, for instance, in a separate administration tool.
- b. An administrator should be able to enroll trainees in courses, to assign assessors and to define assessment reports for the several courses, for instance, in a separate administration tool.

R11 Input of data

- a. Possible mistakes in input should be blocked (e.g., by special menus or blocking options)
- b. Obligatory and optional fields should be made visible.
- c. Multiple users should be able to work with the tool at the same moment.

R12 Network

- a. It should be able to install the tool at systems that are supported by the IT department.
- b. The tool should be available at the intranet for access at several places (e.g., tower, office)

R13 Further development of the tool

- a. It should be possible to extend the functionalities of the tool if necessary.
- b. It should be possible to enlarge the boundaries of the input into the tool
- c. It should be possible that the database of the tool could be connected to other database in use at LVNL in the future, or that data can easily be exchanged.

Properties of the tool

The Program of Requirements was used as the basis for the construction of a *Functional Design*, made by the project team. This Functional Design describes in more detail the properties of the system such as the main functionalities, user interface, database properties, and security levels for authorization. The next step was to compare different existing computer-based systems with each other in order to investigate which system would be most suitable. We choose for the assessment tool Questionmark Perception (QMP), because LVNL was using this tool for theory exams already. Relatively small adjustments were needed to build a specific application within this system; thus, this solution would save costs in comparison with complete new systems to build or to buy. The final step was to define the *Specifications* of the application to be built in QMP. We did this in close collaboration with the programmers. These Specifications served as the basis for the subsequent creation of the system.

The assessment tool is web-based but only available from the intranet (not from the internet), due to general security regulations of LVNL. Several types of users are authorized to enter assessment data. The group of stakeholders approved these authorizations. We explain in several steps how the assessor uses the system by means of screen dumps made from the records of a trainee in ACC OJT (in Dutch). This does not only show the assessment tool, but it also displays the assessment system that was described in section 7.4.

After selecting the correct URL on the intranet, the users login with a user name and a password:

Voer uw gegevens

Login

Naam: ACC_DH

Wachtwoord: ••••••

Aanmelden

Password

Leerling Volg Systeem (LVS)

After registration, the users can select their own name as an assessor, the course involved or the name of a trainee. Next, on the left site the assessor sees an overview of the trainee's previous assessments on which he can click on to have access to a specific report. On the right site the assessor can select the assessment report to fill in. In this case, there is only one report but there could be more dependent on the course involved:

Previous assessment reports

Assessor

Overzicht gegevens
Leerling:

Beoordelaar

Ingevulde beoordelingen

Beoordelaar: Heijnism

Beschikbare formulieren

Empty assessment report

Beoordelaar	Ingevulde beoordelingen	Beschikbare formulieren
Sloorn	ACC Progressierapport QJT 10-2-2006 07:02	ACC Progressierapport QJT
Beertjesp	ACC Progressierapport QJT 26-1-2006 12:33	
Dam J.F. van	ACC Progressierapport QJT 18-1-2006 10:25	
Dam J.F. van	ACC Progressierapport QJT 4-1-2006 07:37	
Beertjesp	ACC Progressierapport QJT 3-1-2006 00:32	
Sloorn	ACC Progressierapport QJT 17-12-2005 12:11	
Heijnism	ACC Progressierapport QJT 12-12-2005 11:00	
Heijnism	ACC Progressierapport QJT 23-11-2005 13:29	
Beertjesp	ACC Progressierapport QJT 13-11-2005 10:23	
Dam J.F. van	ACC Progressierapport QJT 9-11-2005 15:17	
Heijnism	ACC Progressierapport QJT 8-10-2005 08:46	

Terug naar selectiescherm

Leerling Volg Systeem (LVS)

After the assessor has selected a new assessment report to fill in, he gets the template in which he can rate the performance criteria:

Competences & performance criteria

Progressierapport
Executive- & Planning Controller ACC

Insufficient

Sufficient

Remark

Beoordelingscategorieën en items	Onvoldoende	Voldoende	Opn.
Veiligheid			
Past de separatieminima correct toe	☐	☐	☐
Gaat tijdig over van monitoren naar vectoren	☐	☐	☐
Bouwt voldoende safety buffers in	☐	☐	☐
Draagt het verkeer correct en tijdig over aan de naastliggende ATC unit	☐	☐	☐
Efficiency			
Past vectorotechniek correct toe	☐	☐	☐
Past speedcontrol correct toe	☐	☐	☐
Houdt rekening met onderlinge hoogteverschillen en windinvloeden	☐	☐	☐
Houdt rekening met (verschillen in) vliegtuigprestaties	☐	☐	☐
Beperkt zijn eigen workload zoveel mogelijk	☐	☐	☐
Bouwt een sequence van dalende en klimmende verkeersstromen	☐	☐	☐
Uitdrukkingsvaardigheid			
Drukt zich goed uit in (non)-standard fraseologie	☐	☐	☐
Drukt zich kort, bondig, eenduidig en kortaat uit	☐	☐	☐
Heeft een duidelijke, rustige uitspraak en intonatie	☐	☐	☐

The assessor can provide numbers for remarks in the most right column. The last page contains a text box in which the assessor is expected to write down his remarks in plain text. Finally, the report is saved as a PDF-file (Acrobat Reader). After

saving, changes are not possible anymore for security reasons. Now a print can be made to discuss the report with the trainee. Other interested persons such as training managers and training experts have access to the system but with another account (more rights). Exporting the data to Excel or SPSS can generate overviews. QMP can also make some specific reports.

Introduction of the tool

We implemented the web-based assessment tool gradually in order to monitor possible difficulties closely and to take the time to make the users familiarized with the system. First, we started a pilot in two simulator courses (ARR and ACC), in which only a small number of coaches and trainees was involved. We instructed each coach individually to be sure that the coach did understand the purposes and practical use of the system. At the office, a 'helpdesk' was available. After the pilot, we held an evaluation by means of individual interviews and a discussion in a general meeting. The results of this evaluation lead to the specifications of the second version of the tool. In the meantime, we started to implement the actual system in the other simulator courses (GND, APP, TWR). In this way, we made the coaches getting used to the tool in a rather safe and controllable environment: the simulator. The group of coaches in the OJT is larger and more difficult to reach. Before we introduced the tool in the OJT, we explained the purposes, benefits and practical use of the tool in OJT-meetings. After these OJT-meetings we started to implement the tool in the OJT, first in ACC and next in the various courses of TWR/APP. Table 7.12 presents a summary of the implementation activities:

Table 7.12. *Overview of implementation activities of the web-based assessment tool*

Date/period	Activity
December, 2004	Pilot with the first version of the tool in preOJT
January, 2005	Evaluation of the first version with the pilot groups
February, 2005	Implementation in the other preOJT courses
July, 2005	Implementation of the second version in all preOJT courses
May-October, 2005	General introduction at OJT-meetings to all ACC coaches
June, 2005	Implementation of the tool in ACC OJT
July, 2005	General introduction at OJT-meetings to all coaches
September, 2005	Implementation of the tool in the whole OJT

Evaluation against the PoR

The tool has been changed and improved continuously until now. We could say that the project of designing a web-based assessment tool is not finished because a number of requirements in the PoR are not met yet. We evaluated the system (December, 2006) against the PoR by providing a rating for each requirement on a three points scale: 1) not achieved; 2) partly achieved; 3) completely achieved. The results are presented in Table 7.13. It shows that the majority of requirements is (partly) achieved, especially for the user group itself: trainees, assessors, and training man-

agers. They experience as the main advantages that assessments can be filled in rather accurately in an easy way and that assessment reports can be viewed from many places in order to follow trainees' progression. Office personnel can quickly generate overviews of trainees' performance such as averaged ratings on competences. In this way, possible deficiencies can be detected as early as possible and timely interventions can be launched if necessary. The history of assessment reports stored in the database has been used since three years now for taking pass-fail decisions. Furthermore, statistical analyses can be executed by exporting the results to SPSS. This has provided useful information with regards to the evaluation of the assessment system (see chapters 9, 10 and 11).

However, the design of the web-based assessment tool is not finished yet for several reasons. Management information is not directly available in the tool. Other applications (e.g., Excel, SPSS) should be used for reports and overviews such as summaries of individual trainees' performance (e.g., learning curves, averaged ratings on competences), and of courses (e.g., pass-fail percentages, averages across trainees). Improvements of the technical infrastructure at LVNL (intranet) are needed since there exist too many technical problems for the user group.

In conclusion, we learned from the project that using an existing system (QMP) added by some small adjustments is not always the best long-term solution despite of the lower costs. Therefore, attempts have now been undertaken to look for another and better system while maintaining the main advantages and positive experiences by the user group.

Table 7.13. *Evaluation of the web-based assessment tool against the PoR.*

Requirements PoR	Rating	Explanation
R1 Boundaries of input of assessment results	2	The tool has been used in the whole scope of Unit Training but not yet in Initial Training; selection and assessment results (also obtained earlier) could be combined but not in an integrated selection and training system yet.
R2 Filling in assessment reports digitally	3	The assessors can easily fill in assessment reports; the layout is similar with paper reports; quantitative and qualitative information can be provided.
R3 Personal information	2	Only a small part of personal information is included in the tool (name, course); the tool is not connected yet to the HRM-database.
R4 Assessment results	2	The tool stores the results of progression reports, ratings on competences and criteria, but not of performance tests and theory tests yet; general assessment information is stored: assessors' name, date, hours in OJT etc.
R5 Management information	1	Management information can be generated in other applications (Excel, SPSS) but is not included in the tool itself yet, such as training trajectory, pass-fail decisions, dates of entrance and outcome of courses.
R6 Reports and overviews	2	Some reports can be generated by QMP but the majority only can be made in other applications (Excel, SPSS). This involves overviews of training results of single trainees and courses (e.g., averaged ratings, pass-fail, hours in OJT). The assessment reports (as PDF-files) can be shown and printed by the tool itself.
R7 Export and import	3	Assessment results can be exported to other applications (Excel, than SPSS) for statistical analyses, and other assessment results can be added.
R8 Searching	3	QMP contains a searching functionality.
R9 Authorization	3	Different authorizations can be assigned in QMP to various user groups with personal accounts and passwords, arranged by an administrator.
R10 Administration	3	An administrator is responsible for making and adjusting assessment reports in a separate administration tool, and for enrolling trainees into courses and assigning assessors and assessment reports to courses.
R11 Input of data	3	Input of data is arranged with menus, obligatory and optional fields are visible to the users, and multiple users are able to work with the tool at the same time.
R12 Network	2	The tool is available at the intranet, however, there are regularly technical problems with the access to the intranet and the tool itself, due to the LVNL infrastructure, especially during weekends.
R13 Further development of the tool	2	Extensions of the tool are rather restricted due to the properties of QMP. It is difficult to make other templates (reports) and reporting facilities. It is possible to link the database with other databases, but this has not been done yet.

7.6 Conclusions

This chapter described the design of the competence-based assessment system for simulator and on-the-job training for LVNL, both the design process and the product properties. The assessment system is based on competences in the first place. All facets of competence are assessed (technical, cognitive, emotional, social) to get a complete picture of the trainee's performance. Each competence is represented by a set of performance criteria, rated at a 6-points rating scale. In order to follow the trainee's progression on each competence over time and to get a better insight in (deficiencies in) performance in different task situations, the same competences are assessed during training. Augmenting performance standards ('norms'), expressed in exemplary behaviors, are defined for the subsequent training phases in simulator training and OJT. Multiple assessors measure the trainee's progression over one to two week periods in a continuous assessment system, and additionally performance tests are conducted. A web-based assessment tool is used to make assessments, to store the results in a database and to generate overviews of trainee performance. This makes it possible to evaluate the main properties in both a qualitative and a quantitative way; this will be subject of chapters 9, 10 and 11.

8 Implementation

The step of implementation in the redesign process involves the production of the materials, the organization of the new activities, and the introduction of the system to the user group. Most critical in the implementation of our assessment system was to get it accepted by the users, and therefore this is emphasized in this chapter.

A basic principle to enhance implementation is to incorporate future users in the design process, and to present not only completed products but also intermediate (partial) products to those users who are not involved in the design process (Moonen, 2000). This relates to *internal* and *external acceptability*. It is assumed that internal acceptance by those involved in the design and development process would also result in external acceptance by the end users, but this does not always need to be true. According to Moonen (2000), the main objective of implementation is to confront the actual end users with the final (partial) product and to create an opportunity for the end users to adapt the product to their own individualized specifications. A method to achieve this is to make the design process iterative: development of partial products is followed by implementation (e.g., in a pilot) and evaluation, which results in adjustments and improvements of these (partial) products. This evaluation is not an ultimate evaluation of the final product yet (see chapter 13), but it relates to 'simulation' in the design cycle of Roozenburg and Eekels (1991). It precedes the realization and the use of the final product, and allows changes in these (partial) products.

We tried to follow these recommendations in designing the assessment system. This chapter describes the process of implementation. Section 8.1 discusses the steps in the process of implementation. Section 8.2 describes the design of the assessor training.

8.1 The process of implementation

The assessment system was implemented in Unit Training gradually. The first version was already introduced in the OJT ACC in September 2001, but at regional airports the system was not renewed before January 2004. This section describes how we implemented the assessment system in ACC training.

Implementation activities

The first step was to produce the materials (reports, manuals) and to organize that assessors could use the new system. In September 2001, we started with a pilot for OJT ACC with two trainees and a team of five coaches. We chose for this small pilot to be able to make changes in the design before the assessment system was introduced in the training as a whole. We explained how the system works to individual coaches during these first trials. The coaches started to use the system without refusing to cooperate. After a pilot of two months, we evaluated the system by interviewing some trainees and coaches. This resulted in small adjustments such as rephrasing some performance criteria and standards, and making new layouts of the reports. Next, we introduced the first version of the assessment system in a meeting for course managers in order to implement the system in the OJT as a whole, but only for the first sector cluster (see Appendix A). This first version was used from November 2001 until April 2002.

In April and May 2002, we introduced the assessment system to all preOJT and OJT coaches in another meeting in order to enhance acceptance and correct use of the system. We explained how the assessment system was designed, and how it should be used in practice. In the meantime, the first version was evaluated and improved. A second version was used from January 2003 in the preOJT as well as in the OJT, also for the second and the third sector cluster. Simulator tests in a renewed form were applied in June 2002 for the first time. We accompanied the first two times in order to receive feedback from coaches and trainees about the practical use. A year later (June, 2003), we refined the weighting of the performance criteria (see section 7.4) based on collected data. Finally, we introduced the web-based assessment tool Questionmark Perception in December 2004 (see section 7.5). Table 8.1 presents an overview of the implementation activities:

Table 8.1. *Overview of implementation activities in ACC training*

Date/period	Activity
April, 2001	Presentation of 1 st version to ACC course managers (Cuco-meeting)
Sept. – Oct. 2001	Pilot (1 st version) ACC OJT (1 st section cluster) with two trainees and five coaches
September, 2001	Evaluation of the pilot with ACC course managers (Cuco-meeting)
Nov. 2001 – April 2002	Implementation 1 st version (small adjustments) in OJT with all trainees and coaches
April – May 2002	OJT meetings: evaluation 1 st version + presentation for all coaches
June, 2002	Introduction 1 st version simulator test in preOJT with two trainees
January 2003	Implementation 2 nd version in both preOJT and OJT with all trainees and coaches
June, 2003	Introduction of the 2 nd version simulator test in preOJT with four trainees
December, 2004	Start of the implementation of Questionmark Perception (see section 5.6)

Evaluation of the implementation process

We put a lot of effort in the implementation activities for several reasons. We tried to achieve *acceptance* by the user group (course managers, coaches and trainees) and *correct use* of the assessment system (e.g., procedures, use of reports). Before the start of this project, other serious attempts had been undertaken to improve the assessment system, but these attempts had failed due to disagreement of the users with the design method. They did not accept the new proposals. Therefore, we wanted to be extra careful in our project when implementing the system. The most essential implementation activities to enhance acceptance were: 1) close involvement of the users (coaches, course managers) in the design process as members of the design team (see also chapter 7, section 7.1); 2) close involvement of the designer(s) in the training process (e.g., by employment in the organization); 3) gradual introduction of the assessment system in several steps; 4) organization of meetings to explain the system to the users.

We did not experience many difficulties with the willingness of coaches to apply the assessment system. They understood the benefits, which were explained in the meetings, quite well. In chapter 12 we describe how we evaluated the assessment system with the users.

8.2 Training of coaches and assessors

An important part of implementation was the training of assessors. The assessor training is embedded in the coach course given by LVNL.

Coach course

The European regulations (ESARR, 2000) put specific requirements on coaches, assessors and examiners in ATC training to guarantee their quality, also in relation to safety requirements. There exist three special ratings involved with ATC training: 1) coach (OJTI); 2) assessor (ASE); 3) examiner (EXM). LVNL restricts the function of assessor to proficiency checks for operational controllers. The coach course at LVNL aims at the acquisition of the ratings of OJTI and EXM. This course takes three days and is given by training experts (train-the-trainers). The subject 'assessment' takes one full day and covers theory lessons and practical exercises. In addition, each person with an OJTI, ASE or EXM rating must get a refresher training once in three years.

Table 8.2. *Training objectives of assessment training*

Knowledge objectives: Evaluation and assessment. The trainee . . .
Assessment in general
Understands the main problems of (performance) assessment in practical training
Understands the most frequently occurring rating errors
ATC performance model
Understands the ATC Performance Model of the air traffic controller
Is able to explain the ATC competences in relation to the model
Understands the performance criteria belonging to each category (competence)
LVNL assessment system
Is able to describe the importance and principles of continuous assessment
Is able to explain the properties of the LVNL assessment system
Debriefing reports
Is able to explain the purposes and properties of debriefing reports
Understands the guidelines how to fill in a debriefing report
Is able to provide examples of clear and unambiguous language in debriefing reports
Progression reports
Is able to explain the purposes and properties of progression reports
Understands the guidelines how to fill in a progression report
Is able to provide examples of clear and unambiguous language in progression reports
Tests and examination
Is able to explain the purposes and properties of performance tests
Knows the background of the scoring methods derived from the ATC Performance Model
Knows the responsibilities as examiner and the content of the Examination Regulations
Skill objectives: Evaluation and assessment. The trainee . . .
Assessment in general
Writes the several reports in clear language, with meaningful arguments and clear evidence.
Provides not only negative, but also positive feedback supported by tips for improvement
Restricts his/herself to really observed behavior, distinguishing facts from meanings
Discusses the findings in the reports intensively with the trainee
Provides well-founded pass/fail decisions, supported by concrete and clear arguments
Separates the role of coach and assessor clearly in the 'continuous assessment' system
Debriefing reports
Restricts his/herself to the main points of interest, related to (individual) training objectives
Writes debriefing reports in the role of coach instead of assessor
Progression reports
Applies and interprets the performance criteria belonging to each category correctly
Judges against (intermediate/final) phase level, using the behavioral examples as standards
Mentions trainee's strengths and weaknesses clearly, supported by concrete evidence
Judges trainee's progression over a certain learning period clearly
Avoids rating errors as much as possible
Tests and examination
Applies and interprets the performance criteria belonging to each category correctly
Judges against (intermediate/final) phase level, using the behavioral examples as standards
Bases judgments only on observations during the test/exam, not on experiences as a coach
Applies the scoring methods (and checklists if available) correctly
Avoids rating errors as much as possible
Applies the rules of the Examination Regulations correctly

Assessor training

We formulated training objectives for assessment in the coach course. They are included in the Unit Training Plans for OJT and EXM (ESARR, 2000), see Table 8.2.

The assessor training comprises two parts: theory lessons and practical exercises. The theory part contains general assessment theory, a description of the LVNL assessment system, and an explanation of the most frequently occurring rating errors (see also chapter 4). Afterwards, the participants practice in a simplified simulator environment, specifically designed for coach training by Eurocontrol. A performance test is executed and two or three participants, in the role of assessors, have to assess the same (simulated) trainee using the methods and reports that are being used in practice. First, they fill in the final report independently, and next, they are forced to discuss the differences. This also serves as a method to get more familiarized with the meaning of the various performance criteria and standards.

8.3 Conclusions

Much effort was put in the implementation of the redesigned assessment system, as this was considered very important for its correct use and acceptance. Implementation activities comprise the production of the materials, organization of new activities, and the introduction of the assessment system to the users. The close cooperation between the users and designers in the design process as well as a careful introduction by using pilots and organizing presentations in meetings made that the users accepted the system rather easily. Intermediate evaluations resulted in small changes of the assessment system before the final version, as described in chapter 7, was launched.

Coach / assessor and trainee in ACC simulator training at LVNL.

Radar screen in ACC simulator training at LVNL.

PART 3: EVALUATION

The evaluation part is guided by the Program of Requirements (PoR; see chapter 6) and comprises the following parts:

- Evaluation of *psychometric quality* (reliability, validity): chapter 9
- Evaluation of learning processes (feedback, pass-fail decisions):
 - Analysis of *learning curves*: chapter 10
 - Analysis of *competence development*: chapter 11
- *User evaluation* (acceptance, practical usability): chapter 12

The scheme below illustrates the relations of the evaluation parts with the requirements and constraints of the PoR. We should notice that three constraints of the PoR are not included in the scheme because they refer to the quality of the assessment system as a whole: C9 (benchmarking), C10 (continuous evaluation) and C11 (quality assurance). These constraints are included in the overall evaluation against the PoR, described in chapter 13, which covers the PoR completely.

Psychometric quality (CH9)	Learning processes (CH10-11)	User evaluation (CH12)
Reliability -R10: Objectivity -R16: Role of assessor/coach -C2: Reliability	Assistance of learning: -R3: Competences -R4: Learning processes -R6: Follow trainee's progress -R11: Continuous assessment -R14: Training redesign	Acceptance: -R13: Culture -C1: Consequences of redesign -C6: Acceptance -C7: Confidentiality -C8: Input of controllers
Validity: -R1: Quality of controllers -R12: External influences -C3: Validity	Pass-fail decisions: -R2: Pass rate -R5: Pass-fail decisions -R9: Level of performance -R17: Integration of selection and training -R18: Personnel shortage	Practical usability: -R7: Clearness -R8: Consistency -R15: Optimal construction -C4: Practical usability -C5: Transparency

9 Evaluation of psychometric quality

Since psychometric quality is a prerequisite for any assessment system (see also chapter 4; section 4.2), we started the evaluation with investigating reliability and validity as included in several requirements and constraints of the Program of Requirements (see chapter 6). Within the aviation domain a substantial amount of research has been done on reliability and validity of performance ratings in assessment of aircrew (CRM). See for instance, the special issue of *International Journal of Aviation Psychology* (Brannick, Prince & Salas, 2002; Goldsmith & Johnson, 2002; Holt, Hansberger & Boehm-Davis, 2002; Mulqueen, Baker & Dismukes, 2002; O'Connor et al., 2002). The scarce literature on ATC, in which comparable performance assessment systems are applied, pertains to job samples used in selection systems (e.g., Manning, 2000; Eissfeldt, 2002). Unfortunately, there is no research on the reliability and validity of performance assessment systems in ATC training which we can compare our findings with.

This chapter describes how we investigated the following reliability issues⁷: interrater agreement (section 9.1), rating errors (section 9.2), and test reliability (internal consistency, split-half reliability; section 9.3). We should notice that this chapter does not discuss criterion-related validity, although this actually belongs to the psychometric quality of the assessment system. This involves learning processes over time and is therefore subject of chapters 10 and 11.

In this chapter, we present analyses of data from ACC preOJT (simulator training) and ACC OJT, collected in the period January 2001 till December 2006. In some cases, where analyses are restricted to simulator tests, we added data from other courses in Unit Training to get more evidence. This is justified because simulator tests are executed in the same way for all courses, and there are no reasons to expect systematic differences. The structure of the data used is rather complex. A complete description of the available training data, comprising ACC and other courses, is given in Appendix C.

9.1 Interrater agreement

According to Guion (1998) and others interrater agreement refers to the degree of agreement between ratings of different assessors, or the interchangeability of ratings

⁷ Parts of this chapter are published in two conference papers (Oprins, Burggraaff & Van Weerdenburg, 2006a, 2006b) and a journal article (Oprins, Burggraaff & Van Weerdenburg, 2008).

between assessors. Low interrater agreement would imply unreliability as the measure of one assessor cannot be replicated by another assessor. Interrater agreement should be distinguished from interrater reliability. Interrater reliability refers to the extent in which multiple assessors put the ratings of different trainees in relatively the same order (Guion, 1998), but it does not provide information about agreement in level of performance. It should be noted that agreement in performance level is more important for assessment in training than agreement in the rank order of ratings, because it has consequences for feedback, interventions and pass-fail decisions in training. This section examines the interrater agreement between assessors by answering the question: "To which extent are the ratings between assessors interchangeable?"

Analysis of interrater agreement is often part of the evaluation of assessment systems. Within the aviation domain, interrater agreement has been examined for assessment systems designed for CRM-training of aircrew (Mulqueen, Baker & Key Dismukes, 2002; O'Connor et al., 2002; Goldsmith & Johnson, 2002; Brannick, Prince & Salas, 2002; Holt, Hansberger & Boehm-Davis, 2002). In this kind of research, a group of assessors observes the same target (e.g., videotapes of aircrew performance) while they independently rate a set of dimensions that are represented by behavioural markers. Interrater agreement is often examined in calibration sessions, which aim at bringing the assessors together in their judgments (Goldsmith & Johnson, 2002). The dimensions with low agreement are discussed and reviewed if necessary. The dimensions in these studies are quite comparable to the competences rated in our assessment system, such as *situational awareness* and *decision making*. Variations of behavioural anchored scales are commonly applied, like in our assessment system.

Interrater agreement is an important requirement for the quality of our assessment system. While multiple assessors are involved in simulator tests, only one assessor fills in a progression report in preOJT and OJT, based on a period of coaching (see chapter 7). As Brannick, Prince and Salas (2002) emphasize, often only one or two assessors are available in practice, but assessment should not depend on the specific person who assesses. If the interrater agreement were sufficient, a small number of assessors could be involved and this would lower the costs of training.

Furthermore, analysis of interrater agreement in singular competences that are represented by a set of performance criteria provides information about the quality of their definition and operationalization. If there exists too low agreement in specific competence ratings, their meaning should be better clarified to assessors or their formulation should be changed (e.g., Goldsmith & Johnson, 2002). We expect higher agreement between assessors for competences that can be measured more objectively, such as *safety* that rests on counting safety violations, than for competences that cannot be measured as objectively, such as *attention management* that should be inferred from behaviour. Clearer standards can be formulated for more objectively measurable competences, and assessment can be based on observed facts.

Interrater agreement can be expressed in three types of indices: *shape* provides information about the profile of ratings, *dispersion* indicates how widely ratings diverge from their mean, and *level* refers to average performance across items (Church, 1997; Youngstrom, Loeber & Stouthamer-Loeber, 2000). We examined

interrater agreement for all the three types of indices because they are all relevant for the interchangeability of assessors. The assessors' ratings should have a comparable profile (shape) with respect to the set of competences rated for the same trainees so that they all assign low ratings to some and high ratings to other competences. The assessors should also agree in the distance from the competence ratings to their means (dispersion) so that their use of low and high ratings is comparable. In training situations, the level appears to be even more important: trainees' performance level, rated by different assessors, should be rated in a comparable way, because decision making (e.g., pass-fail decisions) is based on this performance level that can be insufficient or sufficient. For these reasons, we examined interrater agreement in terms of shape, dispersion and level indices.

Methods

Data collection

To examine the interrater agreement we used 22 simulator tests made by 16 trainees at the end of the 7th week and at the end of preOJT (15th week) in 'real' ACC training (some trainees did more than one test). 13 different assessors assessed the tests. Because the number of cases is rather small, we present the results of the same analyses for the courses in the TWR/APP trajectory in a separate section, to allow for a comparison with ACC results.

During each test, two assessors observed trainees' performance, using observation checklists. Normally different assessor pairs assessed each trainee; trainees did not execute a simulator test simultaneously but at different times. Afterwards, the assessors started to fill in an individual test report independently of each other, based on their notes, by rating the same set of competences. Next, they discussed the differences between them, and filled in a final test report together (see also chapter 7). We used the assessors' individual test reports to calculate the interrater agreement between the two assessors for one trainee. Some tests were not usable because assessors did not follow the procedure.

The assessors rated a set of 12 competences on a 6-points rating scale. There was one exception: *safety* is the sum of observed safety violations (see chapter 7 and the test report in Appendix B). We recoded *safety* from an absolute scale to a 6-points scale, comparable with the other ratings, to be able to include this competence in the overall data analysis. As the raw scores were approximately normally distributed (not presented in this study), we may expect that recoding did not influence the results substantially. We calculated interrater agreement at two levels of aggregation:

- *Overall performance level*: expressed in two ways:
 - Unweighted sum of competence ratings (range: 1 to 6)
 - Weighted sum of competence ratings, expressed in a 'percentage' (maximum value of 100), although this is not a real percentage but partly based on an ordinal scale (see chapter 7).

- *Singular competence level*: ratings on each competence at the 6-points scale (range: 1 to 6)

Measures

We applied multiple methods to be able to examine the three indices of interrater agreement: *shape*, *dispersion* and *level*. The choice of the specific measures heavily depends on the data structure.

First, we calculated Pearson's correlations across the competence ratings between pairs of assessors for each trainee. They were averaged in order to achieve an overall index for interrater agreement (see also Brannick, Prince & Salas, 2002; Holt, Hansberger & Boehm-Davis, 2002). These correlations are called *q-correlations* instead of the conventional *p-correlations* (Church, 1997; Youngstrom, Loeber & Stouthamer-Loeber, 2000). Q-correlations are calculated over attributes (competences), while p-correlations are calculated over persons. We were not able to calculate the degree of agreement between multiple assessors across persons (trainees), because trainees had different assessors when executing the same simulator test.

We chose for *q-correlations* because they can measure profile similarity (shape, dispersion), they are scale-independent, and only two assessors generally were involved which makes it easy to correlate the pairs. However, *q-correlations* do not provide information about differences in level (Youngstrom et al., 1997), while this is important in training as we argued. Furthermore, *q-correlations* cannot be used to examine agreement on a single attribute (cf. a competence) while this is relevant as well: it is possible that the interrater agreement is higher for some competences than for others, for instance, because their meaning is not clear enough. Therefore, the next step was to search for measures that are sensitive for shape, dispersion and level and that can be applied for singular competences.

A specific index is often applied to examine the interrater agreement of behavioural marker systems (e.g., O'Connor et al., 2002): *within-group interrater reliability*, r_{wg} (James, Demaree & Wolf, 1984). This measure reflects absolute agreement in terms of shape, dispersion and level. Its equations are:

$$(1) \quad r_{wg} = 1 - (S_{x_j}^2 / \sigma_{EU}^2).$$

$$(2) \quad \sigma_{EU}^2 = (A^2 - 1) / 12$$

In formula 1, $S_{x_j}^2$ is the observed variance on X_j (single attribute), and σ_{EU}^2 is the variance on X_j that would be expected if all judgments were due to random measurement error exclusively. In formula 2, A corresponds with the number of response categories for X_j . The r_{wg} can be used if a group of assessors rates a single target on a variable having either a single item (e.g., performance level) or a set of items (cf. criteria) that measure the same construct (James, Demaree & Wolf, 1984). Goldsmith & Johnson (2002) and Mulqueen, Baker and Key Dismukes (2002) explained with examples from CRM assessment how r_{wg} measures absolute agreement among assessors while correlations serve only as a measure of relative agreement. However, the r_{wg} is not a suitable measure for our data structure: we do not have a group of assessors but only two assessors, and they do not assess a single target but dif-

ferent targets (trainees). Other (general) limitations of the r_{wg} are that the appropriate specification of the null distribution is debatable and that the values are scale-dependent since they are a function of the number of response categories; this makes the results difficult to interpret (James, Demaree & Wolf, 1984; Dunlap, Burke & Smith-Crowe, 2003).

Possible alternatives that better match with our data structure, that are sensitive for shape, dispersion and level information, and that can be applied for singular competences can be found in the so-called 'deviation indexes' for interrater agreement. Dunlap, Burke & Smith-Crowe (2003) proposed the *average deviation (AD) index* as an alternative for r_{wg} in order to meet the aforementioned limitations:

$$(3) \quad AD = \sum |X_{jk} - X_j| / N$$

Formula 3 expresses the absolute deviation of each rating from their mean of the group rating, and then averaging the results. X_{jk} is the k^{th} assessors' rating on attribute j and X_j is the mean of the assessors' ratings on attribute j . Like the r_{wg} , the AD -index is more suitable for groups of assessors larger than two, and this measure is difficult to interpret. Therefore we chose for a variant of the AD -index, that is, the *absolute difference (D-) index* (Church, 1997):

$$(4) \quad D = \sum |X_k - X_j| / N$$

Formula 4 implies the averaged sum of absolute differences between the ratings of assessors. Like the AD -index, this D -index is highly sensitive for dispersion, shape and level of profiles. We can interpret the D -index more easily for practical use despite of its scale-dependency. This D -index is also a variant of the *generalized distance (D^2 -) index* (Youngstrom et al., 2000):

$$(5) \quad D^2 = \sum |X_k - X_j|^2 / N$$

Formula 5 refers to the mean squared differences between ratings. However, like the AD -index the D^2 -index is also more difficult to interpret. Another variant is the *difference score* (Youngstrom et al., 2000):

$$(6) \quad D = X_k - X_j$$

Formula 6 represents the relative difference between ratings. This measure is not suitable because this difference score is sensitive to the level but not to the shape or dispersion of profiles.

In sum, we calculated *q-correlations* for shape and dispersion information about agreement in assessors' profiles of competence ratings. Next, we calculated *absolute difference (D-) indexes* (see formula 4) for additional level information about agreement between assessor pairs at several levels of aggregation: overall performance level, represented by the weighted and unweighted sum of competence ratings, and singular competences. We averaged the *q-correlations* and *D-indexes* for as-

sensor pairs to achieve overall indexes for interrater agreement. It is assumed that correlations between ratings of different assessors who assess performance of the same targets (trainees) should be in the order of .55 to .75 (Handyside, 1989; Ridderbos, 1992). This is in accordance with the interpretation of the findings by Brannick, Prince & Salas (2002), Holt, Hansberger & Boehm-Davis (2002) and Youngstrom et al. (2002).

The *D*-indexes are scale-dependent and therefore a minimum value for acceptance does not exist. From practical point-of-view, we consider a *D*-index of 5% to be acceptable for the overall performance level (weighted sum of competence ratings). This measure is used as the final test score, represented in a percentage scale. A deviation of 5% hardly influences (pass-fail) decisions that are based on this final test score. In practice, the most critical zone for sufficient performance is in-between 60% and 70% because the norm for sufficient performance is considered to be around 70%, but deviations of 5% do not have consequences for decision making (e.g., 62% or 67%). Deviations in lower or higher scores (e.g., 43% versus 48%, or 83% versus 92%) do not influence the picture about trainee's performance: the trainees just show very low or very high performance. The acceptable minimum of 5% can also be used for singular competence ratings for the same reasons, but a higher disagreement has less consequences.

Results ACC

Q-correlations

After having calculated bivariate Pearson correlations between the assessors' competence ratings on a 6-points scale, we averaged the correlations of the tests to achieve an overall index for the interrater agreement⁸. For averaging the correlations we transformed each correlation coefficient (*r*) to a *z*-score using Fisher's formula for *z*-transformation:

$$(7) \quad z = 0,5 \log_e [(1 + r) / (1 - r)].$$

This is generally recommended: averaging correlation coefficients leads to underestimation because their sampling distribution is skewed (Silver & Dunlap, 1987; Corey, Dunlap & Burke, 1998). Afterwards, we converted the average of the transformed *z*-scores back because it is in transformed form and not a correlation coefficient. We used the formula for back-transformation to *r*.

$$(8) \quad r = (e^{2z} - 1) / (e^{2z} + 1).$$

⁸ For a particular test, three instead of two assessors were involved. We calculated the correlations between the three assessor pairs and next we averaged the correlations to achieve one measure for interrater agreement for this test.

These calculations resulted in an average correlation coefficient (r_q) of .56 for a total number of 22 ACC tests. We consider this correlation as moderately high under the assumption that this correlation should be .55 or higher as mentioned before.

Some extreme outliers could have lowered the correlations. Therefore, we also examined the distribution of the correlations for each test. Figure 9.1 shows a histogram of the correlation coefficients. It confirms that they are rather normally distributed. Exact 50% of all correlation coefficients ($N=11$) is higher than .55. Correlations higher than .75 are found for 5 tests (=23%) and lower than .25 (=23%) for 5 tests as well. One value was below zero (-.15). We could say that agreement between some assessor pairs is rather high while agreement between other pairs is rather low; there exist many differences between pairs.

Figure 9.1. Histogram of correlation coefficients for ACC ($N=22$)

Absolute difference (D)-indexes

Table 9.1 presents the *D*-indexes for the overall performance level expressed in a weighted sum (in %) and unweighted sum of competence ratings (6-points scale), including the means and standard deviations, for 22 tests. We also transformed the *D*-index, being scale-dependent, for the unweighted sum from the 6-points scale into the same percentage scale (in %) to be able to compare the *D*-indexes with each other.

Table 9.1. Interrater agreement in absolute *D*-indexes for ACC ($N=22$)

Measure	<i>D</i> -index	Mean	Standard Deviation
Unweighted sum of ratings	.27 (5.5%)	4.2	.64
Weighted sum of ratings	4.5%	61.8	12.6

Table 9.1 shows that the D -indexes are rather low; they are respectively 42% of the standard deviations for the unweighted sum and 36% for the weighted sum of ratings. The D -index for the weighted sum of competence ratings is somewhat better than for the unweighted sum. The D -index of this weighted sum is lower than 5%. We consider this value for interrater agreement to be sufficient under the assumption that it should be 5% or lower as we argued. As for q -correlations, it is useful to examine the distributions of D -indexes in order to look for possible outliers. We restricted ourselves to the distributions of the D -indexes of the weighted sum of ratings, because this represents overall performance best. In contrast with the q -correlations (see Figure 9.1), these distributions are strongly skewed to the left and there are only two tests with an extreme high D -index (above 15%). Figure 9.2 presents the histogram of D -indexes for all tests.

Figure 9.2. Histogram of D -indexes for overall performance level for ACC (N=22)

Figure 9.2 shows that the majority of D -indexes (18 tests; 82% of the total number of tests) is below 5%.

Table 9.2 presents the D -indexes calculated for singular competences for 22 tests, rated at a 6-points scale, including the means and standards deviations. This table also includes the averaged D -index across competence ratings serving as an alternative overall measure for interrater agreement in addition to the measures presented in Table 9.1. Again we transformed this D -index to the same percentage scale to be able to compare the various D -indexes with each other. Table 9.2 indicates that the D -indexes for singular competences differ from .28 to .67. These are rather low D -indexes if we realize that the competences are rated on a 6-points scale. They are around 1/3 to 1/2 of the standard deviations. In general, the D -indexes of the various competences are rather comparable. This suggests that there do not exist competences with substantially more agreement than others. This does not support our aforementioned expectation that competences that can be measured more objectively would show lower D -indexes because assessors would agree more easily.

The findings show that the *D*-index for *safety* is relatively low (.36), but competences such as *attention management* and *workload management*, which cannot be measured as objectively, have low *D*-indexes as well (respectively .28 and .32). Finally, there is no clear relationship between *D*-indexes and lower or higher means and/or standard deviations. This implies that competences with relatively high ratings or variance do not have a lower or higher agreement than competences with relatively low ratings or variance. The total *D*-index, averaged across competences, seems to be higher than the *D*-indexes presented in Table 9.1: it is 9.5% if transformed into a percentage. This can be explained by the fact that in Table 9.2 the absolute differences for each competence were counted, while in Table 9.1 there is compensation between negative and positive values since only the final sum of competence ratings was used.

Table 9.2. *D*-indexes of interrater agreement for singular competence ratings and overall means for ACC (N=22)

Competences	<i>D</i> -index	Mean	Standard Deviation
Safety	.36	4.58	.97
Efficiency	.45	3.77	.95
Verbal expression	.46	4.31	.79
Listening	.55	4.65	.60
Co-ordination	.60	4.15	.82
Equipment operation	.67	4.52	.89
Strip management	.67	4.36	1.08
Mental picture	.39	3.78	1.02
Attention management	.28	4.13	.95
Planning	.41	3.82	.88
Decisiveness	.53	4.16	.91
Workload management	.32	4.18	.77
Total (mean)	.47 (9.5%)	4.20	.89

Results TWR/APP (summary)

Because the number of data of ACC simulator tests is too small to get sufficient evidence, we did the same calculations for the simulator tests applied in the courses of TWR/APP: GND, ARR and APP (see appendix C for the data structure). We may assume that the results would be comparable because the assessment methods, scoring techniques and the rated competences are the same in each course as we explained before. The main difference is that in ARR and APP all performance criteria are rated while in GND and in ACC the competences are rated directly. The TWR/APP courses have sometimes three assessors per test instead of two. GND has only one test; therefore, the number of trainees is similar with the number of tests administrated. Table 9.3 presents an overview of the various numbers per course.

Table 9.3. Overview of numbers of tests, trainees, assessors and ratings for each course in TWR/APP

Course	Tests	Trainees	Assessors per test	Ratings
GND	18	18	3	13
ARR	29	11	2 or 3	39
APP	18	9	2 or 3	28

Q-correlations

We calculated averaged Pearson correlations in the same way as explained for ACC. We calculated them per test by averaging the correlation for each assessor pair when three assessors were involved, and we calculated them for assessor pairs only. Table 9.4 presents the (averaged) correlation coefficients (r_q).

Table 9.4. Interrater agreement in average Pearson's correlations (r_q) for each course in TWR/APP

Course	N (tests)	r_q (averaged per test)	N (assessor pairs)	r_q (per assessor pair)
GND	18	.60	54	.60
ARR	29	.53	51	.48
APP	18	.46	38	.43

Table 9.4 shows that the correlation coefficients of the various courses are rather similar to each other and to ACC. It does not make much difference whether they are calculated for each test (averaged) or for assessor pairs. As argued for ACC, they are only moderately high. The values for ARR and APP are even below the required minimum of .55. This implies that the assessors' profiles are moderately similar with respect to shape and dispersion of the ratings on the different competences. Histograms of the correlations for the assessor pairs (not presented here) suggest that they are rather symmetrically distributed, comparable with ACC (see Figure 9.1); there do not exist extreme outliers. Around 25% of the assessor pairs have very low agreement (below .25), while also around 25% have very high agreement (above .75%). The interrater agreement of the remaining pairs, around 50%, can be considered as moderately.

Absolute difference (D)-indexes

Table 9.5 presents the results of calculating the D -indexes for the overall performance level, expressed in a weighted sum (in %) and unweighted sum (in a 6-points scale) of competence ratings. We only present the results calculated per test (averaged for three assessors if needed), not for assessor pairs; the D -indexes per test are very similar with the D -indexes for assessor pairs.

Table 9.5. Interrater agreement in absolute D -indexes for each course in TWR/APP

Course	N (tests)	d_w (weighted sum of ratings)	d_{uw} (unweighted sum of ratings)
GND	18	7.5	.38 (7.6%)
ARR	29	4.5	.29 (5.9%)
APP	18	4.7	.30 (5.9%)

Table 9.5 shows that the *D*-indexes are rather low, especially for the weighted sum of ratings. The results are very similar to each other again and to ACC. Except for GND ($SD = 13.3$), the *D*-indexes for the weighted sum of competence ratings are around 1/3 of the standard deviations of assessors' test versions (ARR: $SD = 13.5$, APP: $SD = 14.0$), which is rather low. Their values are lower than the acceptable value of 5% for ARR and APP. Like for ACC (see Figure 9.2), the distributions of *D*-indexes (not presented here) are strongly skewed to the left. The interrater agreement for GND is lower than for ARR and APP: only 17% of the GND tests have *D*-indexes lower than 5%, while this number is 68% for ARR and 56% for APP. Table 9.6 presents the *D*-indexes for singular competence ratings.

Table 9.6. *D*-indexes (in %) of interrater agreement for singular competences ratings for each course in TWR/APP

Competences TWR/APP	<i>D</i> -indexes GND (N=18)	<i>D</i> -indexes ARR (N=29)	<i>D</i> -indexes APP (N=18)
Safety	.37	.47	.25
Efficiency	.65	.43	.46
Communication	.60	.32	.35
Co-ordination	.91	.46	.60
Equipment operation	.12	.20	.26
Strip/label management	.69	.59	.49
Perception	.76	.40	.36
Attention management	.75	.51	.48
Mental picture	.72	.36	.40
Planning	.82	.47	.62
Decisiveness	.58	.43	.51
Workload management	.63	.38	.55
Total (mean)	.63 (12.7%)	.44 (8.8%)	.44 (9.0%)

Table 9.6 shows that the *D*-indexes vary between .12 and .91 while the total (averaged) *D*-indexes lie between .44 and .63. Comparable with the results presented in Table 9.5, GND has higher *D*-indexes for many competences in contrast with ARR, APP and ACC. Like for ACC, *safety* generally has relatively low *D*-indexes, but it is not clear why some competences have low and others have high *D*-indexes. If we would make a rank order from low to high *D*-indexes, this rank order would differ for each course. In sum, we conclude that both *q*-correlations and *D*-indexes of the three TWR/APP courses are rather similar to ACC, although the *D*-indexes for GND are somewhat higher.

Discussion

We applied several methods to investigate the interrater agreement in order to get information about agreement in level, shape and dispersion of profiles. The main question was to examine to which extent the ratings between assessors were interchangeable.

First, we calculated *q*-correlations (Pearson) between assessor's ratings across competences to compare the distributions of assessor's ratings with each other in

terms of shape and dispersion (Youngstrom et al., 2000). The results have shown that the q -correlations are only moderate; the overall (averaged) correlation coefficient for ACC is .56, and for other courses even lower. Although this is higher than the supposed minimum of .55 (Handyside, 1989; Ridderbos, 1992), a substantial number of assessor pairs have lower correlation coefficients. Findings in related domains confirm that we should consider our results only as moderate. Brannick, Prince and Salas (2002) found correlations between ratings of two assessors on a set of dimensions, comparable with the competences assessed in our study, which differed from .55 to .80 (average: .68). However, we should note that this research was done in a more structured (experimental) setting. Holt, Hansberger and Boehm-Davis (2002) found rather similar correlations in a related study: from .58 to .89 (average: .76). In sum, the findings imply that the assessors' profiles of competence ratings in our study, in terms of shape and dispersion, are only moderately comparable to each other: some assessors give low ratings on some competences while others give high ratings on the same competences for the same trainee, and vice versa.

Second, we calculated absolute differences (D)-indexes (Church, 1997) to obtain level indices that cannot be provided by q -correlations: how much assessors agree with each other. We calculated D -indexes at two levels of aggregation: the *overall performance level* (weighted and unweighted sum of competence ratings), and the *singular competence level*. The D -indexes for the overall performance level are rather low for each course, especially for the weighted sum of ratings: 4.5% for ACC. This value is lower than the acceptable maximum for practical use (5%), and therefore we consider these D -indexes as sufficient. In contrast with the findings for q -correlations, there are only some outliers with very high D -indexes (only two for ACC); the majority of assessor pairs have D -indexes for the weighted sum of ratings below 5%. This weighted sum of ratings is assumed to be a better measure for overall performance than the unweighted sum because more critical competences have a higher weight. That is the reason why the weighted sum of ratings is used as the final test score in practice. Apparently, assessors agree about their 'gut feeling' of trainee performance, expressed by the weighted sum of competence ratings that represent overall performance optimally, while the D -index for the unweighted sum of ratings is relatively higher. An alternative explanation for the higher agreement for the weighted sum could be that there exists more agreement about competences with higher weights. This was examined by calculating the D -indexes at singular competence level.

We found relatively higher D -indexes for singular competences. These findings agree with the results that follow from the q -correlations: some assessors give low ratings on specific competences while others give high ratings on the same competences and this leads to higher D -indexes. The D -indexes of singular competences provide additional level information above shape and dispersion, because they indicate how much the assessors agree with each other, in terms of the applied rating scale. We expected lower D -indexes for competences that can be measured objectively. This expectation was partly confirmed for ACC, because *safety*, the result of counting safety errors, has a relatively lower D -index. This can explain why the weighted sum of ratings has a lower D -index than the unweighted sum: *safety* has a relatively high weight in comparison with other competences. The D -indexes of com-

petences generally do not deviate much from each other. Rating errors, to be discussed in next section, could have played a role here. For instance, too high inter-correlations between the competence ratings are assumed to be an important cause for low interrater agreement (e.g., Brannick, Prince & Salas, 2002; Murphy & Cleveland, 1995). Differences in interrater agreement between dimensions (cf. competences in our study) were also found by others within the aviation domain. It is remarkable that O'Connor et al. (2002) found a large variation in agreement in pass-fail judgments, which was sometimes rather low, while the dimension ratings showed rather high agreement and low variation. Under the assumption that pass-fail judgments relate to overall performance we found opposite results in a certain way. However, the research was done in a different setting and the dimensions to be rated were not precisely the same as the competences rated in our study. Brannick, Prince and Salas (2002) also found differences in interrater agreement between dimensions but they were rather small, comparable with our findings.

It appears that low interrater agreement, both in terms of *q*-correlations and *D*-indexes, occurs more often with low performance than with high performance of trainees. This does not follow from the results presented in the previous section, but we examined this for individual trainees afterwards. Apparently, assessors agree easier when trainees perform well. It suggests that assessors have more difficulties with pointing at the specific deficiencies of trainees. These deficiencies mostly relate to underlying cognitive aspects that should be recognized from behaviour. They rate the competences differently from each other for trainees with low performance, but their general impression about trainee performance, expressed in the overall performance level, is rather similar. This may explain that the interrater agreement at singular competence level is lower than at the overall performance level. Furthermore, an explanation for some tests with very low interrater agreement could be that they were not reliable enough to use in our analysis. However, we deleted the unreliable reports from our data set as mentioned before. They were unreliable because some assessors did not fill in an independent version seriously because it took them too much time. Others had difficulties with following the trainee's actions adequately, because they were sitting behind another radar screen working as a second controller. Although they were assessor as well, they could have missed crucial situations.

There does not exist a clear relationship between the several measures for interrater agreement (*q*-correlations and *D*-indexes) when we compare the different courses to each other. For instance, GND has the highest *q*-correlation but also the highest *D*-index for overall performance, and tests with low *q*-correlations do not necessarily have high *D*-indexes and vice versa. This implies that assessors may differ in ratings on different competences (e.g., low ratings on some and high ratings on other competences), but the ratings may be compensated in such a way that the agreement about the overall performance level is sufficient. We cannot explain why differences between courses exist.

Conclusions

In conclusion, the findings show that the interrater agreement between assessors is sufficiently high for the overall performance level, optimally represented by the

weighted sum of competence ratings. The q -correlations and D -indexes at the singular competence level indicate that the interrater agreement is only moderate with respect to profile similarity (shape, dispersion) across competence ratings. With respect to the main research question, "to which extent are the ratings between assessors interchangeable?", we could say that the weighted sum of ratings, used as the final test score, is interchangeable between assessors but this is not true for singular competence ratings. Assessors seem to find it difficult to point at possible deficiencies of trainees, founded in competences. Improvement of agreement in competences should help to identify trainees' strengths and weaknesses more clearly. This is essential to provide adequate feedback and to support learning processes. Decision making (e.g., pass-fail) will probably not be influenced, because this is based on the overall performance level in which assessors agree better.

Differences between assessors may be caused by different views on the required performance standards, different meanings of the dimensions, the scenarios or tasks to be executed, or typical rating errors (Brannick, Prince & Salas, 2002; O'Connor et al. 2002). Solutions to improve interrater agreement are usually sought in better training, in calibration sessions to achieve more agreement in standards and in meanings of dimensions, and in changes of specific aspects of the assessment system (e.g., Brannick, Prince & Salas, 2002; Goldsmith & Johnson, 2002; Holt, Hansberger & Boehm-Davis, 2002; Murphy & Cleveland, 1995). Holt, Hansberger and Boehm-Davis (2002) found that the interrater agreement, expressed in correlations, substantially increased after more training. Because we did not found large differences in interrater agreement between competences, it was difficult to determine which of them were unclear and should be changed. But we organized an assessor training as a refresher after we did this study. This training deals with subjects such as avoidance of rating errors, main principles of assessment, and meaning of competences. A practical implication is that the involvement of multiple assessors had to be maintained to keep the assessments sufficiently reliable.

Finally, we should realize that assessment of behaviour in practical training settings remains difficult, especially if complex cognitive skills are involved. This is probably the main reason why interrater agreement appears to be not as high for all dimensions (Brannick, Prince & Salas, 2002; O'Connor et al. 2002). A certain degree of subjectivity is inevitable despite of the involvement of well-trained assessors, the choice of clear criteria and other measures to make the assessment system reliable. Yet, we have to keep striving for the highest possible interrater agreement.

9.2 Rating errors

Rating errors are tendencies of assessors to rate a set of items in a particular way. They are considered inadequacies of a rating system since they may affect accuracy in measuring performance (Murphy & Balzer, 1989). Their presence has been examined in various studies (Saal, Downey & Lahey, 1980; Murphy & Cleveland, 1995). Rating errors are assumed likely to occur in performance assessment systems within the aviation domain (Goldsmith & Johnson, 2002; Holt, Hansberger & Boehm-Davis,

2002; Hamman, Beaubien & Holt, 1999). A well-designed assessment system should minimize rating errors, for instance, by defining clear performance criteria and by using reliable rating techniques, because assessment may not be dependent on the specific person who assesses. We examined the research question: 'To which extent do rating errors occur?'

The most common classical rating errors in assessment of behaviour can be divided into *distributional errors* and *correlational errors* (Murphy & Cleveland, 1995; Roe & Daniels, 1994). Distributional errors are present when the distribution of ratings differs from the assumed distribution of performance ratings, while correlational errors refer to a too high relationship between separate aspects of a person's performance. We restrict ourselves to the best-known distributional errors: *leniency error*, *range restriction* and *central tendency*, and we discuss *halo error* as a form of correlational error. There exist many different conceptual and operational definitions of these rating errors, but they are nonetheless highly related to each other (Saal, Downey & Lahey, 1980; Murphy, Jako & Anhalt, 1993).

Leniency/severity

There are many studies on *leniency* or *severity errors*: personal tendencies of raters to provide respectively higher or lower ratings in a systematic way. Their operationalization differs. Two types of leniency measures are commonly used (Kane, Bernardin, Villanova & Peyrefitte, 1995): 1) measures that express leniency as a comparison between mean rating and a representation of the true mean rating, and 2) measures that express leniency as a comparison between mean rating and the skewness of the distribution of ratings. The assumption usually is that the true mean rating is located at the scale midpoint. However, this has been criticized because in practice the majority of rating distributions will be above the scale midpoint and will be negatively skewed (Saal, Downey & Lahey, 1980; Murphy & Cleveland, 1995; Kane, Bernardin, Villanova & Peyrefitte, 1995). This is certainly true in assessment in training, because we assume that only the competences that have not been achieved yet will be rated as insufficiently. Assessors will avoid insufficient ratings for keeping trainees motivated and self-confident.

Another method is to use a relative measure for leniency: a comparison of a rater's mean or skewness to the mean of other raters in the group (Kane, Bernardin, Villanova & Peyrefitte, 1995). This is even more relevant for assessment in training situations, because it is not desirable that some assessors are more lenient or severe than others. This hampers a valid judgment of trainees' performance and lowers the interrater agreement between assessors. Differences in leniency or severity between assessors, also called *systematic differences*, have been examined for assessment of CRM skills in aviation by identifying assessors who gave significantly higher or lower ratings than other assessors (Holt, Hansberger & Boehm-Davis, 2002; Hamman, Beaubien & Holt, 1999).

Range restriction and central tendency

Range restriction and *central tendency* are highly interrelated but they are not synonyms. *Range restriction* occurs when the variability of ratings is too small and therefore the discrimination among ratees in terms of their performance level is reduced.

Central tendency is defined as the tendency to avoid extreme judgments; they occur if the ratings cluster near the scale midpoint (Saal, Downey & Lahey, 1980; Murphy & Cleveland, 1995). Range restriction would be similar to central tendency if the mean ratings coincided with the scale midpoint, but this is usually not the case as was argued above. Central tendency implies range restriction but the converse is not true, because range restriction may reflect leniency, severity or central tendency (Saal, Downey & Lahey, 1980). Qualified assessment systems for training should make it possible to discriminate between trainees and competences by using a sufficiently wide range of the rating scale. Differences between assessors in range restriction or central tendency are undesirable because differences in the use and interpretation of rating scales can result in a lower interrater agreement (Holt, Hansberger & Boehm-Davis, 2002). Consequently, they may substantially affect (pass-fail) decisions and interventions, based on assessors' ratings. A specific measure to examine differences between assessors in the distribution of their ratings, developed for the assessment of aircrew, is called the *congruency index* (Goldsmith & Johnson, 2002; Holt, Hansberger & Boehm-Davis, 2002; Hamman, Beaubien & Holt, 1999). Congruency refers to the degree to which individual assessors distribute their ratings in comparison with some standard distribution (Goldsmith & Johnson, 2002).

Halo error

A number of definitions for *halo error* exist, but the different conceptual definitions are not related to different operational measures of halo systematically (Balzer & Sulsky, 1992; Murphy, Jako & Anhalt, 1993). The original definition for the halo error comes from Thorndike (1920): the tendency to think of a person in general as rather good or inferior and to colour the ratings on specific dimensions by this general feeling. As a result, intercorrelations between different dimensions of performance or behaviours are rather high (Saal, Downey & Lahey, 1980). Balzer and Sulsky (1992) suggest that the real operational definition of Thorndike (1920) would be high correlations between an overall performance rating with the ratings on specific dimensions, while intercorrelations between dimensions would better be referred to as the *logical error*: a rater's tendency to rate dimensions that are perceived as conceptually similar or logically related in a similar way (see also Roe & Daniels, 1994). Although the majority of researchers considers the logical error to be (a kind of) a halo error, Balzer and Sulsky (1992) distinguish two types of halo: *general impression halo* and *dimensional similarity halo*; the latter is assumed to be identical to the logical error.

The two operational definitions of halo are both relevant for assessment in training: assessors' impression of trainees is often coloured by the opinions of their colleagues, especially when a relatively small group of assessors and trainees is involved, or it is influenced by previous experiences in work or training situations (general impression halo). In addition, assessors usually find it difficult to distinguish between competences because they are conceptually related (dimensional similarity halo). Like distributional errors, halo errors are undesirable in our assessment system. Murphy, Jako and Anhalt (1993) suggest that the strengths and weaknesses in trainees' performance cannot be differentiated easily if intercorrelations between dimensions are too high. Discrimination between competences in our assessment

system is needed to provide adequate feedback and to detect possible deficiencies of trainees in training situations. Furthermore, assessors' judgments can be higher or lower than actual performance indicates because of halo errors. This makes assessments less reliable and valid.

Intercorrelations have also been calculated for the assessment of CRM skills of aircrew (Brannick, Prince & Salas, 2002), and for the assessment of ATC performance in a work sample used for selection (Manning, 2000). Manning (2000) correlated eight dimensions of ATC performance with each other, and these correlations were high. The dimensions were rated at a 7-points scale by over-the-shoulder observation (OTS), and they are comparable to our set of competences (e.g., 'maintaining separation', 'maintaining attention and situational awareness', and 'coordinating'). These dimensions were also correlated with dimensions of another instrument, the behavioural and event checklist. The instruments were clearly distinguished from each other because they did not correlate much, being not conceptually related (cf. dimensional similarity halo).

Methods

We used ACC simulator tests (assessors' versions (N = 50) or final versions (N = 36), depending on the specific analysis) and progression reports in preOJT (N = 188) and OJT (N = 407) to examine possible occurrences of distributional and correlational rating errors. We applied various operational measures for leniency and severity, range restriction, central tendency and halo, mostly based on Saal, Downey and Lahey (1980). Some measures were based on studies from the aviation literature to be able to allow for a comparison of our results (e.g., Holt, Hansberger & Boehm-Davis, 2002).

We examined *leniency/severity errors* in two ways. First, we investigated general tendencies of the total group of assessors by examining a possible shift in mean ratings from the midpoint of the rating scale in the positive or negative direction for the overall performance level (weighted sum of competence ratings; see section 9.1) and for singular competences to get an impression of the general distribution of assessors' ratings (Saal, Downey & Lahey, 1980; Murphy & Cleveland, 1995). We used the following formula:

$$(9) \quad D = \Sigma (X_j - 3.5) / N$$

Formula 9 expresses the mean rating X_j minus the scale midpoint, that is 3.5 at a 6-points scale, averaged per assessor for each competence, and next averaging the results across assessors. A positive value implies leniency and a negative value implies severity. However, we expect that the mean ratings are not at the scale midpoint. Second, we examined possible systematic differences between raters (Holt, Hansberger & Boehm-Davis, 2002; Hamman, Beaubien & Holt, 1999): the assessor's mean in comparison with the group mean in order to detect possible lenient and severe assessors. We did an analysis of variance (ANOVA) across assessors' ratings (Saal, Downey & Lahey, 1980; Murphy & Cleveland, 1995) at two levels of

aggregation: the overall performance level (weighted sum of competence ratings) and singular competences. We expect differences between competences: assessors would tend to give higher ratings to less critical and more learnable competences such as *label management* and *equipment operation*, while they would tend to give lower ratings to more critical and less learnable competences such as *mental picture* and *workload management*. When the means of the overall performance level were significantly different, we investigated which assessors deviated significantly from the group mean. Therefore, we did an independent *t*-test of the individual assessor's mean with the group mean (Hamman, Beaubien & Holt, 1999). We present the differences in means across assessors for the overall performance level in a graph to show how much the assessors differ and to which degree leniency (higher ratings) or severity (lower ratings) were involved.

Next, we examined *range restriction* by calculating the standard deviations of the competence ratings assigned to different trainees, averaged for each assessor (Saal, Downey & Lahey, 1980; Murphy & Cleveland, 1995). We already investigated *central tendency* in a certain way by presenting the differences between mean ratings and the scale midpoint with respect to leniency, but we examined in more detail how the ratings were distributed over the 6-points rating scale by presenting the mean proportions of ratings for each scale point. This illustrates how often the extreme scale points 1 and 6 were used. In order to examine possible differences between assessors in their distribution of ratings, we calculated a *congruency index* (CI). The formula given by Hamman, Beaubien and Holt (1999) is:

$$(10) \quad CI = 1 - \sum |P_i - P_g|$$

In formula 10, P_i refers to the relative proportion of the assessor's ratings occurring at that scale point and P_g refers to the relative proportion of the group's ratings occurring at that scale point. We used the proportions for each scale point (at a 6-points rating scale) across all ratings, averaged for each individual assessor across assessments for various trainees, to achieve an overall congruency index.

Finally, we examined the presence of *halo errors* by calculating the intercorrelations between competence ratings, being the most common method to examine halo (Saal, Downey & Lahey, 1980; Murphy & Cleveland, 1995; Murphy, Jako & Anhalt, 1993; Murphy & Balzer, 1989; Goldsmith & Johnson, 2002). We calculated the median correlation between competence ratings as a general indicator for halo (Saal, Downey & Lahey, 1980; Murphy & Balzer, 1989). In addition, we did a factor analysis (principal components) because this provides more information about the internal structure of competence ratings. The fewer components that emerge, the greater the halo. It is assumed that the emergence of only one component that explains a sizeable proportion of the variance reflects a maximum halo effect (Saal, Downey & Lahey, 1980). We expect that we can recognize the conceptually related competences in the components that emerge from the factor analysis. For instance, *verbal expression* and *listening* are conceptually related and would form a component, as well as *attitude* and *team orientation*. This would agree with the classification in the ATC Performance Model (see chapter 5).

Murphy and Cleveland (1995) stress that it is not possible to define a minimum value for concluding if rating errors are present since there exist too many conceptual and operational definitions of these errors. Contextual influences and specific properties of the design, such as the relationships between dimensions, also play a crucial role (Murphy, Jako & Anhalt, 1993). This implies that we should interpret the results in relation to the properties and purposes of the assessment system itself.

Results

Leniency/severity

Table 9.7 presents the relative deviations between the assessors' mean ratings and the scale midpoint as a measure for general tendencies of the total group of assessors.

Table 9.7. *Relative deviations between mean assessors' ratings (6-points scale) and scale midpoint*

Competences	Simulator tests (13 assessors, 50 reports)	PreOJT progr. reports (14 assessors, 188 reports)	OJT progr. reports (39 assessors, 407 reports)
Safety	1.01	.59	.82
Efficiency	.32	.93	.86
Verbal expression	.83	.76	.76
Listening	1.21	.96	.69
Co-ordination	.66	.96	.67
Equipment operation	1.13	1.10	1.24
Strip management	.88	.87	1.13
Mental picture	.31	.79	.76
Attention management	.65	.95	.79
Planning	.28	.94	.83
Decisiveness	.66	.78	.61
Workload management	.69	.86	.86
Attitude	-	1.32	1.50
Team orientation	-	1.20	1.23
Overall performance level	.75	.82	.80

Table 9.7 shows only positive deviations, both for the overall performance level and for singular competences. This implies that all mean ratings are above the scale midpoint, as expected. The values differ from .28 to 1.50. Considering the use of a 6-points rating scale, these values seem to be not very high, but this can be influenced by range restriction (see next paragraph). In accordance with our expectations, the relative deviations are the highest for the competences that are often rated as sufficiently and are considered as less critical for ATC performance: *equipment operation*, *attitude* and *team orientation*. The deviations for separate assessors and singular competences (not presented here) are almost never below zero, less frequently than 1% of the total number of mean ratings across assessors. The lowest value that

we found was only $-.50$. These results indicate that assessors tend to leniency, or simply that the mean ratings are above the scale midpoint.

Next, we examined systematic differences between assessors in leniency/severity by means of an analysis of variance (ANOVA). Table 9.8 presents the results. It shows that the differences across assessors for ACC simulator tests are not significantly different at level $p < .05$, but they generally are for progression reports in preOJT and OJT, both for the overall performance level and for singular competences. This suggests that some assessors are more lenient or severe than others, especially in filling in progression reports. For progression reports in preOJT the differences for *verbal expression* and *mental picture* are not significant at $p < .05$. It is not clear why this is true for these specific competences and not for others, for instance, *attention management* strongly relates to *mental picture* but their results differ. However, we should taken into account that Levene's test (not presented here) showed that the variances are not always similar: for two competences for simulator tests and for circa 50% of the competences for progression reports.

Table 9.8. *Variance analysis (ANOVA): differences in mean ratings across assessors*

Competence	Simulator tests (13 assessors, 50 reports; df1=12, df2=37)		PreOJT progr. reports (14 assessors, 188 reports; df1=13, df2=174)		OJT progr. reports (39 assessors, 407 reports; df1=38, df2=332)	
	F	Sig.	F	Sig.	F	Sig.
Safety	1.733	.099	1.950	.028	1.685	.009
Efficiency	1.164	.343	2.308	.008	2.375	.000
Verbal expression	1.463	.183	1.900	.033	2.123	.000
Listening	1.103	.386	3.782	.000	3.278	.000
Co-ordination	1.581	.140	2.458	.004	3.026	.000
Equipment op.	1.045	.431	3.131	.000	2.948	.000
Label management	.388	.959	3.667	.000	2.448	.000
Mental picture	.688	.752	1.499	.122	2.389	.000
Attention man.	1.105	.385	3.018	.001	3.391	.000
Planning	.442	.935	2.030	.021	4.081	.000
Decisiveness	.833	.617	3.062	.000	2.844	.000
Workload man.	2.139	.038	3.822	.000	2.521	.000
Attitude	-	-	4.849	.000	3.867	.000
Team orientation	-	-	2.486	.004	2.590	.000
Overall performance	.962	.500	2.712	.002	3.110	.000

We did a *t*-test (independent samples) between the individual assessors' ratings and the group ratings, only for the overall performance level, to examine in more detail how many assessors differ in means. We did this only for progression reports in preOJT and OJT, because we did not find significant differences across assessors for simulator tests. We present the means of the individual assessors in Figure 9.3 and 9.4 to illustrate the systematic differences between assessors (Hamman, Beaubien & Holt, 1999). The horizontal lines in Figures 9.3 and 9.4 present the group means.

Figure 9.3. Means of assessors in comparison with the group mean for preOJT progression reports

Figure 9.4. Means of assessors in comparison with the group mean for OJT progression reports

The assessors in red colour differ in their means from the group mean ($p < .05$) significantly as proved by the t -test. We see that four assessors in preOJT and seven assessors in OJT have significantly lower means than the group mean and that one assessor in preOJT and twelve assessors in OJT have significantly higher means than the group mean. The absolute differences with the means generally are not very large, but rather a high number of assessors differ with the others significantly: 36% of the assessors in preOJT and 49% in OJT. This implies that assessors vary in their mean competence ratings. The findings point at possible severity and leniency errors of assessors, although this also depends on the trainees who are assessed by the different assessors.

We examined whether there is a relationship between trainee performance and the specific assessors who have a significantly lower or higher mean by calculating the percentage of assessments of failed versus passed trainees. In preOJT the total percentage of assessments of passed trainees is 77%. This percentage is only 62% for the four assessors with a significantly lower mean (we did not calculate this for the exceptional assessor with a significantly higher mean because he has made only three assessments). In OJT the total percentage of assessments of passed trainees is 65%. This percentage is only 41% for the seven assessors with a significantly lower mean, but 91% for the twelve assessors with a significantly higher mean. This suggests that the trainees were not equally divided over the assessors, especially not in OJT, and this probably may have influenced the results presented in Figures 9.3 and 9.4.

Range restriction and central tendency

Table 9.9 presents the standard deviations, averaged across assessors, for the overall performance level and for singular competences, for examining range restriction. The standard deviations presented in Table 9.9 are rather low, for singular competences and for the overall performance level. The findings presented in Table 9.9 suggest that only a small range of the 6-points rating scale generally is used, especially for progression reports in preOJT and OJT. The standard deviations tend to be lower for competences that are less critical and more often rated as sufficiently, such as *equipment operation* and *attitude*, although the differences with the standard deviations of other competences are not very high. This points at assessors' tendency to range restriction.

These findings are supported by the mean proportions of each scale point across competence ratings for simulator tests and progression reports preOJT and OJT, presented in Figure 9.5. Figure 9.5 shows clearly that especially the extreme low scale points (value 1 and 2) and the highest option (value 6) are hardly used in comparison to the other options. This also points at possible range restriction errors in assessors to ratings. We see that the ratings in simulator tests are more widely distributed than in progression reports, because the options 2, 3 and 6 are chosen more frequently. In addition, Figure 9.5 suggests that central tendency is not present: scale points 4 and 5 are chosen most often in contrast with the scale points that refer to insufficient performance (values 1, 2 and 3). This confirms the results presented in Table 9.7 with respect to leniency: the mean ratings generally are above the scale midpoint in general.

Table 9.9. *Standard deviations of ratings, averaged across assessors*

Competences	Simulator tests (13 assessors, 50 reports)	PreOJT progr. reports (14 assessors, 188 reports)	OJT progr. reports (39 assessors, 407 reports)
Safety	.89	.66	.56
Efficiency	1.00	.37	.51
Verbal expression	.70	.74	.61
Listening	.45	.58	.53
Co-ordination	.69	.38	.50
Equipment operation	.78	.34	.37
Strip management	1.10	.53	.52
Mental picture	1.02	.66	.55
Attention management	.92	.73	.49
Planning	.98	.70	.53
Decisiveness	.88	.58	.60
Workload management	.73	.74	.58
Attitude	-	.55	.49
Team orientation	-	.46	.38
Overall performance	.75	.51	.43

Figure 9.5. Mean proportions of scale points across competence ratings

Finally, we calculated congruency indexes for comparing differences in distributions between assessors. Table 9.10 presents the results for the three types of assess-

ments in ACC. The proportions for each scale point (at a 6-points rating scale) across all competence ratings are averaged for each individual assessor. We also present the minimum and maximum values for the congruency indexes per assessor.

Table 9.10. *Congruence indexes and minimum/maximum values per assessor*

	N (assessors)	Congruency (CI)	Minimum	Maximum
SIM-tests (N=50)	13	.89	.83	.94
PreOJT progression reports (N=188)	14	.87	.84	.94
OJT progression reports (N=407)	39	.88	.82	.94

Table 9.10 presents very high congruency indexes. We see that assessors do not vary much in CI's, since the overall minimum is .82 and the maximum is .94. This suggests that assessors use the ratings scales in a very similar way, and that they do not differ much in their distribution of ratings across assessments. It implies that there do not exist assessors who show more range restriction or central tendency than others.

Halo error

The occurrence of the halo error was examined by calculating correlations (Pearson) between the competence ratings of assessors. For progression reports, we averaged the ratings for each trainee over time. The ratings within one trainee will be highly correlated (cf. 'low' and 'high' performers), although the progression reports were filled in by different assessors and at subsequent moments of time⁹. Table 9.11, 9.12 and 9.13 summarize the results for respectively ACC simulator tests (only the final versions, filled in by the assessors together), progression reports preOJT and progression reports OJT. Tables 9.11 – 9.13 present rather high intercorrelations. The median correlations are .36 for simulator tests, .71 for progression reports in preOJT and .60 for progression reports in OJT. This implies that especially in progression reports (preOJT and OJT) the competences ratings are highly correlated with each other. In general, the highest correlations were found between the competences *safety*, *efficiency*, *attention management*, *mental picture*, *decisiveness*, and *workload management*. This might point at possible halo errors. However, the intercorrelations varied much between competences, especially in simulator tests: from -.30 to .75. The range of correlations for progression reports in preOJT is from .42 to .94, and in OJT from .13 to .85. This suggests that some specific competence ratings certainly were distinguished sufficiently. For instance, lower (not significant) correlations in simulator tests were found between *verbal expression* and the competences *efficiency*, *co-ordination* and *planning*. Some competences are even negatively correlated, for instance, *listening* with *safety*, *efficiency*, *mental picture* and *planning*. Es-

⁹ The data structure can be considered to be a multilevel structure (Hox, 2002): repeated and longitudinal measures (competence ratings) for a group of cases (trainees). We did not apply multilevel analyses here because this would be very complicated in a technical sense for the specific purposes of this study, and we are not interested in the possible influence of other variables at a higher level: usually an important goal of multilevel analysis.

pecially in OJT, *strip management*, *equipment operation* and *attitude* do not correlate as high with all other competences. Furthermore, we should realize that the averaged ratings for each single trainee for progression reports could have made the intercorrelations for progression reports even higher.

Table 9.11. *Correlations (Pearson) between competences for simulator tests (N=36).*

	Safety	Efficiency	Verbal expression	Listening	Co-ordination	Equipment operation	Strip management	Mental picture	Attention management	Planning	Decisiveness	Workload management
Safety												
Efficiency	.60											
Verbal expression	.31	.06										
Listening	-.03	-.30	.35									
Co-ordination	.29	.43	.18	.31								
Equipment operation	.45	.33	.39	.29	.20							
Strip management	.27	.26	.49	.34	.28	.31						
Mental picture	.62	.70	.22	-.05	.55	.35	.34					
Attention man.	.60	.62	.31	.22	.37	.55	.43	.63				
Planning	.63	.75	.10	-.14	.34	.44	.30	.66	.65			
Decisiveness	.55	.52	.26	.23	.33	.36	.35	.58	.52	.58		
Workload man.	.43	.58	.59	.14	.38	.55	.53	.47	.64	.47	.48	

$r > .30$: significant at the 0.05 level; $r > .38$: significant at the 0.01 level; (1-tailed)

Table 9.12. *Pearson correlations between competences for progression reports preOJT (N=34).*

	Safety	Efficiency	Verbal expression	Listening	Co-ordination	Equipment operation	Strip management	Mental picture	Attention management	Planning	Decisiveness	Workload management	Attitude	Team orientation
Safety														
Efficiency	.84													
Verbal expression	.65	.46												
Listening	.78	.84	.69											
Co-ordination	.83	.89	.49	.84										
Equipment operation	.59	.58	.53	.67	.58									
Strip management	.70	.72	.61	.85	.74	.54								
Mental picture	.86	.90	.57	.85	.83	.56	.75							
Attention management	.83	.92	.53	.85	.89	.50	.73	.92						
Planning	.81	.77	.54	.73	.70	.49	.58	.80	.72					
Decisiveness	.85	.91	.60	.87	.82	.63	.73	.92	.91	.82				
Workload management	.89	.90	.65	.86	.86	.59	.80	.94	.93	.79	.92			
Attitude	.61	.53	.71	.74	.55	.76	.61	.54	.49	.59	.58	.62		
Team orientation	.52	.48	.56	.69	.49	.64	.57	.46	.42	.57	.61	.53	.77	

All correlations are significant at the 0.01 level (1-tailed)

Table 9.13. *Pearson correlations between competences for progression reports OJT (N=27).*

	Safety	Efficiency	Verbal expression	Listening	Co-ordination	Equipment operation	Strip management	Mental picture	Attention management	Planning	Decisiveness	Workload management	Attitude	Team orientation
Safety														
Efficiency	.70													
Verbal expression	.64	.65												
Listening	.78	.71	.85											
Co-ordination	.61	.45	.76	.74										
Equipment operation	.62	.53	.23	.27	.14									
Strip management	.60	.52	.32	.41	.15	.84								
Mental picture	.70	.80	.45	.61	.42	.54	.51							
Attention man.	.76	.70	.65	.84	.74	.23	.36	.68						
Planning	.69	.77	.58	.65	.67	.37	.31	.69	.78					
Decisiveness	.59	.65	.80	.77	.82	.05	.21	.33	.69	.66				
Workload man.	.80	.81	.65	.72	.57	.41	.53	.73	.79	.77	.60			
Attitude	.28	.13	.16	.18	.38	.48	.40	.19	.17	.42	.24	.17		
Team orientation	.63	.44	.58	.57	.62	.70	.62	.53	.53	.62	.48	.59	.73	

$r > .40$: significant at the 0.05 level; $r > .51$: significant at the 0.01 level; (1-tailed)

In addition, we examined the internal structure of ratings with a factor analysis (principal components, Varimax rotation). For progression reports, the competence ratings were averaged for each trainee similar to the calculation of correlations. The results are presented in Table 9.14 – 9.16:

Table 9.14. *Rotated component matrix of competences rated in preOJT simulator tests (N = 36)*

	Component	
	1	2
Efficiency	.92	
Planning	.88	
Mental picture	.84	
Safety	.76	
Attention management	.74	
Decisiveness	.66	
Workload management	.60	
Coordination	.45	
Listening		.80
Verbal expression		.77
Label management		.66
Equipment operation		.52

Table 9.15. *Rotated component matrix of competences rated in preOJT progression reports (N = 34)*

	Component	
	1	2
Attention management	.94	
Efficiency	.92	
Mental picture	.91	
Workload management	.88	
Coordination	.86	
Decisiveness	.86	
Safety	.82	
Listening	.74	
Planning	.73	
Label management	.66	
Attitude		.90
Team orientation		.85
Equipment operation		.75
Verbal expression		.71

Table 9.16. *Rotated component matrix of competences rated in OJT progression reports (N = 27)*

	Component		
	1	2	3
Attention management	.89		
Workload management	.86		
Listening	.85		
Mental picture	.84		
Planning	.81		
Efficiency	.80		
Decisiveness	.78		
Verbal expression	.76		
Safety	.73		
Coordination	.73		
Label management		.89	
Equipment operation		.86	
Team orientation			.71
Attitude			.90
Attention management	.89		
Workload management	.86		

The results presented in Tables 9.14 – 9.16 show that two or three components emerged (eigenvalues over 1). For ACC preOJT simulator tests (only final versions), two components emerged, accounting for 62.1 percent of the total variance (39.9 for component 1 and 22.3 for component 2). For ACC preOJT progression reports, two components emerged, accounting for 82.9 percent of the total variance (52.6 for component 1 and 30.1 for component 2). For ACC OJT progression reports, three components emerged, accounting for 84.1 percent of the total variance (48.8 for component 1, 19.1 for component 2, 16.2 for component 3). The results support the findings of the intercorrelation matrixes (Tables 9.11 – 9.13). The majority of competences, as part of component 1 in the factor analyses, are highly intercorrelated. Exceptions exist for competences that are less critical or less related to ATC perform-

ance: *label management, equipment operation, attitude, team orientation, listening and verbal expression*, although differently for simulator tests or progression reports in preOJT and OJT. The high variances found for the first components could point at a possible halo error, but the emergence of the other components could also be caused by differences in the meaning of competences, for instance, why *attitude* and *team orientation* form a separate component.

Discussion

We examined the presence of possible rating errors as part of the evaluation of the psychometric quality of the assessment system. It is important that these rating errors are minimized by the way of designing the assessment system, because they may affect reliability and validity of assessment. We examined the research question: 'To which extent do rating errors occur?'. We may conclude that some distributional and correlational rating errors occur (Murphy & Cleveland, 1995; Roe & Daniels, 1994).

Leniency/severity

We investigated general tendencies for leniency/severity for the total group of assessors by calculating the relative deviations between the assessors' mean ratings and the scale midpoint. The findings, supported by the mean proportions of each scale point used in the ratings, indicate that assessors tend to give relatively high ratings above the scale midpoint. This may be explained by the tendency to leniency of assessors, but only under the assumption that the mean ratings should be around the scale midpoint. Saal, Downey & Lahey (1980) and Kane, Bernardin, Villanova and Peyrefitte (1995) argued that some assessment systems, such as applied for personnel selection and training, are designed to produce a skewed distribution of performance. For instance, the purpose is to have excellent personnel and not to have only personnel who show average performance. These arguments are certainly true for our assessment system: we expect ratings above the scale midpoint because insufficient ratings are only given if the assessor found some deficiencies in performance. Trainees' performance is not compared with average performance expressed as the midpoint on the scale, but individual performance is assessed against standards. This relates to 'criterion-referenced' assessment, in which standards are used, as opposed to 'norm-referenced' assessment, which involves between-persons comparisons (Landy & Farr, 1983; Murphy & Cleveland, 1995; see section 4.2). In addition, assessors tend to keep the trainee motivated and self-confident. If trainees have many (unnecessary) insufficient ratings, they will become less self-confident which may negatively affect their performance. In sum, we conclude that assessors tend to give higher ratings (leniency), but we do not consider this as a rating error because it was expected from the properties of assessment in training situations.

Keeping in mind that all assessors tend to give relatively high ratings, it is even more interesting to examine possible differences between assessors in leniency/severity, also called *systematic differences* (Hamman, Beaubien & Holt, 1999; Holt, Hansberger and Boehm-Davis (2002). These differences are undesirable be-

cause judgments of trainee performance should not be dependent on the specific person who is the assessor. The results of an analysis of variance (ANOVA) showed that assessors have significantly different mean ratings for progression reports in preOJT and in OJT, but not for simulator tests, both for the overall performance level and for singular competences. The results of a *t*-test (independent samples) of individual assessor's mean ratings with the group mean ratings indicated that around 50% of the assessors give significantly lower or higher mean ratings than the group. Holt, Hansberger and Boehm-Davis (2002) found a range in different samples from 60% to 0% (average of 25%). They strive for 10% or less assessors who provide ratings that are significantly lower or higher than the group, and they suggest that this would be achievable by means of extensive assessor training. From this point of view, systematic differences in leniency or severity would be too large in our case. This suggests that there exist so-called 'Santa Claus' assessors (high mean ratings) as well as 'Axeman' assessors (low mean ratings) in the total group of assessors (Hamman, Beaubien & Holt, 1999; Holt, Hansberger and Boehm-Davis, 2002). More clarification of the performance standards and criteria and extensive assessor training can reduce differences in leniency and severity between assessors (Goldsmith & Johnson, 2002; Holt, Hansberger and Boehm-Davis, 2002).

However, the differences between the assessors' means and the group mean generally are not very high, only for some individual assessors. Furthermore, the findings are dependent on actual performance of the ratees (Murphy & Cleveland, 1995); assessors can be more lenient if the ratees show relatively high performance. The results confirmed that severe assessors have assessed relatively low performers (failed trainees) and that lenient assessors have assessed relatively high performers (passed trainees), especially in OJT. Trainees in OJT are assigned to a team of coaches who are their assessors; the trainees are not equally divided over the coaches (see chapter 7). In addition, we designed the assessment system in such a way that multiple assessors are involved, and this compensates for possible differences in leniency or severity. In sum, we conclude that individual differences in leniency do not cause problems for the objectivity of assessments.

The fact that systematic differences in leniency only exist for progression measurements confirms that simulator tests are more objective and reliable, caused by well-balanced test scenarios, clear performance criteria and detailed scoring techniques. Personal tendencies of assessors may not get as much chance to occur in these objective measurements as in more subjective progression measurements over time. Finally, the findings suggest that interrater agreement, only calculated for simulator tests (see section 9.1), could not have been lowered by differences in leniency or severity of the assessors.

Range restriction and central tendency

Examination of the standard deviations, averaged for assessors, suggests that assessors generally show range restriction: they do not distribute their ratings equally over the 6-points rating scale. The mean proportions of scale points across competence ratings support these findings: assessors do not use the extreme low and high scale points. These results relate to the finding that assessors tend to provide relatively high ratings: assessors use the scale points directly above the scale midpoint,

values 4 and 5, most frequently. The same explanation as mentioned with respect to leniency can be given: assessors tend to assess trainees as sufficiently unless trainees show deficiencies. Especially the first two scale points (value 1 and 2) carry a very negative meaning and may affect the trainee's self-confidence. But this does not explain why assessors tend to use only three scale points instead of six on the six-point rating scale. Sufficient variance is important for discriminating between ratees in terms of their performance levels (Saal, Downey & Lahey, 1980). The detection of deficiencies during training will be more clear if the differentiation between scale points is larger, and progression on specific competences can be expressed better if a wider range of scale points is used (e.g., progression of ratings from scale point 3 to 5). Therefore, the range restriction should be lowered. The central tendency error is not present since the average ratings generally are above the scale midpoint.

Again, these errors are less likely to occur in simulator tests, probably because tests are more reliable. Progression measurement over time can make it more difficult for assessors to use the extreme scale points, for instance, because possible mistakes occur only in a specific situation at one day. Assessors might give moderate judgments based on general performance during the whole period. A possible solution for the range restriction problem would be to change the rating scale into a scale with fewer scale points, for instance, a 4-points scale. For learning purposes we consider it very important that assessors will differentiate better and we prefer to stimulate assessors to distribute their ratings better over the rating scale. Therefore, we kept the 6-points rating scale in the system.

Differences between assessors in their distribution of ratings can affect their judgments of trainees in the same way as systematic differences in leniency. We calculated *congruency indexes* (Goldsmith & Johnson, 2002; Holt, Hansberger & Boehm-Davis, 2002; Hamman, Beaubien & Holt, 1999). We found very high indexes for the overall performance level, based on proportions for each scale point across competences ratings: between .87 and .89 for simulator tests and progression reports in preOJT and OJT. Holt, Hansberger and Boehm-Davis (2002) found a range from .67 to .82; thus, the congruency indexes in our study are higher. Holt, Hansberger and Boehm-Davis suggested that their indexes are already very high. They found some exceptions of individuals who had an inadequate understanding of the rating scale or the use of standards for making the ratings, but the assessors in our study all have high indexes. Therefore, we may conclude that the assessors generally do not differ in their distribution of ratings.

Halo error

We examined the presence of the halo error by calculating intercorrelations (Pearson) between the competence ratings. The results showed that the intercorrelations are rather high for progression reports in preOJT and OJT (median respectively .71 and .60) but not for simulator tests (median .36). In comparison with findings in other studies in related domains, these intercorrelations are not extremely high. Brannick, Prince & Salas (2002) found a range from .46 to .93 (mean .76) for correlations between dimensions of CRM skills. Manning (2000) found a range from .80 to .97 (median .91) for eight dimensions of ATC performance that are comparable with our

competences as mentioned before. Manning (2000) found negative correlations with other types of dimensions that have another meaning. Our findings are similar: competences that are conceptually different are low or negatively correlated with each other, such as *efficiency* and *listening* in simulator tests. The intercorrelations differ across competences.

In order to examine the internal structure and interrelations between competences in more detail, we did a factor analysis (principal components). Two components emerged for simulator tests and progression reports in preOJT, and three components for progression reports in OJT. The highest correlations were found between the competences that are most directly related to ATC performance. They generally form the first component in the factor analyses and reflect the classification of the ATC Performance Model (see chapter 5): they refer to the result of traffic handling (*safety, efficiency*) and the cognitive processes that are underlying traffic handling (*mental picture, attention management, planning, decisiveness, workload management*). The competences that are less critical for ATC performance are less correlated with the others and are part of the second or third component in the factor analyses such as *label management, equipment operation, attitude, and team orientation*. This suggests that some competences are correlated with each other because they are conceptually related and not because of the presence of halo. In case of halo, we probably would have found higher intercorrelations between all competences, for instance, also between the competences *mental picture* and *attitude*.

This latter problem refers to a more general issue concerning the interpretation of halo mentioned by various authors: the difference between true and illusory halo (Murphy, Jako & Anhalt, 1993; Murphy & Cleveland, 1995; Balzer & Sulsky, 1992). Ratings on separate dimensions should be correlated if they are related concepts (true halo), while illusory halo refers to the part of correlation that is due to cognitive distortion, deficiencies in measurement and observation, and memory errors on the part of the rater. It is assumed extremely difficult to separate these types of halo. Murphy, Jako and Anhalt (1993) suggest that the only way to separate true from illusory halo is in a laboratory situation, since in field studies there will always be an unknown mix of true and illusory halo. In our case, it is reasonable to assume that the competences as classified in the ATC Performance Model are not (only) be intercorrelated because of halo, but (also) because they are conceptually related to each other. Then we could say that the set of competences that form the first component in the factor analysis would represent the construct 'air traffic management'.

In addition, distributional errors, being personal tendencies of raters, may affect the presence of halo errors (Murphy & Cleveland, 1995; Balzer & Sulski, 1992), especially because the correlational measures are usually aggregated over multiple raters. For instance, if each rater shows extreme range restriction and raters differ in leniency, the intercorrelations will be high. This points at a possible halo error although the dimensions are uncorrelated for each individual rater (Balzer & Sulski, 1992). We found comparable results that may have increased the intercorrelations between competences: the assessors show range restriction and differ in leniency. Balzer and Sulski (1992) suggest that therefore intercorrelations should not be ag-

gregated over multiple raters. However, a higher number of ratees is needed than available in this study for examining this.

Furthermore, we should note that some authors raised the question to which extent halo errors would really affect the psychometric quality of assessment. Murphy, Jako and Anhalt (1993) argue that halo may even contribute to the accuracy and utility of ratings. They suggest that although halo error makes it more difficult to distinguish individual's strengths from his weaknesses, it could make it easier to distinguish between generally good or inferior individuals. Usual measures for reliability such as internal consistency will lead to better results if the dimensions are highly intercorrelated. Therefore, certain intercorrelations are likely to exist, especially between competences that are conceptually related.

In addition to these psychometric issues, we should realize that learning processes may have played a role. We argued that competence-based training implies that the whole of knowledge, skills and attitudes will become more integrated (see chapters 2 and 7). This suggests that the intercorrelations between competences will increase over time, and that they will be higher in OJT than in preOJT. However, our findings did not support this view. Another explanation might be that the differences between trainees are higher in OJT than in preOJT due to more varying learning tasks and differences in entrance level. We also expect differences between the competences over time, because some of them will be acquired and integrated earlier than others. We did not examine this because it was too complicated for investigating the presence of rating errors. Yet, time could have affected the results. Chapter 11 elaborates issues with respect to competence development further.

In a qualitative way, we believe that halo errors are certainly present in practice. General impressions about trainees, such as opinions of assessors and previous experiences, easily influence assessors. The group of assessors and trainees is very small and they are colleagues of each other. This makes it extremely difficult for assessors to form their own image of trainees' performance. This latter type of halo error relates better to the original definition of Thorndike (1920) or the so-called 'general impression halo', while the high intercorrelations between competences refer more to the 'dimensional similarity halo' (Balzer & Sulski, 1992).

In sum, we may assume that the presence of halo errors is not as high that they may affect the judgments of trainees substantially. They should be intercorrelated to a certain extent, influenced by learning processes. Nevertheless it remains important that assessors differentiate between competences sufficiently to be able to detect possible deficiencies of trainees.

Conclusions

In conclusion, some rating errors have been found to exist and may affect the reliability of the assessments. Assessors generally give ratings higher than the scale midpoint, but we don't consider this to be caused by the leniency/severity error. It is the result of assessment in training in which trainees' performance is compared to specific standards. Some assessors are more lenient or severe than others, although this also depends on which trainees they assess (cf. low versus high performers). Assessors show range restriction but not central tendency: they tend to use only

three scale points of the 6-points rating scale (values 3, 4 and 5), but they do not differ in their distributions of ratings. Halo errors appear to occur to a certain extent in progression reports. Some competences should be intercorrelated because they are conceptually related. In general, rating errors are present more often in progression reports than in simulator tests. This emphasizes that simulator tests are more reliable because of the use of multiple assessors, well-balanced test scenarios and detailed scoring techniques.

The presence of rating errors can be reduced by a critical review of the performance criteria and standards. Calibration sessions can contribute to more agreement in the interpretation of criteria and standards (e.g., Goldsmith & Johnson, 2002). Another solution is to provide extensive assessor training (Goldsmith & Johnson, 2002; Holt, Hansberger & Boehm-Davis, 2002; Murphy & Cleveland, 1995). Assessor training should make assessors more aware of the rating errors that occur most frequently and how these errors can be avoided. As mentioned before, we organized refresher training for the assessors.

9.3 Test reliability

A test should be constructed in such a way that the scores are reproducible and that one would get the same results when repeating the test (Drenth & Sijtsma, 1990). We could not repeat the same assessment in practical training settings (e.g., test-retest reliability), but we had other possibilities to estimate the reliability.

Internal consistency, measured by calculating Cronbach's alpha, refers to the degree in which all items in a scale or test measure the same concept (Drenth & Sijtsma, 1990; Wigdor & Green, 1991). It is often referred to as the homogeneity of a test. Cronbach's alpha will increase if the length of the test increases and if there exist highly intercorrelated items (Schmidt, 1996). These two issues have consequences for the interpretation of the results. Item-total correlations are often calculated to eliminate items that do not correlate with the assessed construct strongly (Clark & Watson, 1995). Internal consistency is relevant for the evaluation of our assessment system since a set of performance criteria should measure the same competence. A criterion that does not correlate with the competence sufficiently should be deleted. In the same way, Brannick, Prince and Salas (2002) examined this for a set of CRM dimensions for assessment of aircrew. The internal consistency cannot be determined for the set of competences as a whole, because the competences are considered to be different constructs (Murphy & Cleveland, 1995).

We cannot determine test reliability in the same way for simulator tests, because the competences are rated instead of the criteria (see chapter 7). In this case, a reliability estimate can be derived in another way. As explained in section 9.1, two or three assessors start to fill in a test report independently before they make a final test report together based on a critical discussion of discrepancies. We consider the reports of the individual assessors as parallel test versions. These versions can be used to estimate the overall test reliability by means of the Spearman-Brown reliability coefficient, which is based on correlations between half parallel test parts (Drenth

& Sijtsma, 1990). The results also provide information concerning the number of assessors that are needed for sufficiently reliable assessments.

Methods

For ACC progression reports, we calculated the Cronbach's alphas for each competence (e.g., Drenth & Sijtsma, 1990). We averaged the ratings for each trainee in preOJT ($N = 188$) and OJT ($N = 407$) because the ratings within one trainee will be highly intercorrelated. For the same data set, we calculated item-total correlations (r_{it}) for each competence, that is, the correlations between the criterion and the competence score (e.g., Clark & Watson, 1995). Various values have been proposed for acceptability of internal consistency reliability. Schmidt (1996) mentions a required level of Cronbach's alpha of .70, while Clark & Watson (1995) state that there are no longer any clear standards. Nunnally (1978) recommended minimum standards of respectively .80 and .90 for basic and applied research, but nowadays also reliabilities of .60 or .70 are accepted. In addition, Clark and Watson (1995) suggest that the (average) item-total correlation should fall in the range of .15 and .20 for a broader construct (e.g., extraversion) and in the range of .40 and .50 for a narrower construct (e.g., talkativeness). In contrast, Schmidt (1996) and others argue that we should not only rely on minimum values because influences of factors such as test length, intercorrelations and unidimensionality should also be considered.

For simulator tests, we estimated reliability coefficients under the assumption that the assessors' versions can be considered as parallel test versions, and that the scores are counted without applying differential weighting. We calculated the reliability coefficient with the Spearman Brown formula (e.g., Drenth & Sijtsma, 1990; Schmidt, 1996):

$$(11) \quad r_{KK} = K r_{xx'} / (1 + (K - 1)r_{xx'}).$$

In our case, we consider the average correlation coefficient between the assessors' individual reports and the final test report, which was filled in by the two or three assessors together, to be the best estimation of test reliability ($r_{xx'}$), because this comes closest to the 'true' test score. The reliability of the test as a whole (r_{KK}) will become higher when more assessors are involved; K involves the number of assessors.

Results

Internal consistency (progression reports)

Table 9.17 presents the Cronbach's alphas and the range of item-total correlations respectively for ACC progression reports preOJT and OJT, averaged for each trainee.

Table 9.17. Cronbach's alpha and item-total correlations for progression reports preOJT (N = 27) and OJT (N=19)

Competences	N (items)	Cronbach's alpha		Range of item-total correlations (r_{it})	
		PreOJT	OJT	PreOJT	OJT
Safety	4	.90	.88	.75 < > .81	.71 < > .87
Efficiency	6	.92	.88	.67 < > .87	.60 < > .79
Verbal expression	3	.90	.92	.73 < > .86	.83 < > .88
Listening	3	.96	.93	.91 < > .93	.79 < > .93
Co-ordination	4	.79	.93	.20 < > .79	.86 < > .96
Equipment oper.	2	.89	.86	.80 < > .80	.80 < > .80
Strip management	2	.70	.63	.56 < > .56	.47 < > .47
Mental picture	5	.93	.90	.74 < > .92	.41 < > .87
Attention man.	3	.97	.95	.95 < > .97	.87 < > .96
Planning	3	.79	.91	.54 < > .70	.72 < > .95
Decisiveness	3	.88	.87	.74 < > .80	.70 < > .86
Workload man.	2	.80	.83	.74 < > .74	.73 < > .73
Attitude	3	.98	.98	.95 < > .98	.94 < > .97
Teamwork ability	2	.94	.76	.90 < > .90	.62 < > .62

Table 9.17 shows that the Cronbach's alphas and the item-total correlations generally are very high. The majority of alphas is higher than .80 and a substantial number of alphas is even higher than .90. We should realize that we can expect lower alphas for competences with only two items. This is confirmed by the results for *strip management* in both preOJT ($r = .70$) and in OJT ($r = .63$). The item-total correlations of the two criteria that belong to this competence are also relatively low for preOJT ($r = .56$) and OJT ($r = .47$). In addition, we found two items that really have a too low item-total correlation and that must be deleted or changed. The first belongs to *coordination* in preOJT: "coordinates timely and with the appropriate ATC unit" ($r = .20$). This can be explained by the fact that the other items refer more to communication than to this time aspect of coordination. The second is part of *mental picture* in OJT: "Guards the identification process and the label presentation" ($r = .41$). This criterion refers more to the specific action than to the cognitive process of forming a mental picture of the traffic situation.

Test reliability (simulator tests)

First, we estimated test reliability (r_{xx}) by calculating the correlations (Pearson) between the assessors' individual versions, filled in independently, and the final test versions, filled in together. We averaged these correlations in order to get an overall measure for reliability for each course, after having applied Fisher's z- transformation (see section 9.1). Next, we estimated test reliability of the test as a whole (r_{KK}), based on the involvement of two assessors (K). We do not only present the results for ACC but also for other simulator tests in the same way as for examining interrater

agreement (see section 9.1) in order to increase the number of cases. Table 9.18 presents the results¹⁰.

Table 9.18. *Test reliability coefficients (r_{xx}), averaged for each course in ACC and TWR/APP*

Course	N (assessors)	r_{xx} (average)	r_{KK}
GND	54	.78	.88
ARR	60	.73	.84
APP	46	.66	.80
ACC	45	.80	.89

Table 9.18 presents rather high reliability coefficients (r_{xx}), although there are differences between courses. These findings imply that the contribution of each assessor to the final version is rather high. The reliability of final judgments (r_{KK}) for each course, in which the test is lengthened with a factor two in this case, is estimated to be higher than .80. The reliability will become even higher when three assessors are involved, due to the higher value of K , under the assumption that the scores are counted (unweighted).

Discussion

The results of the reliability analyses lead to the conclusion that the reliability of the assessment system is generally sufficient in terms of internal consistency for progression reports in preOJT and OJT and test reliability for simulator tests.

The Cronbach's alphas for each competence are very high, generally higher than the widely accepted minimum of .80 (Schmidt, 1996; Clark & Watson, 1995). They are higher than the alphas found for the same analyses performed for comparable CRM-dimensions by Brannick, Prince and Salas (2002). They found values that varied from .28 to .96. We found only four criteria with an item-total correlation below .50, which is assumed the minimum value for the narrowest construct (Clark & Watson, 1995). These items have to be deleted or moved to another competence in order to make the measurements more reliable. This will be done in a next version of the assessment system. The results suggest that the sets of performance criteria for each competence have been well chosen, because they measure the same constructs. However, we should realize that these results are dependent on the number of items and possible intercorrelations between items (Clark & Watson, 1995; Schmidt, 1996). The number of items (cf. performance criteria) for each competence is very low (only 2 to 6 per competence), thus this could hardly have affected the high Cronbach's alphas. In section 9.2, we showed that the intercorrelations between competences are high. Similar intercorrelations were found for the performance criteria, although we did not present the results. This also explains why the Cronbach's alphas are high.

¹⁰ Table 9.18 also includes the tests in which only one assessor's version was available. In that case, two assessors were involved, but one of them did not fill in his own report independently. Afterwards, they still filled in a final version together, thus these tests were usable for this purpose.

We estimated test reliability for simulator tests with the Spearman-Brown formula by calculating the correlations between the assessors' individual versions with the final test versions, filled in together. The resulting reliability coefficients are considered sufficiently high for various courses, not only for ACC (between .66 and .80 for four courses). Next, we used these reliability coefficients to estimate the reliability coefficient for the test as a whole under the assumption that the assessors' versions can be considered as parallel test versions (Drenth & Sijtsma, 1995). These are higher than .80 in all courses for the involvement of two assessors. Brannick, Prince and Sales (2002) calculated how many assessors they would need for a high coefficient for split-half reliability. They estimated the reliability to be .80 for one assessor, .89 for two assessors and .92 for three assessors. These results are rather similar to our results. We conclude that the reliability of simulator tests is sufficiently high. A practical consequence of the findings is that we should maintain at least two assessors in simulator tests in order to keep the tests sufficiently reliable.

9.4 Conclusions

This chapter described how we evaluated the psychometric quality of the assessment system. We investigated interrater agreement, rating errors, and test reliability. We conclude that the assessment system is generally reliable. Reliability is higher for simulator tests than for progression measurements because of well-balanced test scenarios, multiple assessors and objective scoring techniques. We could say that the assessment system is well designed in terms of coherence and transparency of competences, performance criteria and performance standards. The classification of the ATC Performance Model (see chapter 5), serving as a framework for the assessment design, was clearly recognized. However, the assessors' qualities can be improved. Agreement in their profile of competence ratings was not always sufficient. However, decision making (e.g., pass-fail) is not affected because agreement in trainees' overall performance level is sufficiently high. They seem to have problems with pointing at specific deficiencies of trainees, founded in competences. Furthermore, assessors make rating errors to a certain extent, such as range restriction and halo errors. However, they do not differ in leniency/severity (cf. systematic differences) or in their distribution of ratings. Assessor training was organized after this study, which pays attention to the avoidance of rating errors, the meaning of standards and competences (calibration), and general assessment techniques.

This chapter has ignored the influence of time while the progression measurements were successively made over a training period. We argued that learning processes could have influenced the results such as the intercorrelations between competences. Analysis over time, including criterion-related validity, requires specific methods and is subject of the chapters 10 and 11.

10 Evaluation of learning processes: analysis of learning curves

We posed as one of the main purposes of the project that the redesigned assessment system should contribute to more effective and efficient learning processes. Patterns and individual differences in learning processes (e.g., slow starters, learning plateaus) and in performance (strengths and weaknesses; possible deficiencies) should be clearly distinguished to provide a basis for adequate feedback and interventions. A well-designed assessment system will work adequately if the assessment results represent learning processes optimally. Assessment compares the trainees' actual performance to required performance at successive moments of time. Under the assumption that performance is the result of learning, learning curves can be derived from a sequence of performance measures over time. If trainees are learning, then their performance will increase over time.

If these learning curves were sufficiently representative for learning processes, as we expect from general learning theory, then the second step would be to apply the assessment system as an instrument to gain insight in learning processes of individual trainees (how they learn). Consequently, training can be maximally adapted to trainees' needs (e.g., by optimal task selection, coaching, and remedial teaching). Finally, pass-fail decisions might become more valid if they are based on particular characteristics derived from trainees' learning curves.

We examined to which extent the assessment system is able to represent learning processes adequately by answering two main questions¹¹:

1. Does the assessment system represent patterns in learning processes?
2. Does the assessment system represent individual differences in learning processes?

The first question is about similarities in learning processes while the second question focuses on differences in learning. We compared learning curves produced by the assessment system with learning curves that we expect from general learning theory. Next, these derived learning curves can be used to examine criterion-related validity under the assumption that they are sufficiently reliable and valid to represent learning processes. We try to answer the following questions:

¹¹ Part of this chapter is also described in Oprins, Burggraaff & Van Weerdenburg (2007a).

3. Are learning processes in a phase predictive for learning in successive phases?
4. Which properties of learning processes are most predictive for pass-fail decisions?

Finally, evidence of predictive validity makes it possible to define cut-offs for pass-fail decisions.

We should notice that this chapter is restricted to learning curves based on overall performance. Chapter 11 discusses the development of singular competences, which is also crucial for learning.

10.1 Learning curves

Learning curves and learning theory

Learning curves are usually presented as growth curves that show progression over time based on repeated executions of the same task at successive moments of time. The main purpose of this kind of research is modelling learning processes. The power law of practice (Newell & Rosenbloom, 1981) has been a commonly investigated model for learning curves (see chapter 2). This law states that the logarithm of the reaction time for a particular task decreases linearly with the logarithm of the number of practice trials taken. Qualitatively, the law simply says that practice improves performance. Its explanation is founded in theories about skill acquisition that relies on two modes of information processing: controlled and automated information processing (Shiffrin & Schneider, 1977; Schneider & Shiffrin, 1977). As a result of learning, information processing develops from controlled information processing, which requires much effort and attention, into the direction of automation in which attentional demands are strongly reduced. Initial performance is therefore slow and error prone, but with consistent practice performance speed and accuracy will increase rapidly. The power law specifies how this learning process develops: it predicts that performance improves rapidly in initial trials, but that the rate of climb diminishes with more and more trials because everyone has his limits in (cognitive) abilities. The asymptote can also be referred to as a final learning plateau (e.g., Schneider, 1990).

Skill acquisition is often referred to as growth of performance since better performance is usually the result of learning processes. Therefore, growth curves of learning can be made based on performance measures of repeated task execution. The following information is often provided by these growth curves (e.g., Browne & Du Toit, 1991):

- Asymptote: final performance level
- Growth: between the initial state and final performance level
- Rate of growth (change) over time

This results in the model of learning curves in Figure 10.1, representing the power law of practice:

Figure 10.1. Growth curve of learning

Besides the variables used in this growth model, *stability* in learning is another variable that has been used in analysing learning curves (e.g., Schijven & Jakimowics, 2004). It refers to the extent to which a trainee performs constantly.

It is argued that the growth curve of learning (Figure 10.1) especially applies to simple skill acquisition. The specific form of the learning curve is supposed to be strongly dependent on task complexity, task consistency, and individual differences in learning. Complex tasks usually pose more demands on cognitive abilities than simple tasks (Ackerman, 1989). They are assumed to consist of consistent and non-consistent task components (Ackerman, 1989; Schneider, 1990). Consistent task components show large improvements with practice whereas non-consistent tasks do not (Schneider, 1990): consistent tasks require automation while non-consistent tasks require more controlled information processing (see also chapter 2; section 2.3). Furthermore, many individual differences in learning exist such as differences in cognitive abilities, learning styles, personality and other personal properties (e.g., Ackerman, 1988; Jonassen & Grabowski, 1993). These individual differences lead to different initial performance levels, acquisition rates and final learning asymptotes in the learning curves.

Nevertheless, the power law has been used for examining complex skill acquisition, for instance, in a simplified air traffic control task, the Kanfer-Ackerman task (Kanfer & Ackerman, 1988; Ackerman, 1988, 1989; Lee & Anderson, 2001; Taatgen, 2001; Taatgen & Lee, 2003; see section 2.3). Learning curves as presented in Figure 10.1 have been recognized in these studies. We saw that Lee & Anderson (2001) decomposed the Kanfer-Ackerman task and examined learning processes by making learning curves of the sub tasks. They found that learning at the different levels of decomposition is well fit by the power function. Other models have been investigated for complex skills in the same ATC task (Browne & Du Toit, 1991) and in related domains such as flight skill acquisition (e.g., Roessingh, 2002).

However, we assume that complex skill acquisition in real-life training (simulator, on-the-job training) is more complex and less standardized than learning in simulated tasks, which are usually performed by novices and examined in a laboratory situation. In simulator training the task situations are offered to trainees in a sequence of exercises, but in the OJT the task situations cannot be predicted at all. Learning is strongly dependent on the quality of coaches and other influences such as the mental pressure to succeed training or the complex working environment

(physical conditions, colleagues, etc.). We can say that the real ATC task consists of many 'non-consistent' task components that do not necessarily improve by practice in accordance with the power law.

As a consequence, we can imagine that the learning process as a whole has many underlying, smaller learning curves in accordance with the findings of Lee & Anderson (2001). A trainee needs time to assimilate new knowledge and skills with previous experiences in order to automate skills. Therefore, there exist several asymptotes or intermediate learning plateaus during the learning process for each individual. The ATC task can be subdivided into many task components as mentioned before. Under the assumption that these task components obey the power law of practice, and that they are learned one after each other, then we expect that the overall learning curve consists of a sequence of smaller learning curves for each task component to be learned as illustrated in Figure 10.2:

Figure 10.2. Learning curve (complex skills)

However, in simulator training and OJT learning tasks are usually not delivered in such a pre-structured sequence. Many tasks are trained simultaneously since complex skill acquisition implies that a trainee should be able to do multiple tasks at the same time. Learning curves would be more smoothed than visualized in Figure 10.2, because they are the result of many overlapping curves. In addition, the order and tempo of learning is different across trainees because of individual differences in learning (cognitive abilities, learning styles, personality, pre-education, external influences etc.). For instance, 'slow starters' may be present (trainees who learn in a slower tempo) or trainees who show an intermediate learning plateau for some time because they need more time to automate specific skills than others. Figure 10.3 illustrates how some variations of learning curves, more smoothed than presented in figure 10.2, are expected in complex skill acquisition in real ATC training, under the assumption that all trainees start at the same (zero) performance level:

Figure 10.3. Three variations of learning curves (complex skills)

Figure 10.3 presents a fast learner who achieves the final performance level early (blue), a so-called 'slow starter' who needs more time and who shows an intermediate learning plateau (green), and a learner who never achieves the required level and probably fails (red). Ideally, the assessment system should optimally reflect the kind of learning curves depicted here, but our assessment system cannot produce these learning curves in the same way.

Learning curves and the assessment system

Under the assumption that performance is the result of learning, and that performance measurement at successive moments of time can be used to make learning curves, it should be possible to derive learning curves from the assessment results. As described in chapter 7 (section 7.4), trainees' performance is measured at regular intervals (around once a week) but these intervals differ across trainees, due to different training schedules. Sometimes the intervals differ so much that only a rank order can be extracted from these successive moments of time. Multiple assessors assess (averaged) performance by rating a set of competences at a 6-points rating scale during a certain training period. Distinct learning processes take place in between the moments of assessment that cannot be captured completely in these ratings, thus trainees' performance is not really measured continuously. Theoretically, we may not draw a straight line through the points of measurement, because we do not know the precise form of this line.

An important characteristic of the designed assessment system is that assessors assess trainees' performance against augmenting performance standards defined for each subsequent training phase, both in simulator training (preOJT) and in OJT (see section 7.4). These standards are continuously mapped on the same 6-points rating scale. This means that this rating scale, with a value of 4 or more being sufficient, is constantly being recalibrated. When ratings would stay 'sufficient' over time, this implies that the trainee succeeds to meet the increasing standards over time. This trainee is learning and shows progression, for instance, because he is able to handle increasing traffic complexity more efficiently without help of the coach. Furthermore, the same behaviour that is rated as sufficient at the start of training (value > 4) may be rated as insufficient later on (value < 4), because the expecta-

tions of trainee performance (standards) increase over time¹². Therefore, the 'learning curve' produced by our assessment system should be seen as a derivative of the real learning curve, representing a 'second order learning curve'. We describe how we produce the recalibrated learning curves in two steps. First, we draw learning curves that would be expected if the scales were not recalibrated, see Figure 10.4.

Figure 10.4. Learning curves in relation to performance standards at four moments of time

Figure 10.4 represents three learning curves of trainees who are assessed against performance standards at four moments of time with different intervals. The horizontal black lines visualize the standards defined for a training phase. The learning processes in-between the moments of measurement are not visible in assessments. For example, the specific form of the red learning curve in-between the second and third moment of measurement (t_2 and t_3) cannot directly be derived from the assessments, for instance, the specific moment at which the trainee has achieved the standards. The blue learning curve represents a trainee who constantly performs above the required standards, the green learning curve reflects a 'slow starter' who finally achieves the required standards.

The next step is to realize a recalibration into the same rating scale. This implies that the increasing standards in Figure 10.4 are transformed into one horizontal line, since the same value of 4 at the 6-points scale represents 'sufficient' performance over time. Figure 10.5 visualizes the recalibrated learning curves as produced by our assessment system.

¹² In a certain way, educational systems also make use of recalibrated scales (e.g., in the Netherlands a value of 6 is sufficient while task difficulty increases). But learning curves as presented in this chapter have not been found.

Figure 10.5. Recalibrated learning curves, produced by the assessment system

Figure 10.5 shows rather the same learners as in figure 10.4: the blue learning curve represents a learner who constantly performs above standards, the red learning curve refers to a learner who only starts to perform below standards at the first two moments of measurement (t1 and t2), and the green learning curve reflects a learner ('slow starter') who performs worse but finally achieves the required standards. Straight lines connect the points of measurement to suggest a learning process, but these lines do not necessarily represent the real underlying learning processes. We do not draw a line from the beginning since we don't know at which performance level the trainee has started. The horizontal line stays the same during the whole training period, because the performance standards constantly are recalibrated across successive phases. Both figures show four moments of measurements at the end of each phase, but in reality trainees are assessed more often during training at different moments of time. In the further study we use the term learning curves to refer to the recalibrated learning curves as discussed here.

Prototypical learning curves

The learning curves produced by the assessment system should differentiate between trainees who differ in overall training success. Therefore, we made an a priori classification of trainees and we compared their learning curves. This classification, based on overall training success, is as follows:

- *High performers*: passed without problems
- *Moderate performers*: passed with difficulties (slow starters, learning plateaus)
- *Low performers*: failed

This classification was based on expert judgment of three training managers who are responsible for following trainees' progress. This expert judgment serves as an external criterion for trainee success. The managers relied on their general image of trainees' performance and progression over time and made the classification independently of the (quantitative) training results.

Based on general learning theories, we defined prototypical learning curves to which the learning curves of actual trainees can be compared. They serve as hy-

potheses in the analyses described in next section. We describe the properties of the prototypical learning curves for preOJT and OJT ACC below.

PreOJT

We expect that trainees start at a comparable (zero) performance level in preOJT, under the assumption that they do not differ in prerequisite knowledge and skills. They probably perform around or above standards in the first period of training because the first simulator exercises are rather easy for all trainees. In the recalibrated learning curves, the majority of trainees would show decreasing performance during training, because it becomes more difficult to achieve the required performance standards due to the increasing complexity of task situations. But we expect this to differ for the three groups of trainees. The high performers should show constantly sufficient or high performance with only a small decrease in performance. Performance of low performers should decrease; they do not achieve the required standards and therefore they fail (strict cut-offs do not exist as explained in section 7.4). The moderate performers should take an intermediate position, showing more variations (e.g., slow starters, intermediate learning plateaus).

These expectations suggest that the overall differences between trainees of the three groups will increase. Examples of hypothetical learning curves for the three groups of trainees are represented in different colours in Figure 10.6. It should be noticed that the points of measurement are connected with straight lines like in Figure 10.5, although we don't know the learning curves in-between, and that the time intervals differ for each trainee. Figure 10.6 shows that the trainees start with sufficient performance. We see smaller differences between trainees at the start than at the end of preOJT.

Figure 10.6. Hypothetical learning curves of high (blue), moderate (green) and low (red) performers in preOJT

The high performers (blue) perform constantly around or above the required standards. The figure represents two examples of moderate performers (green): an intermediate learning plateau (no. 1) with temporary performance below the standards, and a slow starter (no. 2) who slowly grows towards the required standards. The low performers (red) constantly perform below the required standards and show decrease in performance. Number 3 fails before reaching the end of the training period.

We summarize the distinction between high, moderate and low performers in a more generic way by means of the expected 'zones' (cf. hypotheses) in Figure 10.7:

Figure 10.7. Expected zones for learning curves of high (blue), moderate (green) and low (red) performers in preOJT.

OJT

We expect other learning curves in OJT. We expect more differences in the entrance level between trainees, because the high and moderate performers in preOJT start in OJT with differences in final performance level in preOJT as shown in Figure 10.6. They can fail in OJT, becoming low performers. The assessors are supposed to assess against phase level in contrast with preOJT. Trainees transfer to a next phase if they only have got sufficient competence ratings (value 4 or more). Thus, the length of each phase is flexible (see also section 7.4). The majority of trainees start with a certain amount of insufficient ratings due to the method of assessment against phase level: they cannot have reached the final phase level yet and cannot transfer to the next phase immediately. As a consequence, the majority of trainees, especially moderate and high performers, should show increasing performance over time towards phase level. The difference between high and moderate performers should be that high performers should have a more constantly high performance level and that the moderate performers should show more variations like in preOJT (e.g., slow starters, intermediate learning plateaus). The low performers should show a low mean performance level, which tends to decrease; consequently, they fail during or at the end of training. These expectations suggest that the differences between the trainees who belong to the moderate and high performers should decrease over time, because they all achieve the final performance standards. Figure 10.8 presents some examples of hypothetical learning curves for the three groups of trainees in OJT.

Figure 10.8. Hypothetical learning curves of high (blue), moderate (green) and low (red) performers in OJT

The high performers (blue) constantly perform around or above the performance standards, although they show more increase in performance in OJT than in preOJT, due to assessment against the standards. Within the group of moderate performers, who also show increase in performance, we recognize a slow starter (no. 1) and an intermediate learning plateau (no. 2). The low performers show decreasing performance or performance constantly below the standards. One of them failed during training (no. 3). Figure 10.8 also shows how the entrance level differs across trainees, while there exist less differences across trainees at the end for the high and moderate performers. The expected 'zones' (cf. hypotheses) in a more generic way for the high, moderate and low performers in OJT are presented in Figure 10.9.

Figure 10.9. Expected zones for learning curves of high (blue), moderate (green) and low (red) performers in OJT.

Variables of learning curves

The learning curves are mainly determined by performance level and progression. In the analyses discussed in next section, we applied several measures that define the specific form of the curves:

1) Performance level:

- *Mean performance level*: weighted sum of competence ratings over time (range 1 – 6)
- *Insufficient performance*: occurrences of insufficient performance level over time (i.e. value of weighted sum of competence ratings < 4), presented in a percentage

2) Progression:

- *Growth*: final performance level minus initial performance level
- *Rate of growth*: beta coefficient of the linear regression model (obtained by curve fitting)

It is important to realize that the measures are not similar to those applied in growth models of learning (see Figure 10.1), although some terms are the same. We used *mean performance level* and *insufficient performance* as two measures for performance level. We chose for insufficient instead of sufficient performance because insufficient ratings are more important for learning purposes: the detection of possible deficiencies of trainees. We also used two measures for progression. *Growth* refers to the difference between the final and entrance level, but this does not provide information about the form of the curve in-between. The *rate of growth* indicates the direction of the curves better. We achieved the indexes for the rate of growth (beta coefficients) for each trainee by means of curve fitting in SPSS. We chose for the linear regression model as an estimate for progression, although this model does not fit significantly for each learning curve¹³:

$$(12) P_i(\text{performance}) = \beta_0 + \beta_1 (t) + e_i$$

In this equation β_0 is the intercept that refers to the initial state of performance, β_1 is the slope that refers to the rate of growth over time (t), and e_i is the residual. As an example, Figure 10.10 presents the learning curve of trainee 19 in OJT (learning phase 1). At the horizontal X-axis the variable 'time' is used for successive moments of measurement (1st to 14th for this trainee), and at the vertical Y-axis the variable 'performance' refers to the weighted sum of competence ratings.

¹³ We checked other (non-linear) models, but generally they did not fit better. We should note that we cannot apply curve fitting for groups of trainees because they are assessed at different moments of time.

An alternative would be to do a multilevel analysis because this offers the possibility to analyse longitudinal measures that are taken at both varying and fixed occasions (e.g., Hox, 2002). However, multi-level analysis is usually applied to explain variables at a higher level and to derive a regression model. We do not have variables to be explained at a higher level but we are mainly interested in differences within and between groups. A multilevel analysis would be very complicated for our purpose so we did not apply it, also for practical reasons.

Figure 10.10. Learning curve of trainee 19 in OJT (learning phase 1)

The linear regression model obtained with curve fitting, significant for this specific trainee at $p < .001$, is presented in Table 10.1:

Table 10.1. Linear regression model for trainee 19 in OJT (learning phase 1)

R Square	F	df1	df2	Sig.	Constant	b1
.532	11.353	1	10	.007	3.189	.110

The various measures for performance and progression for this trainee are¹⁴:

- Mean performance level = 3.91
- Insufficient performance = .58
- Growth = 1.56
- Rate of growth (beta coefficient): .110

We calculated these measures for each individual trainee and we used them to do various quantitative analyses (see sections 10.2 and 10.3).

¹⁴ In first analyses, we also included the variable *stability* as applied in some analyses of learning curves (e.g., Schijven & Jakimowics, 2003), represented by the standard deviation (or variance) of the performance level. But finally we deleted this variable, because it did not differentiate sufficiently across the three groups. In addition, we examined whether the number of decreases and increases in the curves would differentiate across trainees, but this did not have enough value either.

10.2 Methods

Data collection

For studying learning curves, we used training results obtained in area control (ACC) training during the period January 2001 till December 2006. We included 182 progression measurements made for 31 trainees in preOJT (simulator training) and 403 progression measurements made for 27 trainees in on-the-job training (OJT) that follows preOJT. These progression measurements were made at three or more successive moments of time during preOJT or OJT for each trainee, with irregular time intervals. Cases with fewer measurements were deleted from the data set.

Only 22 trainees are the same persons in preOJT and in OJT because some of them failed in preOJT, or the assessment system was implemented when trainees were in OJT already. We split up the OJT into the two learning phases for which the standards were defined (see section 7.4). This was not necessary for preOJT due to its different method of assessment and shorter time period.

Methods of analysis

In order to examine possible patterns in learning in conformance with research question 1 ("Does the assessment system sufficiently represent patterns in learning processes?"), we visually compared the learning curves of the trainees, classified into high, moderate and low performers, with the hypothetical learning curves (with the expected zones) presented in Figures 10.7 and 10.9. In addition to this visual comparison, we did quantitative analyses to get more evidence for the patterns in learning curves. We did a discriminant analysis to check whether the classification into three groups, based on expert judgment as an external criterion, was correctly predicted by the variables that define the learning curves (see section 10.1): mean performance level, growth, insufficient performance and rate of growth¹⁵. We also examined how the four variables were related to each other by calculating the correlations (Pearson) between them. An analysis of variance (ANOVA) between the three groups was mentioned to provide insight into the differences in means of the four variables across the three groups.

Next, we investigated possible differences in learning curves in order to answer research question 2 ("Does the assessment system sufficiently represent individual differences in learning processes?"). Variability in learning across trainees is often measured by calculating the standard deviation of mean performance (Ackerman, 1988). Ackerman suggests that these standard deviations reduce with practice because initial task performance is more dependent on individual abilities and final task performance is the result of learning. However, this is only true when the tasks are

¹⁵ A better alternative would be a multinomial logistic regression. However, this was not possible, probably due to a too low number of cases in relation to the variables. Curve fitting was not possible as well because trainees' progression has been measured at different time intervals, and most individual models were not significant. We did not apply a multilevel analysis for other reasons (see footnote 7).

repeatedly practiced (consistent tasks). The real ATC task consists of many non-consistent task components that depend more heavily on innate abilities, and therefore it is reasonable to expect that differences in performance become larger over time in preOJT and smaller for moderate and high performers in OJT. We made graphs of mean performance level for the three groups in order to investigate how these groups differ over time. Next, we made graphs of the standard deviations of mean performance level, in accordance with Ackerman (1988), in order to examine how differences between individuals change over time.

Finally, under the assumption that the recalibrated learning curves sufficiently represent learning curves, we made a first start with exploring criterion-related validity by answering research question 3 ("Are learning processes in a phase sufficiently predictive for learning in successive phases?") and question 4 ("Which properties of learning processes are most predictive for pass-fail decisions?"). In addition to a visual examination of learning curves, we calculated correlations (Pearson) between preOJT and OJT (and between two phases within OJT) for the aforementioned four variables (see previous section). We also correlated the same variables with the pass-fail criterion in preOJT and OJT in order to explore which properties of learning curves are most predictive for pass-fail.

10.3 Results

Classification into groups

Two preOJT managers and one OJT manager classified the trainees into the three categories high, moderate and low performers based on their expert judgment without a close look to the trainees' assessment results (see also section 10.1). If there was disagreement between the preOJT managers and the OJT manager, we followed the judgments of the manager(s) who were responsible for the specific part of training (preOJT or OJT), because they have a better insight into trainee performance in their own part of training. In preOJT we considered a trainee as a moderate performer if the two preOJT managers disagreed about a trainee being a moderate or high performer, independently of the judgment of the OJT-manager, because at least there was some doubt about the trainee's qualities. In OJT we followed the judgment of the OJT-manager if there was disagreement between him and one or two preOJT managers for the same reason. This resulted in the classification of trainees presented in Table 10.2 (preOJT) and Table 10.3 (OJT).

Table 10.2. *Classification of trainees in ACC preOJT based on expert judgment*

ACC preOJT	N (high performers)	N (moderate performers)	N (low performers)
Total agreement	9	7	9
Agreement preOJT managers	3	-	-
Disagreement preOJT managers	-	3	-
Total	12	10	9

Table 10.3. *Classification of trainees in ACC OJT based on expert judgment*

ACC OJT	N (high performers)	N (moderate performers)	N (low performers)
Total agreement	9	1	11
Disagreement with OJT manager	1	5	-
Total	10	6	11

Tables 10.2 and 10.3 show that there was agreement between the managers for the majority of trainees (25 trainees in preOJT (= 65%) and 20 trainees in OJT (= 74%)), but not for all trainees. This can be explained by the fact that the classification between moderate and high performers is not strictly defined; they all passed but with or without problems.

Learning curves

Next, we made graphs of individual learning curves in each group, split up for preOJT and the two learning phases in OJT. We choose a range of a maximum of 7 moments of measurement in preOJT and a maximum of 14 moments in both learning phases in OJT, although a minority of trainees was assessed more often. The X-axis presents the successive moments of measurement, and the Y-axis refers to the performance level, based on a weighted sum of competence ratings. The variable 'time' is only based on a rank order since the intervals vary. We included the standard of sufficient performance in the graphs (value = 4 on a 6-points scale). The results of ACC preOJT are presented in Figures 10.11 – 10.13, and the results of ACC OJT are presented in Figures 10.14 – 10.19.

ACC PreOJT

Figure 10.11. PreOJT high performers

Figure 10.12. PreOJT moderate performers

Figure 10.13. PreOJT low performers

The correspondence with the hypothesized, prototypical learning curves (Figure 10.6 – 10.7) can be recognized. Some exceptions within a group exist, for instance, trainee 9 in Figure 10.11 would rather belong to the moderate performers instead of high performers, and trainee 30 in Figure 10.12 to the high performers instead of the moderate performers. The high performers generally show increasing performance above the standard 'sufficient'; this could point at a higher (implicit) norm (e.g., value > 4.8) for really high performers. As expected, the group of moderate performers (Figure 10.12) has the most variations of learning curves within the group. Trainee 11 is a slow starter, and trainee 5 shows an intermediate learning plateau. The pattern of the low performers (Figure 10.13) is rather similar for all trainees with one exception (trainee 2). Trainees 10, 17, 22 and 27 failed during preOJT already after having received four progression measurements, and they have a rather similar pattern in their learning curves. We have to realize that these failures are not only based on progression measurements but also on a simulator test. Therefore, some other trainees who seem to be worse go further in training, but they may finally fail at the end of preOJT.

ACC OJT (learning phase 1)

Figure 10.14. OJT learning phase 1, high performers

Figure 10.15. OJT learning phase 1, moderate performers

Figure 10.16. OJT learning phase 1, low performers

ACC OJT (learning phase 2)

Figure 10.17. OJT learning phase 2, high performers

Figure 10.18. OJT learning phase 2, moderate performers

Figure 10.19. OJT learning phase 2, low performers

In OJT the hypothesized, prototypical learning curves (Figure 10.8 – 10.9) can also be recognized, although the differentiation between the groups is less clear than in the preOJT and with more peaks and downs in the curves. This probably has two

main causes: trainees' performance is less stable due to the unpredictable task situations in OJT, and there is more variation in the number of progression measurements for each trainee, since the length of OJT and the time intervals highly differ for each trainee, more than in preOJT. The graphs for the three groups are quite similar for learning phase 1 and learning phase 2 as expected, because the assessment and training methods during the whole OJT are the same. The group of high performers (Figure 10.14 and 10.17) performs rather constantly above the sufficient standards as in the preOJT. If there were a (implicit) norm as argued for preOJT, this would be a bit lower (around value > 4.5). It is remarkable that the trainees never perform below the standards except trainee 13 in learning phase 1 (Figure 14), in contrast with the moderate and low performers. We would expect an increase towards the required standards because the trainees are assessed against phase level with all competence ratings having to be sufficient (value > 4), but this increase is only visible for the moderate and low performers. Apparently, some competence ratings can be insufficient, also for high performers, but the weighted sum of ratings stay sufficient for the high performers.

As expected, the variation between trainees within the group of moderate performers is really high (Figure 10.15 and 10.18). Trainees perform around the standards but with many peaks and downs. Trainee 21 shows a learning plateau in learning phase 1 (Figure 10.15) and trainee 14 seems to be a slow starter both in learning phase 1 (figure 10.15) and in learning phase 2 (figure 10.18). The low performers perform also around the standards in learning phase 1 (Figure 10.16), but below the standards with decreasing performance in learning phase 2 (Figure 10.19). We should realize that some trainees failed in learning phase 1 already but not all of them. Some may have achieved the required performance level for phase 1 but fail in learning phase 2. The distinction in learning curves between trainees failed in learning phase 1 or in learning phase 2 is not directly visible in Figure 10.15. A very strange outlier in Figure 10.16 and 10.19 is trainee 30. This trainee performs below standards only twice in both learning phase 1 and 2, only in the beginning, but he has still failed. This trainee failed on qualitative grounds: his performance was not stable enough to work safely in all situations, but it was difficult to make this clear in the competence ratings by the assessors. They explained the problem in the additional comments.

Discrimination between groups

We did a discriminant analysis to check whether the classification based on expert judgment was correctly predicted by the variables described for the learning curves (see section 10.1): *mean performance level* (averaged sum of competence ratings), *growth* (final minus initial performance level), *insufficient performance* (occurrence of insufficient performance in percentage), and *rate of growth* (beta coefficient in linear regression model). Additional analyses were calculations of correlations (Pearson) between the variables for each group, and an analysis of variance (ANOVA) across the groups.

PreOJT

The first discriminant function is significantly different across groups (chi-square = 54.17, $df = 8$, $p = .000$). The discriminant function coefficients indicate that mean performance is the best predictor for the classification into groups, respectively followed by insufficient performance, rate of growth and growth. In total, the discriminant function successfully predicted group membership for 83.9%:

Table 10.4. Predicted group membership for preOJT (N=31)

	Low performers	Moderate performers	High performers	Total
Low performers	7	2	0	9
Moderate performers	1	8	1	10
High performers	0	1	11	12

However, the variables are not totally independent of each other, and this may have affected the results. Table 10.5 presents the correlations (Pearson) between the averaged variables over time. As expected, the two measures for performance (mean performance level, insufficient performance) are highly correlated as well as the two measures for progression (growth, rate of growth).

Table 10.5. Correlations (Pearson) between the variables for preOJT (N=31)

	Mean performance	Insufficient performance	Growth	Rate of growth
Mean performance				
Insufficient performance	-.89**			
Growth	.43*	-.35		
Rate of growth	.46**	-.31	.60**	

* Correlation is significant at the 0.05 level (2-tailed); ** Correlation is significant at the 0.01 level (2-tailed).

It is remarkable that mean performance level is also highly correlated with rate of growth but insufficient performance is not; apparently, trainees with high mean performance level tend to show increasing performance and vice versa. This can be recognized in the graphs of learning curves, especially of high and low performers (see Figure 10.11 and 10.13).

An analysis of variance (ANOVA) shows that the means of the four variables differ significantly ($p < .05$) across the three groups as presented in Table 10.6. However, Levene's test shows that the variances of the variable insufficient performance are not homogenous in contrast with the other three variables (significant at $p < .05$):

Table 10.6. Analysis of variance (ANOVA) for preOJT (N=31)

	F	df1	df2	Sig.
Mean performance	44.843	2	28	.000
Insufficient performance	11.289	2	28	.000
Growth	5.907	2	28	.007
Rate of growth	18.457	2	28	.000

OJT learning phase 1

The first discriminant function is significantly different across groups (chi-square = 22.36, $df = 8$, $p = .004$). The discriminant function coefficients indicate that insufficient performance is the best predictor for the classification into groups, respectively followed by mean performance level, rate of growth and growth. In total, the discriminant function successfully predicted group membership for 73.1%:

Table 10.7. *Predicted group membership for OJT learning phase 1 (N=26)*

	Low performers	Moderate performers	High performers	Total
Low performers	7	1	2	10
Moderate performers	1	4	1	6
High performers	1	1	8	10

The correlations (Pearson) presented in Table 10.8 show that the two measures for both performance and progression are highly correlated as expected, but that the measures of performance and progression are only slightly correlated. The latter result is different from preOJT but it is more conform our expectations:

Table 10.8. *Correlations (Pearson) between the variables for OJT learning phase 1 (N=26)*

	Mean performance	Insufficient performance	Growth	Rate of growth
Mean performance				
Insufficient performance	-.88**			
Growth	-.31	.26		
Rate of growth	-.19	.18	.77**	

* Correlation is significant at the 0.05 level (2-tailed); ** Correlation is significant at the 0.01 level (2-tailed).

An analysis of variance (ANOVA) shows that the means of two variables differ significantly ($p < .05$) across the three groups as presented in Table 10.9; Levene's test showed that the variances of the four variables are all homogeneous:

Table 10.9. *Analysis of variance (ANOVA) for OJT learning phase 1 (N=26)*

	F	df1	df2	Sig.
Mean performance	8.154	2	23	.002
Insufficient performance	.124	2	23	.884
Growth	1.487	2	23	.247
Rate of growth	14.456	2	23	.000

OJT learning phase 2

The first discriminant function is significantly different across groups (chi-square = 45.49, $df = 8$, $p = .000$). The discriminant function coefficients indicate that mean performance is the best predictor for the classification into groups, respectively followed by insufficient performance, growth and rate of growth. In total, the discriminant function successfully predicted group membership for 90.5%:

Table 10.10. *Predicted group membership for OJT learning phase 2 (N=21)*

	Low performers	Moderate performers	High performers	Total
Low performers	5	1	0	6
Moderate performers	0	6	0	6
High performers	0	1	8	9

The results of the correlations presented in Table 10.11 show the same results as for learning phase 1: the two measures for both performance and progression are highly correlated as expected, but the measures of performance and progression are hardly correlated:

Table 10.11. *Correlations (Pearson) between the variables for OJT learning phase 2 (N=21)*

	Mean performance	Insufficient performance	Growth	Rate of growth
Mean performance				
Insufficient performance	-.92**			
Growth	.20	-.10		
Rate of growth	.30	-.19	.82**	

* Correlation is significant at the 0.05 level (2-tailed); ** Correlation is significant at the 0.01 level (2-tailed).

An analysis of variance (ANOVA) shows that the means of all variables differ significantly ($p < .05$) across the three groups (see Table 10.12). Levene's test showed that we may not assume that the variances of mean performance level and insufficient performance are homogeneous at $p < .05$.

Table 10.12. *Analysis of variance (ANOVA) for OJT learning phase 2 (N=21)*

	F	df1	df2	Sig.
Mean performance	28.566	2	18	.000
Insufficient performance	3.812	2	18	.042
Growth	7.640	2	18	.004
Rate of growth	8.960	2	18	.002

The results for both preOJT and the two learning phases in OJT suggest that mean performance level and the rate of growth are the best predictors for the classification into the groups, although growth and insufficient performance generally contribute to this classification as well. Although the rate of growth is only a rough indication for progression as we mentioned before, it seems to be sufficiently relevant to recognize patterns in learning curves. But we should note that the two measures for performance level and for progression are highly correlated. We may conclude that if we have to choose one of the two measures, than we will choose for respectively mean performance level and rate of growth to define the learning curves and to discriminate between the groups of trainees. Finally, we should realize that the groups are the same for both learning phases in OJT. This may have affected the results be-

cause some trainees may fit better in two different groups for the two phases. This detailed distinction was too difficult to make for the training managers, because they were not able to separate learning phase 1 and 2 clearly in their judgments.

Differences in learning curves

We examined whether the three groups differentiate in means sufficiently by making graphs of the mean performance level at successive moments of time. This actually summarizes the graphs presented in Figures 10.11 – 10.19. We expect that the means are rather different across the three groups. We also expect that mean performance over time decreases in preOJT, especially for moderate and low performers, and that mean performance increases in OJT, especially for high and moderate performers as explained in the previous section. These expectations are mainly confirmed in the graphs of preOJT and OJT (learning phase 1 and 2) presented in Figures 10.20 – 10.22.

Figure 10.20. PreOJT means over time

Figure 10.21. OJT learning phase 1, means over time

Figure 10.22. OJT learning phase 2, means over time

In PreOJT the means of the three groups start above the standards, but the differences become larger over time as we expected. The high performers generally show mean performance far above standards, and the moderate performers only a bit above standards with decreasing performance because it becomes more difficult for

them to achieve the standards. The low performers perform below standards with decreasing performance. The differences between the three groups in both learning phases of OJT are also very clear and comparable with the findings in preOJT. The means generally increase over time in OJT because trainees are assessed against the standards, as we explained in section 10.1. Figure 10.21 presents a remarkable exception for the low performers in learning phase 2: its mean performance increases over time. This is probably caused by the fact that some trainees within these groups achieved the standards in learning phase 1, but finally failed in learning phase 2. Furthermore, the moments of measurement differ more in OJT than in preOJT; this number decreases over time.

Next, we examined how the differences in mean performance level across trainees develop over time. We made graphs of the standard deviations of the performance level at successive moments of time. We expect that the overall differences, expressed in the purple line that represent 'total', increases over time in preOJT, but decreases for the moderate and high performers in OJT. These expectations are confirmed for the preOJT but not for the OJT as illustrated in the graphs presented in Figures 10.23 – 10.25.

Figure 10.23. PreOJT standard deviations over time

Figure 10.24. OJT learning phase 1: standard deviations over time

Figure 10.25. OJT learning phase 2: standard deviations over time

In fact, only the graphs of the total group of performers are relevant for examining the differences between individuals over time. Figure 10.23 presents increasing standard deviations over time with an exception at t7. This can be explained by the

fact that some trainees are not included because they received fewer progression measurements, or they failed already. Figure 10.24 and 10.25 show that the standard deviations remain rather similar over time, while we expect a decrease, except for the last moments of time in learning phase 1. These results may also be based on a smaller number of trainees. We suggested that the variations between trainees would be the highest for the moderate performers; this is confirmed by the results presented in Figure 10.23 and 10.24, although not very clearly.

Analysis of criterion-related validity

Finally, we explored whether performance in preOJT is predictive for performance in OJT. We examined the relationship between the classification into groups for the trainees who followed both preOJT and OJT. Table 10.13 presents the classification of trainees:

Table 10.13. *Relationship classification into groups between preOJT and OJT (N=19)*

PreOJT	High performers	Moderate performers	Low performers
	OJT	OJT	OJT
Moderate performers	0	3	7
High performers	7	2	0

Although the number of trainees is very small, the relationship between preOJT and OJT seems to be close. Table 10.13 shows that moderate performers in preOJT have never become high performers in OJT and that high performers in preOJT have never failed.

Next, we investigated this relationship in more detail by calculating correlations (Pearson) of the four variables between preOJT and OJT (both learning phases); see Table 10.14.

Table 10.14. *Correlations (Pearson) of performance and progression between preOJT and OJT (N=22)*

PreOJT	Mean performance		Insufficient performance		Growth		Rate of growth	
	OJT		OJT		OJT		OJT	
	Phase 1	Phase 2	Phase 1	Phase 2	Phase 1	Phase 2	Phase 1	Phase 2
Mean perform.	.75**	.73**	-.62**	-.78**	-.07	.06	.15	.26
Insuff. perform.	-.26	.15	.09	-.27	.11	.33	.21	.22
Growth	.19	.50*	-.22	-.52*	.34	.20	.27	.20
Rate of growth	-.58**	-.78**	.58**	.81**	.23	.17	-.01	-.09

** Correlation is significant at the 0.01 level (1-tailed); * Correlation is significant at the 0.05 level (1-tailed)

Table 10.14 suggests that the performance level in preOJT is rather predictive for performance (mean, insufficient performance) in OJT, despite of the low number of cases. The rate of growth significantly correlates with the same variables but in the opposite direction. This can be explained by the fact that the assessment method is

different in preOJT and OJT. Moderate and high performers tend to show increasing performance in OJT, but not in preOJT. Table 10.15 presents the same correlations but now for OJT learning phase 1 and 2 (again for a low number of cases).

Table 10.15. *Correlations (Pearson) of performance and progression between OJT phase 1 and phase 2 (N=18)*

Learning phase 1	Learning phase 2			
	Mean performance	Insufficient performance	Growth	Rate of growth
Mean perform.	.65**	-.18	.01	-.64**
Insuff. perform.	-.33	.15	.06	.35
Growth	-.07	.18	.05	.12
Rate of growth	-.75**	.26	-.05	.71**

** Correlation is significant at the 0.01 level (1-tailed); * Correlation is significant at the 0.05 level (1-tailed)

Table 10.15 shows that mean performance level and rate of growth are highly correlated between phase 1 and 2 in OJT. As in preOJT, these variables correlate with each other in the opposite direction, probably for the same reasons. Insufficient performance and growth are less predictive as shown in Table 10.111.

Next, we calculated the correlations (Pearson) between the variables of learning curves in preOJT and pass-fail in preOJT as well as in OJT, see table 10.16:

Table 10.16. *Correlations (pass-fail) between learning curves and pass-fail preOJT and OJT*

PreOJT	Pass-fail preOJT (N=31)	Pass-fail OJT (N=22)
Mean performance	.71**	.56**
Insufficient performance	-.65**	-.41*
Growth	.54**	-.15
Rate of growth	.60**	.22

Table 10.16 suggests that the best predictor for pass-fail in preOJT is performance (mean and insufficient performance), for both preOJT and OJT. Progression (growth, rate of growth) in preOJT appears to be also a good predictor for the pass-fail decision in preOJT, but it is not very predictive for pass-fail in OJT. But again we should realize that the number of cases is very low. Tables 10.17 and 10.18 present the results of the same analyses but now for OJT (learning phase 1 and 2):

Table 10.17. *Correlations (pass-fail) between learning curves and pass-fail OJT learning phase 1*

OJT phase 1	Pass-fail phase 1 (N=26)	Pass-fail phase 2 (N=21)
Mean performance	.55**	.38*
Insufficient performance	-.59**	-.46*
Growth	-.05	-.13
Rate of growth	-.02	-.34

As in preOJT, performance (mean and sufficient performance) appears to be a better predictor for pass-fail than progression (growth and rate of growth), especially in OJT

learning phase 1. In learning phase 2 all variables seems to be predictive for the pass-fail decision at the end of this learning phase.

Table 10.18. *Correlations (pass-fail) between learning curves and pass-fail OJT learning phase 2*

OJT phase 2	Pass-fail phase 2 (N=21)
Mean performance	.74**
Insuff. performance	-.65**
Growth	.61**
Rate of growth	.54**

10.4 Discussion

This chapter proposed a new method to derive learning curves from assessments. We made 'recalibrated' learning curves, based on assessment against performance standards that are constantly recalibrated into the same rating scale. This was needed to examine whether learning curves, produced by the assessment system, were sufficiently representative for learning processes as part of the evaluation. In this way, this study adds to the studies with the Kanfer-Ackerman task, also an ATC task, aimed at modelling learning processes (Kanfer & Ackerman, 1988; Ackerman, 1988, 1989; Lee & Anderson, 2001; Taatgen, 2001; Taatgen & Lee, 2003).

Representative learning curves

One of the main purposes of assessment is to support learning by providing adequate feedback and by undertaking appropriate training interventions. Therefore, a well-designed assessment system should represent learning processes optimally. This implies that the assessment results should sufficiently discriminate between trainees and that properties of learning processes, based on general learning theory, should be recognized. We formulated two research questions. To answer the first research question, "Does the assessment system represent patterns in learning processes?", we defined three prototypical patterns of learning curves from high, moderate and low performers that served as the hypotheses of the analyses. The learning curves of real trainees, classified in the three groups by training managers, were visually compared with these prototypical learning curves. The findings showed that the discrimination between the three groups was sufficient and in accordance with our expectations.

We also analysed the learning curves quantitatively in answering the aforementioned question with a discriminant analysis. Therefore, we defined two main variables that were each measured in two ways: performance (mean performance level, occurrences of insufficient performance) and progression (growth, rate of growth). With respect to performance, the groups were mainly distinguished by mean performance level while the occurrences of insufficient performance played a minor role. This can be explained by the fact that a clear cut-off for insufficient performance, based on a weighted sum of competence ratings, does not exist yet. The assessors

only rate singular competences but they do not know whether their assessment as a whole would be sufficient or insufficient. We may expect that the assessors' competences ratings are not completely reliable (see chapter 9), which affects the reliability of the weighted sum as well. Besides, the occurrences of insufficient performance do not provide information about how much performance is below or above standards. The mean performance level appears to be a better measure because it is less dependent on the specific standards for insufficient and sufficient performance. It is measured on a continuous scale and it provides information about how much the trainee performs below or above the standards.

Progression plays a role as well, especially the rate of growth (beta coefficient of the linear regression model for an individual trainee). We should realize that progression in the recalibrated learning curves differs from that in general growth curves: an horizontal line implies that the trainee shows progression and is still learning since the required standards increase over time. Growth (difference between final and initial performance level) is less distinctive. This can be explained by the fact that using only two measures, at the first and the last moment of time, is not reliable enough. Individual learning curves show many variations (cf. high variance), caused by instable performance of the trainee (Schijven & Jakimowics, 2004), and by the varying task situations especially in OJT. Rate of growth indicates a certain direction of the learning curve, based on the whole range of measures. This appears to be a better measure for progression than growth. However, we should note that the beta coefficients are not always reliable either, since the linear regression model does not fit for each trainee.

The findings are in accordance with general learning theory and with the power law of practice (Newell & Rosenbloom, 1981). Performance usually improves by practice, and therefore high and moderate performers show progression over time in our recalibrated learning curves (cf. horizontal or increasing lines in the graphs). Low performers achieve a final learning plateau earlier than the moderate and high performers. The decreasing lines make this visible: the gap between their performance level and the required standards becomes higher over time. The graphs also illustrated that the learning curves are not as smooth as we would expect for simple skill acquisition (see section 10.1). The derived learning curves of trainees show many peaks and downs, and outliers exist that do not fit well in the expected patterns. Under the assumption that the reliability of the learning curves is sufficient, this confirms the idea that ATC is a very complex skill, especially when it is acquired in real environments. It comprises a substantial number of non-consistent (less trainable) task components, and is therefore highly dependent on individual and innate abilities (Schneider, 1990; Ackerman, 1988). This explains why a rather high number of candidates, the group of low performers, fails.

The second research question was: "Does the assessment system represent individual differences in learning processes?". This question can also be confirmed, based on a visual examination of learning curves of individual trainees. We specifically mentioned two types of trainees: slow starters who start with performance below standards but improve later on, and trainees with an intermediate learning plateau who need more time to assimilate new knowledge and skills with previous ones.

We recognized these learners in the graphs: we saw slow starters with an increase in performance later than others, and trainees with an intermediate learning plateau with temporary performance below standards.

We expected that individual differences increase over time in simulator training (preOJT). This was confirmed by a visual inspection of the means of performance across the three groups and of the standard deviations, as measures for individual differences (Ackerman, 1988), which increase over time. Under the assumption that trainees do not differ much in prerequisite knowledge and skills, they start at a comparable sufficient level of performance. Task difficulty increases and trainees' performance becomes more dependent on innate abilities (Ackerman, 1988, 1989). The results are less clear for the on-the-job training (OJT), but more factors influence (differences in) performance there. Trainees do not necessarily start at a comparable entrance level because their final levels in simulator training differ. The standards cannot be defined as clearly in OJT because the task situations cannot be planned. This also may cause more problems with the reliability of the assessments in OJT than in preOJT. In sum, the assessment system sufficiently discriminates between individual trainees.

Pass-fail decisions

Under the assumption that the learning curves are sufficiently reliable to represent learning processes, we used them to do some preliminary analyses in order to examine criterion-related validity. However, we should notice that the number of trainees who succeeded the whole Unit Training is too low yet to have valid evidence. We formulated two research questions: "Are learning processes in a phase predictive for learning in successive phases?", and "Which properties of learning processes are most predictive for pass-fail decisions?". We found that moderate performers in preOJT have never become high performers in OJT and that high performers in preOJT have never failed. We correlated the variables that defined the learning curves, performance (mean performance level, occurrences of insufficient performance) and progression (growth, rate of growth) for preOJT with OJT and between the two learning phases in OJT. The findings suggest that learning processes in earlier phases are predictive for future learning, although the number of cases is very small. The results also confirmed the findings mentioned above. Mean performance level seems to be the best predictor followed by rate of growth, not only for pass-fail within the phase but also for performance and progression in successive phases. Ideally, a cut-off could be made if we would have more data. We should also take into account the problem of restriction-of-range: we do not know how the failed trainees would have performed in successive training.

A first attempt for such a cut-off could be based on the findings presented here by defining some weights for the four variables. Possible cut-offs are visualized in the graphs of the failures (see Figures 10.26 – 10.28), although there are exceptions and the number of trainees is much too low yet. For OJT learning phase 1 we only present the failures instead of the total group of low performers (see Figure 10.16), because some of them passed in phase 1 and failed in phase 2.

The cut-off lines start at the minimum performance level (value 4) and decrease over time. They confirm that the cut-off should be based on both performance and on progression. This could be expressed in a quantitative cut-off if more data were available. This cut-off would probably be different for preOJT and OJT, and for the two learning phases in OJT. For instance, such a cut-off could be based on sufficient mean performance level over time (value 4 or higher), few occurrences of insufficient performance (e.g., lower than 50%), and positive progression (growth and rate of growth higher than zero). However, we have to be careful with defining a fixed cut-off, because we must avoid false negatives due to the existing shortage on controllers, and because the reliability of the data might not be sufficient enough yet. Furthermore, we experienced that qualitative information is indispensable for getting a complete image of trainees' performance (see section 7.4). An alternative would be to define a cut-off with thresholds. This implies that trainees directly fail below a certain value and directly pass above a certain value, while pass-fail decisions in-between can be based on an extensive qualitative analysis. We applied such a system in a selective module (Oprins et al., 2007).

Figure 10.26. PreOJT failures

Figure 10.27. OJT (learning phase 1), trainees who failed in OJT learning phase 1

Figure 10.27. OJT (learning phase 2), trainees who failed in OJT learning phase 2

Practical implications

In sum, the findings have shown that learning curves, produced by the assessment system, can represent learning processes as expected from general learning theory, especially complex skill acquisition (e.g., Ackerman, 1988, 1989; Schneider, 1990; Newell & Rosenbloom, 1981). Both patterns and individual differences in learning can be recognized in the learning curves, and this contributes to the quality of the assessment system. Exploring criterion-related validity suggests that the learning curves are predictive for pass-fail and future performance to a certain extent. These findings may have some practical implications in the future, although we need a higher number of cases (trainees) to get valid evidence.

A trainee can be classified in one of the three groups based on the characteristics of his/her (recalibrated) learning curve, and ultimately predictions can be made about future learning. Although selection always precedes ATC training, high predictive validity still remains difficult to achieve (e.g., Eissfeldt, Heil & Broach, 2002), and the training will be selective to a certain extent. We strive for avoidance of false positives (undeserved passed) in earlier training phases in order to save costs and to make training more efficient. The fact that progression (especially rate of growth) plays a role as well in addition to performance level suggests that the learning potential of trainees is also assessed: a trainee does not fail if he is still learning and has not reached its final learning plateau yet. This relates to 'dynamic assessment' (Schneider Lidz, 1987; Campione & Brown, 1990; see also chapter 4; section 4.1): the learning process itself is evaluated as well.

Besides prediction of pass-fail, specific training interventions can be undertaken in an early stage, based on trainees' progress. For instance, if a trainee performs below standards but shows progression and has some learning potential, then remedial teaching or extra training can help him/her to succeed. Or if we would know that a trainee is a slow starter, then we could give him/her more time. This implies that the learning curves can be used for diagnostic information and for adapting training to individuals' needs. This relates to diagnostic assessment: finding causes for insufficient performance (Regian & Schneider, 1990; Seamster, Redding & Kaempff, 1997; see also section 4.1). However, for this purpose the learning curves alone do not provide enough information. We also need information about the development of specific competences, for instance, for defining and repairing specific deficiencies. The distinction between more and less trainable skills (cf. consistent vs. non-consistent task components; Ackerman, 1989; Schneider 1990) is important in this context as well. Individual differences in competence development are discussed in chapter 11.

In addition, the learning curves can be used for evaluating training design based on learning processes of groups of trainees. We did the same analyses of learning curves for other courses in Unit Training but we did not present the results in this study. The ground control (GND) course in preOJT takes five weeks and all trainees are assessed at the same moments of time: at the end of each week. We found that almost all trainees ($N = 26$), classified into the three groups of low, moderate and high performers, perform worse in the second and in the fourth week. We used these results to evaluate parts of the training critically: the task sequence

(simulator exercises), the standards defined for each week, and the quality of specific coaches and assessors. Based on these findings, changes have been made in the training such as building a more gradual sequence of exercises.

However, to achieve these practical applications we need more information about variations in learning curves of a substantially higher number of trainees. More insight into complex skill acquisition in general and especially in ATC is required since a uniform theory does not exist yet (Lee & Anderson, 2001; Taatgen & Lee, 2003). A more fundamental problem may be that we will never be able to make predictions based on learning curves and to adapt training to individuals as have been done in studies of aptitude-treatment interaction (Snow, 1989; see chapter 2, section 2.3) because there are too many individual differences in learning (e.g., Petraglia, 1998; Bereiter, 2002). This also explains why it is so difficult to predict training success in selection. Furthermore, the concept of learning plateaus makes it difficult to make predictions during the learning process; they can be final or intermediate, caused by assimilating new knowledge and skills with existing ones as we explained above.

Finally, we should realize that the recalibrated learning curves produced by our assessment system are limited in various ways. First, a certain amount of unreliability will always exist in the assessors' competence ratings in these kinds of performance assessment systems. Assessors have to rely on observable behaviour while they are assessing cognitive processes (e.g., Murphy & Cleveland, 1995). We defined performance standards for successive phases, but especially in OJT they can only serve as rough guidelines because the task situations are not structured. Second, the recalibrated learning curves do not represent the whole learning processes because they do not provide information about learning processes in-between the moments of measurements. It would be useful to get more detailed information, for instance, at which specific moment a trainee overcomes an intermediate learning plateau and starts to show progression again. This can be caused by various factors: specific learning tasks, coaching techniques, his/her own learning process etc. Third, we classified the trainees into three groups but in fact there is a continuum between the groups. The classification was based on training success, but the groups are not strictly separated except for the group of low performers (failures). More research is needed into these types of recalibrated learning curves to get more insight into the underlying learning processes.

10.5 Conclusions

In conclusion, patterns and individual differences in learning processes and in performance can clearly be recognized in (recalibrated) learning curves produced by the assessment system. This implies that our designed assessment system could support learning processes as stated as important purpose of assessment. We consider the learning curves as sufficiently representative for the learning processes. We posit that the assessment results can be used for providing feedback to trainees and for adapting the training to their needs, thereby making training more effective and effi-

cient. The next step is to use the learning curves to achieve more valid pass-fail decisions. Although preliminary validation results suggest that trainees' performance and progression are predictive for successive training phases, more longitudinal research is needed. Furthermore, the analysis of competence development is important since competences refer to trainees' strengths and weaknesses; this is subject of chapter 11.

11 Evaluation of learning processes: competence development

A well-designed, competence-based assessment system should not only be able to represent overall learning by means of learning curves as discussed in chapter 10, but also the development of specific competences¹⁶.

11.1 Development of competences

We argued in chapter 2 (section 2.2) that there exist differences in trainability between competences in conformance with our view on competences. Competences are the result of learning processes but a person should dispose of certain personal characteristics (Roe, 2002, 2005). We assume that all competences are trainable, but that not everyone is able to learn them within a restricted period of time. Competences are supposed to comprise various mixtures of 'consistent task components' that improve by more practice and 'non-consistent task components' that depend more heavily on individual abilities (Schneider, 1990; Regian & Schneider, 1990; see section 2.3). Therefore, a strict separation between trainable and non-trainable competences does not make sense (Bereiter, 2002). Learning theory says that learning develops from controlled to automated information processing (Anderson, 1983; Hacker, 1998). Rasmussen (1986) distinguishes three phases in this development: from knowledge-based to respectively rule-based and skill-based level of behaviour. It is assumed that the transition to the third phase does not necessarily happen for non-consistent task components (Ackerman, 1989).

As described in chapter 3 (section 3.1), air traffic control is supposed to be a complex skill that consists of a combination of consistent and non-consistent task components (Schneider, 1990). Isaac and Ruitenberg (1999) state that skills such as equipment operation, strip- and label management and radiotelephony are assumed to be performed at the skill-based level while the application of standard operations (e.g., maintaining separation minima) occurs at the rule-base level, and solving unfamiliar or novel traffic situations (e.g., emergencies) at the knowledge-based level. From the perspective of the aforementioned learning theory, this implies that the trainability of skills performed at the skill-based level is higher than respectively rule-based behaviour and problem solving at the knowledge-based level.

¹⁶ Part of this chapter has also been described in Oprins, Burggraaff & Van Weerdenburg (2007a)

Situational awareness (SA) is considered very critical in ATC (Shebilske, Goettl & Garland, 2000; Isaac & Ruitenbergh, 1999; Garland, Stein & Muller, 1999). In our assessment system, this mostly relates to the competences mental picture and attention management (cf. ATC Performance Model, see chapter 5). Isaac and Ruitenbergh emphasize the importance of 'visual imagery' as a basis for planning and decision making: building a three-dimensional mental picture. They suggest that ATC selection should pay attention to imagery and visualization. This is usually done by tests that measure spatial orientation. Furthermore, situational awareness heavily depends on controllers' memory capacities (Isaac & Ruitenbergh, 1999; Garland, Stein & Muller, 1999). The acquisition of these kinds of competences depends on innate abilities, but they can be improved by training to a certain extent. Shebilske, Goettl and Garland (2000) argue that attention management strategies should be trained to enhance situational awareness. Garland, Stein & Muller (1999) mention the principle of 'chunking' as a mean to expand the limits of working memory as is also elaborated by Davison and Hansman (2003). Controllers learn to see sets of aircraft as one chunk or a single information unit in working memory, for instance, a potential conflict or a standard flow of aircraft (see also section 3.1).

Another critical competence appears to be *workload management*. This relates to the application of strategies (see section 3.1): it is assumed that experienced controllers have learned to adapt their strategies towards less effort when the number of aircraft increase beyond their comfortable workload level (Sperandio, 1978). We argued that workload is supposed to have a strong subjective component: workload as experienced by controllers is not only dependent on traffic load but also on factors such as emotions, fatigue and motivation, or so-called 'psychophysiological state' (Hockey, Coles & Gaillard, 1986; see also section 2.1). People are considered to react differently to these situations due to differences in personality (Hettema, Leidelmeijer & Geenen, 2000). We explained in section 2.1 that individuals differ in state-orientation versus action-orientation when they face emotional-related situations (Kuhl, 1994). State-orientation is characterized by passivity, indecisiveness and hesitation, while action-orientation is characterized by activity, decisiveness and initiative. Action-orientation seems to be crucial in ATC and relates to the competence decisiveness in our assessment system. Therefore, ATC selection also focuses on personality characteristics such as stress-resistance, vigilance, and dominance (e.g., Eissfeldt, Heil & Broach, 2002).

We explained in chapter 5 (section 5.2) how trainability relates to the classification of the ATC Performance Model: less trainable competences mainly belong to the part of 'information processing' in the model as opposed to the 'actions' and 'influencing factors'. In this view, we assume that less trainable competences, assessed in ACC training, are: *mental picture, attention management, planning, decisiveness, and workload management*. More trainable competences are supposed to be: *verbal expression, listening, co-ordination, equipment operation, attitude and team orientation*. *Safety and efficiency* are referred to as the 'outcome' in the model and are therefore difficult to classify in one of these rubrics. However, we should realize that a strict separation between trainable and non-trainable competences could not be made, because each competence comprises many components that differ in trainability as we mentioned before.

In order to adapt training to individual needs, a well-designed assessment system should discriminate between competence developments across trainees. This implies that the assessment results should indicate, for instance, that one trainee has deficiencies in the competence *workload management*, while another trainee has difficulties with *attention management*. Training interventions should aim at repairing these deficiencies as far as possible, but this depends on their trainability.

If we had tangible evidence from assessment results about competences that are less trainable and are more predictive for pass-fail, we could make pass-fail decisions more valid, based on history of previous trainees. This relates to predictive validity (see chapter 10). Ultimately, as argued in the previous chapter, cut-offs could be made based on a weighting of competence ratings. The competences could also serve as criteria to be used in the validation of selection (Roe, 2005) if we assume that the competences assessed in training are good predictors for operational performance. It is still difficult to get reliable and valid criterion measures for validation of selection, and research into the main reasons for failing is scarce (e.g., Eissfeldt, 2002). MRU (1993) did a study on the Swedish ATC system. They made a ranking of reasons for failing of 17 trainees in basic training during three years based on a qualitative analysis. They found main problems in respectively rigid working methods, passivity/decision making, low stress tolerance, and lack of theoretical knowledge. Switching from low to high workload was also one of the reasons, but it was ranked less high.

The purpose of this chapter is to examine the degree to which the assessment system differentiates between individual trainees with regard to the development of competences. We pay special attention to the distinction between passed and failed trainees, because this relates to the predictive validity of the assessment system. The analyses aim at answering the following research questions:

1. Does the assessment system represent individual differences in competence development so that training can be maximally adapted to trainees' needs?
2. Which competences are most predictive for pass-fail decisions (cf. main reasons for failing)?

We use the same classification into the three groups of high, moderate and low performers as described in chapter 10. We could say that this chapter analyses the differences in learning processes between trainees in more detail than only focusing on learning curves.

11.2 Methods

We used the same data as in chapter 10: progression reports made for ACC preOJT (N = 188) and OJT (N = 407). For the majority of analyses, we averaged the competence ratings for each trainee in preOJT (N = 34) and in OJT (N = 27) because the ratings are supposed to be intercorrelated per trainee, being at a higher level in the

data structure (see also Appendix C)¹⁷. We made a distinction between (mean) *competence level* and *competence development* over time, comparable with respectively *performance* and *progression* as applied in chapter 10 (see section 10.1). The use of measures is dependent on the purposes of analyses.

Individual differences in competence development

In order to answer research question 1, we did various analyses. First, we examined to which extent individual trainees differ in mean competence level over time by executing an analysis of variance (ANOVA). We did this separately for each group: high, moderate and low performers, and also for the total group of trainees. We assume that trainees do not only differ in mean competence level across groups but also within groups due to individual differences in learning. We also expect that differences in less trainable competences (e.g., *mental picture*, *workload management*) are higher than in more trainable competences (e.g., *equipment operation*, *label management*). Less trainability would result in more variance because this is supposed to be more dependent on individual characteristics as argued.

Next, we examined differences between the three groups of high, moderate and low performers by calculating the occurrences of insufficient competence ratings. As competences were highly correlated (see chapter 9, section 9.2), we did this analysis for insufficient competence ratings instead of the means, because this differentiates better. Furthermore, insufficient ratings are particularly important for learning processes because they point at possible deficiencies of trainees. This analysis also provides information about the main reasons for failing since the group of low performers only consists of failures in contrast with the other two groups (see chapter 10). We counted the occurrences of insufficient competence ratings (value < 4), we averaged them, and we present them in graphs, separately for each group.

However, both analyses ignore the influence of time while trainees may differ in development of specific competences. We examined possible differences between trainees in competence development by calculating rank order correlations (Spearman) of the variable 'time' (i.e., successive moments of progression measurements) with the competence ratings. We did the analyses separately for the three groups in order to compare the groups with each other. This should provide information about the profiles of competence ratings within each group. We expect that low performers (failures) show decreasing performance, especially for less trainable competences.

Prediction of pass-fail

We explored which competences predict pass-fail better than other competences. We should realize that the number of trainees is too low yet for getting dependable evidence of predictive validity.

¹⁷ The data structure can be considered to be a multilevel data structure (e.g. Hox, 2002): repeated and longitudinal measures (competence ratings) for a group of cases (trainees). We did not apply multilevel analyses here because this would be very complicated in a technical sense for the specific purposes of this study, and we are not interested in the possible influence of other variables at a higher level: usually an important goal of multilevel analysis.

First, we did a discriminant analysis to explore how the competence ratings, averaged over time, relate to pass-fail (criterion variable) and to which extent this classification was made correctly¹⁸. This should lead to more insight into the main reasons for failing. We also did an analysis of variance (ANOVA) to examine whether the mean competence ratings differ significantly between pass and fail. Additionally, we correlated the competence ratings, averaged over time, with the pass-fail criterion (Pearson). We expect that less trainable competences are more strongly related to pass-fail than others. Finally, we did a qualitative analysis of the main reasons for failing using qualitative information (i.e., textual remarks on progression reports, oral information from assessors). This plays an essential role in pass-fail decisions (see chapter 7). We have built up history of trainees since five years and we analysed the reports of trainees who had problems or who failed (N=36; ACC and TWR/APP together).

11.3 Results

Individual differences in competence development

Differences in mean competence ratings

Table 11.1 and 11.2 present the results of an analysis of variance (ANOVA) across trainees for respectively ACC preOJT and OJT for the three groups high, moderate and low performers. The results presented in Table 11.1 show that the largest differences in mean competence level were found within the group of moderate performance in preOJT. This confirms the findings of the analysis of learning curves (see chapter 10): this group has many variations in learning, especially in contrast with the group of low performers. However, this does not clearly appear from the results of OJT presented in Table 11.2. Table 11.2 presents more significant differences in mean competence level across trainees in the various groups in OJT, especially for low performers. This can be influenced by the fact that the entrance level is not as similar as in preOJT since this depends on the final competence level in preOJT.

The findings in both Tables 11.1 and 11.2 do not support our expectation that trainees differ more in less trainable competences. For instance, in preOJT the differences in *attitude* and *team orientation* are not significant at $p < .001$ within the group of high performers but they are within the group of moderate and low performers. *Label management* in preOJT is significantly different between trainees within each group in contrast with less trainable competences such as *planning*. *Planning* in OJT is significantly different across trainees as well as *attitude* and *team orientation* within the group of moderate performers, but this is not the case with other critical competences such as *mental picture* and *decisiveness*. Apparently, many indi-

¹⁸ A multinomial logistic regression, with the three groups as dependent variable, would be a good alternative. However, this was not possible, probably due to a too low number of cases in relation to the variables.

vidual differences in averaged competences exist across trainees, even within groups, but the results do not point at a differentiation between competences with respect to trainability. This can be influenced by the intercorrelations between the competence ratings (see section 9.2). We also did an analysis of variance (ANOVA) for the whole group of trainees in both preOJT and OJT (not presented here). As expected, all competences are significantly different at $p < .001$ in preOJT and OJT.

Table 11.1. Analysis of variance (ANOVA) mean performance level across trainees in preOJT

Competence	High performers (df1=11, df2=62)		Moderate performers (df1=9, df2=54)		Low performers (df1=8, df2=35)	
	F	Sig.	F	Sig.	F	Sig.
Safety	3.175	.002	2.075	.048	.931	.503
Efficiency	2.241	.023	7.377	.000	2.555	.026
Verbal expression	1.658	.105	8.710	.000	2.745	.018
Listening	2.348	.017	11.167	.000	2.262	.046
Co-ordination	5.370	.000	14.217	.000	3.098	.011
Equipment op.	2.529	.011	8.359	.000	1.665	.142
Label management	2.892	.004	8.136	.000	4.422	.001
Mental picture	3.806	.000	2.888	.007	2.202	.051
Attention man.	5.746	.000	6.177	.000	1.737	.124
Planning	2.487	.012	1.316	.251	2.597	.024
Decisiveness	2.949	.003	4.360	.000	1.206	.324
Workload man.	5.407	.000	4.166	.000	1.813	.108
Attitude	1.506	.152	11.114	.000	5.014	.000
Team orientation	1.092	.383	9.455	.000	4.399	.001

Table 11.2. Analysis of variance (ANOVA) mean performance level across trainees in OJT

Competence	High performers (df1=9, df2=142)		Moderate performers (df1=5, df2=98)		Low performers (df1=10, df2=129)	
	F	Sig.	F	Sig.	F	Sig.
Safety	.875	.549	3.310	.008	3.935	.000
Efficiency	5.435	.000	2.762	.022	3.908	.000
Verbal expression	7.444	.000	11.050	.000	15.040	.000
Listening	6.161	.000	4.688	.001	9.377	.000
Co-ordination	5.109	.000	5.042	.000	14.916	.000
Equipment op.	7.238	.000	1.672	.148	8.097	.000
Label management	2.220	.024	3.023	.014	8.840	.000
Mental picture	3.352	.001	.936	.461	5.086	.000
Attention man.	5.188	.000	2.003	.084	6.920	.000
Planning	5.108	.000	5.123	.000	11.449	.000
Decisiveness	2.406	.014	1.406	.228	9.149	.000
Workload man.	3.659	.000	2.196	.060	2.651	.006
Attitude	15.411	.000	22.917	.000	1.572	.000
Team orientation	1.662	.000	5.594	.000	5.498	.000

Differences in insufficient competence ratings

We calculated the occurrences of insufficient competence ratings and we made graphs of these insufficient ratings (in percentages) for preOJT and OJT, averaged per trainee, see Figures 11.1 and 11.2. The results suggest that the occurrences of insufficient ratings differentiate strongly between groups. In general, the number of

insufficient ratings is around twice as high in preOJT as in OJT. This can be explained by the fact that the requirements in OJT are higher. In OJT, trainees must be able to avoid any mistake when they are working independently with real-life traffic. Insufficient ratings are not allowed anymore while this is allowed in preOJT to a certain extent. In preOJT, the low performers have at least 50% insufficient ratings for respectively: *mental picture*, *safety*, *workload management*, *attention management*, *efficiency* and *decisiveness*. Especially *mental picture* separates low from high performers clearly. Moderate and high performers never have 50% or more insufficient ratings. In OJT, the low performers have at least 30% insufficient ratings for respectively: *listening*, *co-ordination*, *decisiveness*, *mental picture*, and *attention management*. As expected, in preOJT and OJT the less critical competences are almost never rated as insufficiently: *equipment operation*, *attitude* and *team orientation*. There are differences in competences between preOJT and OJT: it is remarkable that *listening* and *coordination* are rated more often as insufficiently in OJT than in preOJT. The other competences appear to be rather comparable in both training phases.

Figure 11.1. Averaged sum of insufficient competence ratings, averaged per trainee (N=34), for preOJT

Figure 11.2. Averaged sum of insufficient competence ratings, averaged per trainee (N=27), for OJT

Differences in competence development

Tables 11.3 and 11.4 present the rank order correlations (Spearman) of the variable 'time' with the competence ratings averaged per trainee, respectively for preOJT and OJT (two learning phases). The results summarized in Table 11.3 clearly present the differences in competence development between the three groups for preOJT, taking into account that the standards are recalibrated into the same rating scale. The high performers have positive significant correlations for many competences in contrast with the low performers. This implies that competences of high performers develop over time in contrast with the low performers. Thus, these findings are in conformance with the findings of the analyses of learning curves (see chapter 10). The largest differences between the three groups are respectively found for *mental picture*, *planning*, *safety* and *decisiveness*. This was expected because they belong to the less trainable competences. The moderate performers take an intermediate position: the competence ratings stay rather similar over time.

The findings for OJT, presented in Table 11.4, generally show positive correlations, especially in learning phase 2. This can be explained by the fact that trainees are assessed against phase level: they will pass the learning phase if they have achieved the standards. The low performers who fail in learning phase 2 show progression in learning phase 1 towards the required level, otherwise they would have failed in learning phase 1 already. The fact that the requirements in learning phase 2 are higher than in learning phase 1 may explain why moderate performers in learning phase 2 show more progression towards the required level than in learning

phase 1. In this way, the findings presented in Table 11.4 confirm the results of analyses of learning curves as well (see chapter 10). Differences between competences are not as clear as in preOJT, although attitude and team orientation, more trainable competences, stay rather constant over time and do not differ much across groups. However, this can also be influenced by the high correlations between the competence ratings as mentioned before (section 9.2).

Table 11.3. Rank order correlations (Spearman) between time and competence ratings for preOJT

Competence	High performers (N = 74)	Moderate performers (N = 63)	Low performers (N = 44)
Safety	.32**	-.13	-.29*
Efficiency	.43**	-.02	-.22
Verbal expression	.09	-.16	-.19
Listening	.17	-.10	-.24
Co-ordination	.27*	.03	-.04
Equipment operation	.11	-.01	.11
Label management	.23*	.09	.04
Mental picture	.33**	-.14	-.54**
Attention management	.35**	-.11	-.19
Planning	.26*	-.20	-.44**
Decisiveness	.28**	-.11	-.30*
Workload management	.07	-.06	-.19
Attitude	-.27*	-.15	-.17
Team orientation	-.02	-.14	-.11

** Correlation is significant at the 0.01 level (1-tailed); * Correlation is significant at the 0.05 level (1-tailed)

Table 11.4. Rank order correlations (Spearman) between time and competence ratings for OJT

Competence	High performers		Moderate performers		Low performers	
	Phase 1 (N = 71)	Phase 2 (N = 81)	Phase 1 (N = 46)	Phase 2 (N = 64)	Phase 1 (N = 85)	Phase 2 (N = 46)
Safety	.12	.23*	-.25	.37**	.21*	-.22
Efficiency	.33**	.28**	-.09	.50**	.27*	.00
Verbal expression	.24*	-.07	-.00	.42**	.18*	.11
Listening	.23*	.15	-.16	.25*	.18*	.13
Co-ordination	.19	.28**	-.20	.27*	.17	.10
Equipment op.	.05	.03	-.06	.30**	.28**	.01
Label management	.24*	.13	.06	.20	.26**	.05
Mental picture	.22*	.26*	.11	.41**	.31**	.02
Attention man.	.31**	.39**	.01	.45**	.26**	-.07
Planning	.13	.08	-.11	.30**	.31**	.04
Decisiveness	.12	.38**	-.10	.48**	.22*	.11
Workload man.	.28*	.04	.22	.51**	.10	-.04
Attitude	-.16	-.31**	-.06	.33**	.18	.08
Team orientation	.03	.07	.03	.08	.10	.16

** Correlation is significant at the 0.01 level (1-tailed); * Correlation is significant at the 0.05 level (1-tailed)

Prediction of pass-fail

Discriminant analysis

We did a discriminant analysis to explore how the mean competence ratings predict pass-fail and to which extent this classification was made correctly. We also did an analysis of variance (ANOVA). In preOJT, the discriminant function is significantly different across groups (chi-square = 26.26, $df = 14$, $p = .024$). The discriminant function coefficients indicate that the following competences are respectively the best predictors: *efficiency, workload management, mental picture, attention management, decisiveness*. The three worst predictors are respectively: *team orientation, attitude, verbal expression*. These findings are in accordance with our expectations. In total, the discriminant function successfully predicted group membership for 97.1%, which is a very high percentage (see Table 11.5):

Table 11.5. Predicted group membership for pass-fail in preOJT

	Pass	Fail	Total
Pass	24	1	25
Fail	0	9	9

An analysis of variance (ANOVA) shows that all competence ratings are significantly different at $p < .01$ for pass-fail with a few exceptions: *verbal expression* ($F = 4.258$, $sig. = .047$), *equipment operation* ($F = 5.239$, $sig. = .029$), *attitude* ($F = 2.733$, $sig. = .108$), *team orientation* ($F = 1.722$, $sig. = .199$). In OJT, the discriminant function is also significantly different across groups (chi-square = 23.64, $df = 14$, $p = .051$). The discriminant function coefficients indicate that the following competences are respectively the best predictors for pass-fail: *listening, mental picture, safety, attention management, workload management*. The three worst predictors are respectively: *attitude, equipment operation, label management*. As in preOJT these findings are in accordance with our expectations. In total, the discriminant function successfully predicted group membership for 96.0% (see Table 11.6):

Table 11.6. Predicted group membership for pass-fail in OJT

	Pass	Fail	Total
Pass	15	1	16
Fail	0	9	9

An analysis of variance (ANOVA) shows that all competence ratings are significantly different at $p < .01$ for pass-fail with a few exceptions: *co-ordination* ($F = 5.760$, $sig. = .025$), *equipment operation* ($F = 2.547$, $sig. = .124$), *label management* ($F = 2.580$, $sig. = .122$), *decisiveness* ($F = 5.950$, $sig. = .023$) *attitude* ($F = .735$, $sig. = .400$), *team orientation* ($F = 3.571$, $sig. = .071$).

Correlations with pass-fail

Table 11.7 presents the correlations (Pearson) of the variable pass-fail with the mean competence ratings. Table 11.7 shows that the highest correlations in preOJT are respectively found for: *efficiency*, *mental picture*, *workload management*, *attention management* and *decisiveness*. These are respectively for OJT: *safety*, *mental picture*, *workload management*, *attention management*. Many correlations are high, probably influenced by high intercorrelations between the competence ratings.

Table 11.7. *Correlations (Pearson) between pass-fail and competence ratings for preOJT and OJT*

Competence	PreOJT (N = 34)	OJT (N = 27)
Safety	.68**	.68**
Efficiency	.74**	.55**
Verbal expression	.34*	.25
Listening	.55**	.48**
Co-ordination	.64**	.25
Equipment operation	.38*	.30
Label management	.47**	.26
Mental picture	.73**	.64**
Attention management	.72**	.57**
Planning	.62**	.55**
Decisiveness	.70**	.16
Workload management	.73**	.60**
Attitude	.28	.18
Team orientation	.23	.29

**Correlation is significant at the 0.01 level (1-tailed); *Correlation is significant at the 0.05 level (1-tailed).

The several findings of the quantitative findings in this section suggest that mental picture appears to be the competence that most strongly distinguishes between passed and failed trainees, both in mean competence level and in competence development over time. Other critical competences seem to be: *workload management*, *attention management* and *decisiveness*, and the least critical competences seem to be: *equipment operation*, *attitude* and *team orientation*.

Qualitative analysis of main reasons for failing

We analysed the main reasons for failing in a qualitative way for 36 trainees (both ACC and TWR/APP). Although these reasons for failing are often the same and strongly related to each other, we can distinguish four main problems, apparently founded in abilities and personality.

Situational awareness (mental picture, attention management):

Trainees who don't have overview of traffic situations ('the picture'), especially in more complex and busy traffic. They do not see enough patterns in aircraft, they miss potential conflicts, and they cannot divide their attention over several situations. Coaches say that they do not 'see it'; they do not oversee the complete traffic situation. They tend to focus on specific (conflict) situations too much and have problems with monitoring and scanning. Sometimes they cannot keep situations in their mem-

ory for some time (e.g., missing calls and readbacks from pilots). Consequently, they cannot anticipate on future traffic situations adequately, and they tend to work 'ad-hoc' without a plan how to handle the traffic. Besides the traffic situation itself, they miss information in their working environment. Traffic, pilots, colleagues or events may surprise them.

Planning and creative problem solving

Trainees who appear to have overview of traffic situations, but who are not able to find effective and efficient solutions for a diversity of (conflict) situations. They don't have developed adequate 'strategies' for conflict solving. They have problems with balancing between safety and efficiency, and they do not have a large variety of solutions for different situations. Coaches are afraid to let these trainees work independently, because they appear to be not able to solve novel situations with which they did not have coped before. These problems do not only occur in complex or busy traffic, but most often in non-standard and unexpected situations. They tend to rely on fixed rules and procedures. A related problem is that they may be not flexible enough to think about alternatives and switch to another plan when the first solution does not work adequately.

Decisiveness

Trainees who appear to have 'the mental picture', who have insight into the solutions for specific (conflict) situations, but who do not take enough initiative. They hesitate too much, are too passive and do not take action quickly. Their self-confidence may be rather low. Further, they do not show enough 'leadership', needed to be a controller. They appear to be not used to this role personally. Consequently, their work tempo may become too low and they are getting behind.

Workload management

Trainees who are sensitive to stress in more complex situations. They are not able to manage their own workload adequately, especially with increasing complexity of traffic. They appear to be insecure and not self-confident. Stress reactions may occur that influence their decision making substantially. They tend to get overloaded rather easily: a 'tunnel vision' may occur and they cannot think logically anymore. Generally, these trainees cannot switch between working methods that are dependent on the traffic load and their own workload (workload management strategies), which affects the balance between safety, efficiency and their 'personal efficiency' (mental workload).

These four main problems are not easy to separate because they are highly connected and influence each other continuously. Nevertheless, we notice a substantial differentiation between individuals. Table 11.8 presents a classification of problems of trainees (N=36), made by an independent training expert from LVNL. Table 11.8 supports the findings of the quantitative analyses: the main reasons for failing appear to be respectively workload management, mental picture, planning and attention management.

Table 11.8. *Frequencies of reasons for failing (competences) in qualitative analyses of failed trainees*

Competences	Frequencies
Safety	16
Efficiency	8
Communication	10
Co-ordination	5
Strip/label management	1
Equipment operation	-
Mental picture	19
Attention management	17
Planning	18
Decisiveness	13
Workload management	24
Attitude	4

11.4 Discussion

This chapter aimed at answering the two research questions: (1) "Does the assessment system sufficiently represent individual differences in competence development so that training can be maximally adapted to trainees' needs?" and (2) "Which competences are most predictive for pass-fail decisions (main reasons for failing)?" We conclude that the assessment system sufficiently distinguishes in competence development between trainees, and also between failed versus passed trainees. These are important requirements of the assessment design, because assessment should support learning processes and should be predictive valid. Therefore, the assessment results should reflect learning in an optimal way. However, a larger number of data is needed to get convincing evidence of predictive validity.

Individual differences in competence development

We examined individual differences in mean competence level and in competence development over time in several ways. We found significant differences in mean competence level between individuals within the three groups of high, moderate and low performers and for the total group of trainees, but not for all competences. We found larger individual differences in OJT than in preOJT. This can be explained by the fact that OJT is less structured than preOJT. In OJT trainee performance does not only depend on the competence level, but also on the task situations that vary much. For instance, we expect to find lower competence ratings in more difficult task situations. Furthermore, the entrance competence level in preOJT is more comparable across trainees than in OJT as we explained before.

We expected larger differences in competences that are more critical and less trainable, such as *mental picture* and *attention management*, than in competences that are less critical and more trainable, such as *label management* and *equipment operation*. The first group of competences is supposed to be more dependent on innate abilities of individuals because they comprise more non-consistent task components as explained (Schneider, 1990; Regian & Schneider, 1990; Ackerman, 1989). However, a comparison of their means did not show a clear difference be-

tween competences. The findings also showed that insufficient competence ratings differentiate better than their means, influenced by the high intercorrelations between competences (see section 9.2). We consider this differentiation sufficient for detecting possible deficiencies of trainees. The total number of insufficient competence ratings is much lower in OJT than in preOJT because of the higher standards in OJT: all competences must be rated as sufficiently to pass (see chapter 7, section 7.4). It is remarkable that *listening* and *coordination* are rated as insufficiently more in OJT than in preOJT. This can be explained by the fact that these two competences cannot be trained to such a high level in simulator training (preOJT) as in OJT. In preOJT, other trainees, pseudo-pilots or instructors play the controllers to coordinate with, but the reality is different. Listening is a problem in OJT more often, probably because the working environment, with more noise and interaction with others, may hinder this. We can say that these differences are founded in learning processes: *listening* and *coordination* are mainly acquired in OJT and not in preOJT yet.

Furthermore, we found clear differences in competence development between the three groups of trainees (high, moderate and low performers) over time as expected. The low performers show decreasing performance on the majority of competences in contrast with the moderate and high performers. The differences between the three groups are the highest for less trainable competences, such as *mental picture*, *planning*, and *attention management*, in conformance with our expectations. We also found substantial differences for *workload management* and *decisiveness*. These findings support the idea that these two competences are dependent on personality properties such as dealing with stress-critical situations (Hettema, Leidelmeijer & Geenen, 2000; Kuhl, 1994) and traits such as stress resistance, vigilance and dominance that are part of selection requirements (e.g., Eissfeld, Heil & Broach, 2002).

Prediction of pass-fail

The various analyses (discriminant analysis, correlations of competence ratings with the pass-fail criterion, qualitative analysis) showed that pass-fail decisions are predicted rather correctly in preOJT and OJT, based on the competence ratings. We may conclude that the following competences are respectively the main reasons for failing: *mental picture*, *workload management*, *attention management* and *decisiveness*. Especially *mental picture* does not only differentiate in mean competence level but also in competence development (progression) over time.

These findings do not completely agree with the findings of MRU (1993). For instance, in their study *workload management* was ranked less high while theoretical knowledge played a more important role. However, their results were not based on Unit Training but on Initial Training that precedes Unit Training. Initial Training focuses on theoretical knowledge and training of basic skills. It is reasonable to assume that workload management becomes more relevant in Unit Training, especially in OJT. Workload management involves coping with real-life situations. Switching between low and high workload (cf. vigilance) and the feeling of responsibility (cf. stress resistance) are crucial for successful air traffic management. The findings agree with the idea that situational awareness, to which *mental picture* and *attention*

management belong (cf. ATC Performance Model; see chapter 5), heavily depends on innate abilities such as spatial orientation and memory capacities (Shebilske, Goettle & Garland, 2000; Garland, Stein & Muller, 1999).

Practical implications

The fact that the assessment system sufficiently distinguishes between individual trainees with regard to their competence development has some practical implications.

We argued that one of the main purposes of assessment is to identify trainees' strengths and weaknesses, founded in competences. This serves as feedback to the trainees in the first place so that they know which competences they have to improve. It helps to repair possible deficiencies by offering specific coaching, remedial teaching, adapted task selection and other interventions. Our findings suggest that this may be possible in the near future. It implies that this makes assessments diagnostic, based on the detection of causes for insufficient performance (Regian & Schneider, 1990; Seamster, Redding & Kaempff, 1997; see section 4.1). However, we assume that competences can be improved by training, but to a certain extent only. As argued, situational awareness can be acquired by training attention management strategies (Shebilske, Goettl & Garland, 2000), or by applying chunking methods (Garland, Stein & Muller, 1999). Workload management requires the acquisition of strategies (Sperandio, 1978).

The quality of coaching in simulator training and in OJT adds to improvement of competences. However, often the focus is too much on assessment. Coaches have a double role as a coach and assessor in our system (see section 7.4), but it appears to be difficult for them to separate these roles. We have started assessor training to improve their assessment skills (see chapter 8). In addition, we teach them how to coach on specific competences such as situational awareness and workload management. This is difficult, because many of them are conceptually related and highly correlated with each other (see section 9.2). Learning processes, especially in later stages such as OJT, result in an integration of competences. These relationships should be taken into account when coaching on specific competences. Finally, the acquisition of competences such as situational awareness and workload management relates to strategy training. The acquisition of strategies is supposed to be a challenge and a key for success in training (Nunes & Mogford, 1993; Schneider, 1990; see also section 3.2).

In addition, task selection can be made more adaptive to the trainees' needs, especially in the simulator, by differentiating in task sequence, task complexity, and length of training. Possibilities in OJT are to lengthen the training period, to put trainees back in the simulator to train specific task situations, or to let them work on specific working positions or at specific periods. Additional differences between trainees that should be taken into account in task selection are personal characteristics such as learning styles and emotional states. This relates to the principle of 'dynamic task selection' (Salden, 2005; Salden et al., 2004, 2006; see section 3.3). It implies that task selection is adjusted to the performance level of trainees and that the cognitive load (cf. mental effort) that is imposed is controlled. If self-directed learning skills are

developed well, trainees can also have an input in the choice of the learning tasks as well (Van Merriënboer & Kirschner, 2007). However, adaptive task selection is difficult to realize in practice. In addition to problems with analysing trainees' personal characteristics, practical problems play a role such as the restricted availability of the simulator and organizational problems around the training schedule. Finally, more research is needed to the relationship between learning tasks and competence development. Studies like those of Salden et al. (2004, 2006) are executed in a laboratory situation with novices as trainees. In reality, especially in OJT, usually more variables are involved that are difficult to control.

The findings in this study confirm that most critical competences seem to be more predictive for pass-fail than others. We have started to use the assessment results of previous trainees for making more profound pass-fail decisions (see section 7.4). However, we explained that we need more data for defining a cut-off. Under the assumption that the competence ratings will be sufficiently reliable in the future, cut-offs can not only be based on overall performance and progression (see chapter 10), but also on a weighting of competence ratings. In this context, we should take into account trainability of competences. Ultimately, it would also be possible to improve selection instruments based on evidence of success in training, using the competences as criteria in validation research (Roe, 2005, 2007).

11.5 Conclusions

We conclude that the assessment system sufficiently represents individual differences in competence development. This makes it possible to provide adequate feedback to trainees and to maximally adapt training to their needs. It implies that the requirement of supporting learning processes is met. We discussed differences in trainability between competences. For instance, situational awareness and workload management seem to be very critical and are found to be the main reasons for failing in ATC training in this study. The next step is to put more effort into research on how to train competences better by specific coaching, task selection and other measures. In addition, the findings with respect to criterion-related validity contribute to the psychometric quality of the assessment system (see also chapter 9). Pass-fail decisions were correctly predicted by the competence ratings. More longitudinal research is needed to get convincing evidence for predictive validity, and ultimately for defining a cut-off.

12 User evaluation

The main purpose of the user evaluation was to evaluate the practical usability of the assessment system as perceived by the users (trainees, coaches, and managers). This evaluation has taken place continuously and has been alternated with the various implementation activities described in chapter 8. This is in contrast with the (quantitative) evaluations presented in chapters 9, 10 and 11 that were mainly executed at the end of the design process. This chapter discusses the evaluation methods (section 12.1) and the main results (section 12.2) that have led to some adjustments of the assessment system.

12.1 Methods

We applied multiple methods to evaluate the usability of the assessment system. First, all trainees receive a questionnaire to evaluate each course in general. One part of this questionnaire involves assessment and contains questions about the practical usability of assessment reports and procedures, the role of assessors, and the clearness and univocal meaning of performance standards and criteria. The questions about assessment for ACC OJT, as part of the evaluation questionnaire as a whole, is presented in Table 12.1 as an example (translated from Dutch into English).

Second, trainees are interviewed during training to evaluate training; assessment is part of this interview. Third, the usability of the assessment system is evaluated in several group meetings with coaches and training managers. Fourth, ACC coaches received a specific questionnaire about assessment in 2004. This questionnaire comprises comparable questions as presented in Table 12.1. Finally, the most valuable information is received in informal interviews with coaches and trainees, who are our colleagues. Table 12.2 provides an overview of the evaluation methods for the various user groups (a separate evaluation was held for the web-based assessment tool as discussed in section 7.5).

Table 12.1. Questions about assessment as part of the evaluation questionnaire for OJT

Assessment in OJT	Totally disagree	Disagree	No opinion	Agree	Totally agree
The objectives to be obtained at the end of each phase are clear.					
The behavioural examples help me to find out points of attention.					
The final level of preOJT agrees with the starting level of OJT.					
The differences between learning phase 1 and 2 are clear and fine.					
The length of the different phases is well-chosen.					
The manual is clear, understandable and complete.					
Progression reporting					
My strengths clearly appear from the assessments.					
My weaknesses clearly appear from the assessments.					
My progression clearly appear from the assessments.					
The assessments make clear what I have to improve and how.					
The coaching is adapted to the findings in the assessments.					
The assessments were objective.					
The assessors understand the procedures of assessment correctly.					
The agreement between assessors is sufficient.					
The assessors pay enough attention to the successive standards.					
The assessors provide a useful debriefing based on assessments.					
The frequency of reporting is sufficient.					
Assessment reports					
The categorization of competences and performance criteria is fine.					
The meaning of competences and performance criteria is clear.					
It is useful that the same competences are assessed as in preOJT.					
The behavioural examples help to know the meaning of the criteria.					
The rating scales applied are fine.					

Table 12.2. Overview of evaluation methods with the users

	Trainees	Coaches	Managers
Questionnaire	X	X	
Interviews (formal and informal)	X	X	X
Group meetings		X	X

12.2 Results

Trainees, coaches and managers have been very positive about the renewed assessment system in general. They have concluded that the system is practically usable. We summarize the main results, pertaining to ACC, by classifying them in the main issues that have been evaluated in various ways.

Competences and performance criteria

The competences and performance criteria have been evaluated in great detail, especially in meetings with coaches and managers. Many criteria were removed,

added or changed in the first year after the implementation of the assessment system. The use of the same performance criteria makes trainees and coaches more familiar with their meaning. Coaches and trainees argued that this has increased the agreement between assessors. The competences are recognizable for assessors and cover ATC performance sufficiently. However, it took some time before *progression* and *general impression* (see progression reports in Appendix B) were understood correctly. We explained these issues in an accompanying text in the manual and in refresher training of coaches. The weighting of performance criteria in simulator tests has been evaluated by tryouts of the tests and by comparisons of the assessor's subjective impression with the final test scores. A few versions of the weighting scheme were made because in the first tests the differences between the two scores were too high in general (see chapter 7).

Rating scales

Generally, the coaches and trainees have been happy with the 6-points rating scale in ACC progression reports and in simulator tests. The fact that insufficient ratings are clearly separated from sufficient ratings makes that strengths and weaknesses of trainees are directly clear, while motivation and argumentation can be given with textual comments.

Performance standards

Trainees emphasized the benefits of the defined performance standards to be attained at the end of each phase, especially in OJT. It is clearer for them what they have to learn in a specific phase. Coaches and assessors confirmed that they have a more common picture about the requirements in each training phase. This has certainly increased the agreement between them, although there will always exist grey areas in which assessors will differ from each other. Assessors rely on the global description of each phase in terms of the three principles traffic handling, traffic complexity, and aid of the coach. The behavioural examples that represent the performance standards have intensively been used in the first period of implementation. Now, they have become most useful when trainees do not perform well; then the coaches rely on arguments that they can find in the performance standards. New coaches need the behavioural examples to get familiar with the standards and with the meaning of each competence.

Structure of training (phasing)

Phasing was especially new in OJT. Coaches and trainees argued that the renewed structure has improved the OJT substantially. The division into phases in combination with the performance standards is considered to be obvious and clear. The length of each phase appears to be well-chosen. The structure and the objectives of each phase are rather clear for coaches and trainees.

Procedures

There have not been problems with the procedures at all; they appear to be very clear. We repeated the instruction in refresher training for coaches, and we have

explained the procedures to trainees in detail at the start of each course to ensure that trainees and coaches are kept informed.

Layout of assessment reports

Coaches and trainees haven't had problems with the layout of the assessment reports. Only some small improvements were made after a few months of implementation.

Feedback towards trainees

Trainees indicated that they get better feedback about their strengths and weaknesses. The textual comments are more specific, probably not only due to the renewed assessment system, but also due to the assessor training (see chapter 8). Coaches indicated that they can more easily express and validate their 'gut feeling' with aid of the competences. They can provide better arguments for (in)sufficient performance. This makes it easier to define appropriate training interventions and to give tips to trainees how to improve performance. It is reasonable to assume that the involvement of controllers in the design process has contributed to a better recognition of behaviour that refers to a specific competence, formulated in their own jargon. Furthermore, coaches argued that they can more easily follow trainee's progression on each competence during a certain training period and in different task situations. They have more insight into the trainee's individual learning process. This improves the feedback towards trainees and makes training interventions more effective.

Profound pass-fail decisions

More information is available to make profound pass-fail decisions, such as the history of training results of previous trainees and more insight into reasons of failing, aided by the web-based assessment tool. This makes these decisions more valid according to training managers. Training results can be compared with those of previous trainees, and managers have a better overview of trainee's performance. In addition, coaches provide better arguments for (in)sufficient performance, and they can more easily express their 'gut feeling' as mentioned above, helping with the argumentation for pass-fail decisions. More insight into learning processes, such as the degree of progression, provides information whether the trainee is still learning or has achieved a final learning plateau. This information is essential for considering the usefulness of further training.

12.3 Conclusions

We have qualitatively evaluated the practical usability of the assessment system with coaches, trainees and managers. The conclusion is that the system is practically usable and that it has improved learning processes. The set of competences and performance criteria, accompanied by the behavioural examples that represents the performance standards for each training phase, contribute to consistent judgments between assessors. Its clarity improves feedback during training and results in appropriate training interventions. Finally, pass-fail decisions are more profound be-

cause assessors can provide better arguments and because these decisions are based on the history of results of trainees in earlier training; we have gained more insight into the main reasons for failing.

13 Evaluation of Program of Requirements

The Program of Requirements (PoR) was constructed in 2001, in collaboration with the group of stakeholders, and has guided the redesign of the LVNL assessment system (see chapter 6). This section discusses how we evaluated the realized assessment system against this PoR. It serves as the final evaluation of the quality of the assessment system as a whole in accordance with the design cycle of Roozenburg & Eekels (1991) and our design methodology as explained in chapter 1 (section 1.4). First, we explain the methods (section 13.1); next, we describe the results (section 13.2); finally, we discuss the main findings (section 13.3).

13.1 Methods

We evaluated the PoR with a small part of the group of stakeholders, the two training managers of the two training trajectories in Unit Training (ACC and TWR/APP). It would have been better to do this with the whole group of stakeholders who helped to draft the PoR in 2001, but the majority of stakeholders consisted of managers who were replaced by others at the moment of evaluation against the PoR (July, 2007). The two training managers can be considered as the main stakeholders because they are the persons who are responsible for Unit Training. It should be noted that this evaluation is not restricted to ACC, main case in this study, but that it also includes the TWR/APP trajectory within Unit Training since the PoR was made for this whole scope.

First, the training managers evaluated the PoR independently of each other by providing ratings at a 5-points scale from 1 (not achieved at all) to 5 (completely achieved) for each requirement and constraint in the PoR. Next, a meeting was organized to discuss the individual findings. The final ratings for the whole PoR were made together, added by arguments. This was helpful because many requirements and constraints can be interpreted in a different way, and sometimes the results were different for the two trajectories TWR/APP and ACC. For the same reasons, we only present the final ratings as defined in the meeting. The comments are also based on the findings of the other evaluation parts, which were presented earlier (chapters 9 – 12).

13.2 Results

Table 13.1 and 13.2 present the final ratings for each requirement and constraint in the PoR at the 5-points rating scale as made in the meeting with the training managers (see for a detailed formulation of the requirements and constraints: Table 6.1). For each item of the PoR an evaluative comment is added.

Table 13.1. *Evaluation of the requirements of the PoR*

Requirements	Score ¹⁹
R1 Quality of controllers	4
The assessment system has partly contributed to the deliverance of sufficiently qualified trainees from preOJT to OJT and from OJT to OPS besides other factors. The pass rates are still too low, especially in OJT, since the implementation of the assessment system in 2003. In ACC they are until the end of 2006: 74% in preOJT (N = 34) and 59% in OJT (N = 27). The overall pass rates are comparable with the TWR/APP trajectory. The OJT pass rates strongly have increased in 2006 (around 80% in various courses within Unit Training). The shortage on controllers has remained, but there are also other factors such as a too low number of trainees entering in Unit Training due to a too low income in selection and a too low pass rate in a selective module in Initial Training. Besides, the requirements for trainees have become higher over time due to the increasing complexity of the ATC task. The quality of trainees entering in OJT and of trained personnel certainly is sufficient: the final OJT level is clearly defined.	
R2 Pass rate	3
The required pass rate in OJT (90%) is not obtained yet, although it is increasing especially since 2006 (see R1). Therefore, we could say that performance in preOJT does not predict performance in OJT sufficiently yet. Performance in the first modules does not predict performance in subsequent modules very well, although some starting modules do not aim to predict performance (e.g., trainees cannot fail in the first assistant module (VLA) in ACC; it only serves as an internship). However, the number of trainees who followed the whole trajectory is too low yet for definite conclusions. More qualitatively, we have built up history of trainees that has helped to predict future performance better since 3 years.	
R3 Competences	5
The assessment system is based on competences defined in the workshops while maintaining the jargon of controllers for recognizability of behaviour in assessments. Performance criteria and performance standards to be achieved at the end of each phase were derived, including the final competence level. The assessment system measures to which extent trainees have acquired the competences in a certain stage. This implies that the 'gut feeling' of controllers has been explicated in terms of competences. This helps them to identify trainees' strengths and weaknesses better, needed for adequate feedback and appropriate training interventions. Competences optimally reflect individual abilities to perform effectively and include technical, cognitive, emotional and social aspects to get a complete picture about trainees' performance. In this way, the assessment system certainly contributes to sufficiently competent controllers as output from training.	

¹⁹ The anchors of the 5-points rating scale are: 1 = not achieved at all; 2 = hardly achieved; 3 = partly achieved; 4 = largely achieved; 5 = completely achieved.

Requirements	Score
R4 Learning processes	4
<p>Performance standards to be achieved at the end and in intermediate phases were derived from training objectives. They make clear to trainees and coaches what has to be acquired in a phase in preOJT and in OJT. Training interventions such as retraining, remedial teaching, alternative trajectories and specific coaching are based on trainee's deficiencies that follow from assessments. However, the focus is on assessment still too much since we do not know yet how to improve deficiencies related to specific competences (cf. situational awareness, workload management) or how to train strategies better by more adaptive/individual training (task selection, coaching). Learning curves were made of trainees' results and represent learning processes adequately in accordance with learning theory. They can be compared with the average or optimal curve to indicate how much has to be learned. Individual differences in learning, such as slow starters and learning plateaus, can be identified and progression can be assessed to a certain extent. But these learning curves have not been used yet in practice, only for scientific research. Progression is also rated separately, and it is assessed in a more qualitative way. However, its meaning is interpreted differently by some assessors (e.g., the amount of progression during the period of assessment, or sufficient progression with respect to the required learning curve). Some problems remain exist with the reliability of the competence ratings due to different interpretations and rating errors.</p>	
R5 Pass-fail decisions	4
<p>Pass-fail decisions are built in early training phases, although the primary aim is not selection, in order to avoid useless effort and costs in further training of prospectless trainees. For instance, the end of preOJT is always a pass-fail moment, and during preOJT and OJT pass-fail decisions can be made continuously. However, predictive validity has not been proven (see R1, R2). Trainees still fail at the end of OJT. Pass-fail decisions are based on a profound quantitative and qualitative analysis of trainees' performance; history built up already helps to make these decisions more valid. Procedures and possibilities for failed trainees are clearly described in the manual and they are also presented to the trainees before they start with the training.</p>	
R6 Following trainee's progress	4
<p>Trainees' progression is followed quantitatively by storing assessment results in a database, and qualitatively by an exchange of information with coaches and trainees. The fact that the same competences are assessed from selection to ITr and UT adds to following trainees' progression on each competence during the whole training period. Based on the competences, training is adapted to individual needs to a certain extent (see R4). But clear reports cannot be directly generated from the database yet, ITr-results are available in another way, and selection and training results are stored separately (not combined in one database yet). The web-based assessment tool shows some technical problems. Therefore, not all assessments are directly stored in the database. Some are put in the system after the training. We are improving the data storage and we are designing a new tool since a year now.</p>	
R7 Clarity	4
<p>The performance standards resulted in more clarity for trainees and coaches what is expected in a phase (see R4), although a certain degree of disagreement and differences in interpretations of the standards still exist. Clarity of performance criteria and standards is rather high due to formulation by controllers (see R3). Assessors were trained in how to interpret the criteria and how to write a clear report. They were getting to the criteria that stay the same during training. But it remains difficult to point at the specific causes for insufficient performance because of the cognitive character of the ATC task.</p>	

Requirements	Score
R8 Consistency	4
Within the TWR/APP or ACC trajectory the assessment system is very consistent, but there remain differences across the Unit Training as a whole (e.g., assessor/coaching system, procedures, 2- or 6-points rating scales). Assessment can be improved by taking over strengths from each other. Basic principles (competences, phasing, performance criteria, standards) are consistent across the various courses and training phases, not only within Unit Training but also in Initial Training.	
R9 Level of performance (standards)	4
Behavioural examples serve as performance standards, made by controllers. This has lead to clearness what has to be learned in each phase (see R4, R7), although there still exists differences in the interpretation of these standards. For the assessors it remains difficult to assess the amount of progression needed to achieve these standards. Due to individual differences in learning, various (allowed) learning curves exist. Slow starters, learning plateaus in-between and other variations occur as proven by analyses of learning curves of the trainees. Causes of insufficient performance are not always clear because of invisible, underlying cognitive processes. Learning curves of more trainees should be analysed to get evidence of prediction of progression, (variations in) learning curves and differences in trainability of competences. More insight into complex skill acquisition is needed too.	
R10 Objectivity	4
Complete objectivity in performance assessment is hardly possible, especially in complex cognitive tasks such as ATC. Nevertheless, major efforts were made to maximize objectivity: clear performance criteria en standards, strict procedures, multiple assessors, and profound argumentation of the assessors' 'gut feeling' in feedback and in pass-fail decisions. Reliability analyses (see C2) point at high interrater agreement but also at the occurrence of some rating errors (e.g., halo, range restriction). Assessors are strongly influenced by previous experiences, meanings of colleagues and other factors, which is inevitable in training with a small group of assessors. Analyses indicate that the degree of objectivity is higher in performance tests (simulator) than in progression measurements.	
R11 Continuous assessment	5
Continuous assessment has been maintained: trainees' performance is measured over a longer period to get a complete picture about progression on each competence in various task situations. Additionally, performance tests are executed for more objectivity (see R10): well-balanced test scenario, multiple assessors, clear observation checklists, and detailed scoring techniques.	
R12 External influences to be eliminated	3
Personal preferences in work styles, strategies and coaching are not eliminated yet. They often cause problems for trainees due to the inevitable subjectivity (see R10) and the double role of coach/assessor (see R16). Attention is paid to possible external factors (e.g., private circumstances, working environments) that may influence trainees' performance during training. Then retraining, extra counselling or other measures are undertaken. In addition, special task situations in OJT (e.g., extreme weather circumstances), which also relate to external influences, cannot be eliminated in the real working environment. . Assessors should take into account these circumstances in their judgments.	

Requirements	Score
R13 Culture	3
Assimilation to the culture of the operational departments is not directly assessed, but competences such as teamwork ability and attitude are rated. By maximizing objectivity (see R10) possible influences of culture are ignored as much as possible, but assessors still are part of this culture.	
R14 Training redesign	5
Training objectives and performance criteria are directly related, derived from the competence analysis. Competence-based assessment implies that the integrated whole of knowledge, skills and attitudes is assessed; therefore, small changes in training design do not have much effect on the assessment design. The assessment system is flexible enough to change parts of it if necessary.	
R15 Optimal construction	5
The assessment system in various courses is designed from an overall (scientific/ practical) view on assessment, based on competences, performance criteria and standards, continuous assessment and performance tests, assessor role and procedures.	
R16 Role of coach/assessor	3
The role of coach vs. assessor is defined for progression measurements (coaching and assessing) vs. performance tests (assessing), but separating the role of coach and assessor remains difficult and needs improvement. The focus is on assessment often too much. Trainees see their coach as an assessor and feel being assessed continuously. New coaches are trained in assessment techniques, and existing coaches have followed a refresher training, aimed at the introduction of the redesigned assessment system. Training also contains the prevention of rating errors in order to enhance objectivity (see R10). Not all coaches (TWR/APP) are as intensively trained in assessment (e.g., report writing).	
R17 Integration of the selection and training system	4
The selection system DATCOSS is redesigned simultaneously with the training system, as a combined system, to increase training output. The competences serve as criteria in selection and are assessed in work samples. Selection and training results are stored in a database for validation research, but they are not combined yet. Continuous feedback from UT/ITr to selection (reasons for failing, quality of 'aces') and vice versa (coaching/ training advices) is provided by quantitative analyses and qualitative information exchanges. But this can be improved. Validation research of selection has not been started yet, due to a too low number of trainees as outcome from selection; this will start very soon.	
R18 Contribution to personnel shortage	3
The existing shortage of controllers is not solved yet for several reasons (see R1, R2). It has not been proven that the assessment system has contributed to deliver a sufficient number of qualified controllers. More efficiency in terms of shorter training or less effort is not achieved at all, but this may be influenced by increasing task demands during the last years. In contrast, lengthening training and extra costs are invested (remedial teaching, extra coaching, counselling) for saving trainees who may have a chance to succeed, because false negatives should definitely be avoided due to the shortage on operational controllers, if necessary at the cost of false positives.	

Table 13.2. *Evaluation of the constraints of the PoR*

Constraints	Score
C1 Consequences of redesign	4
The implementation of the assessment system has had much impact on the users. Unnecessary changes were avoided as far as possible. The implementation was done very gradually. Design, implementation and evaluation activities were alternating. A profound introduction was held before the system was implemented to the whole user group. Controllers participated in these introductions and in the other design activities; thus, intensive communication with them was realized. Because the changes for TWR/APP coaches were larger, they faced more problems in the implementation of the system than ACC (e.g., problems with acceptance, misunderstanding of the procedures).	
C2 Reliability	4
Various reliability analyses were done (see chapter 9). Interrater agreement across assessors was sufficient for practical settings, but the involvement of multiple assessors is recommended. Results of internal consistency indicated that some criteria that belong to a competence should be adjusted. Some rating errors exist (e.g., halo error, range restriction) but in general they did not have affected the assessments substantially. Reliability of performance tests was higher than of progression measurements due to its objective character. Test reliability of performance tests was very high.	
C3 Validity	3
Predictive validity cannot be examined adequately yet due to the low number of trainees as output from training. However, the pass rate increases (see R1, R2). Pass or fail is correctly predicted by the competence ratings with the most critical and less learnable competences as main reasons for failing (see chapter 11). Progression is also predictive for pass-fail. Analyses of learning curves of more trainees are needed to get evidence of predictability of learning processes, time influences and trainability of specific competences. History of previous trainees (quantitative/ qualitative) has been used to take more profound and valid pass/fail decisions since three years.	
C4 Practical usability	4
The tools for trainees and assessors are practically usable in general (see chapter 12): assessment reports, performance criteria and standards, manuals, checklists. The web-based assessment tool is easy to use but some technical problems exist. Another tool is being designed (see chapter 7).	
C5 Transparency	3
Some procedures for OJT TWR/APP (not for ACC) are rather complicated and therefore not always applied correctly, such as the frequencies of filling in specific reports and rules for transitions to next phases. The tools used in assessments (see C4) are very clear and transparent for trainees and assessors.	
C6 Acceptance	4
The assessment system is generally accepted by the users (coaches and trainees), because they understand the purposes and advantages of the new system. However, some complex procedures for TWR/APP (see C5) are not fully accepted yet, and technical problems with the web-based assessment tool result in some problems with the acceptance (see C4).	
C7 Confidentiality	4
The authorizations in the web-based assessment tool make it possible to keep personal information completely confidential. There exist organizational regulations for confidentiality that are obeyed. However, the possibilities for confidentiality could be applied in a better way. This will be done in the newly designed web-based assessment tool.	

Constraints	Score
C8 Input of controllers	5
The assessment system was designed in intensive collaboration with controllers as participants of the design teams (coaches, training managers). The products (criteria, standards) were partly made by the controllers themselves who served as subject matter experts.	
C9 Benchmarking	5
Profound literature research was done and existing ATC assessment systems were examined to consider the possibilities and constraints of designing an assessment system, based on modern design methodology, general learning theory, and studies on assessment design.	
C10 Continuous evaluation	5
The assessment system has been evaluated continuously up-to-now by means of user evaluations. Quantitative data analyses were also done extensively. This evaluation should be a continuous process with a feedback loop towards ITr and selection and vice versa, aided by the web-based assessment tool.	
C11 Quality assurance	5
Evaluations were done by training experts in collaboration with controllers (training managers) who should guarantee quality of the assessment system. This will be kept in the future.	

In sum, there are no requirements or constraints that were rated as 'not' or 'hardly achieved' (respectively value 1 or 2). Seven of them were rated as 'partly achieved' (value 3), fourteen as 'mainly achieved' (value 4), and eight as 'completely achieved' (value 5). We conclude that the results are quite positive.

13.3 Discussion

In chapter 6 we classified the various requirements and constraints into categories. We use the same classification to discuss the evaluation against the PoR below. First, requirements R1, R2 and R18 refer to a possible increase of the output of qualified (competent) controllers from training. These requirements were all rated as 'partly achieved': the existing shortage on controllers is not solved yet, and the number of trainees who followed the whole training trajectory is still too low for analysing predictive validity adequately. The pass rate is increasing, especially since 2006, but we should wait for more convincing evidence of a higher number of trainees. In addition, other factors play a role such as recruitment and selection. Another influencing factor is the complexity of the ATC task; this has not been reduced yet since the start of the project.

Second, requirements R4, R5, R6, R9, R14, and R16 relate to the main purposes of assessment: to support learning processes and to make reliable and valid pass-fail decisions. The majority of these requirements were rated as 'largely achieved' for several reasons. The performance standards have resulted in more clarity for trainees and coaches on what is expected in intermediate phases and at the end of training. Measurement of progression over time serves as an indicator of learning. Learning curves can be derived from the assessment results. They are stored in a database filled with the help of the web-based assessment tool. This

makes it possible to follow trainees' progress intensively. Individual differences in learning (e.g., slow starters, learning plateaus) can be recognized, and individualized interventions can be undertaken if there are problems in the learning process. In this context, R3 is a distinctive requirement: the assessment system is competence-based. The same set of competences is assessed over time so that trainees' progression on each competence can be followed. Trainees' strengths and weaknesses can be identified, and this allows to provide focused feedback and to adapt training to trainees' needs. Furthermore, the competence ratings are used to make profound pass-fail decisions. The history of previous trainees helps to provide arguments for these decisions. Requirement R17 states that selection and training should be a combined system: they are based on the same set of competences. However, more research on how to adapt training to individuals, on learning curves and on trainability of competences is needed before these requirements will be completely met.

Third, requirements R7, R8 and R15 state that more consistency and clarity was needed in comparison to the previous assessment system. The assessment system is consistently designed within Unit Training, for instance, by making use of the same set of competences and by defining performance standards in a comparable way. The system is rather transparent for trainees and assessors.

Fourth, the psychometric requirements and constraints are included in R10, R12, C2, C3 and C10. These were analysed in more detail in chapter 9 and rated as 'partly achieved' or 'largely achieved', depending on the specific requirement or constraint. We may conclude that the assessment system is sufficiently reliable in terms of interrater agreement, the occurrences of rating errors and test reliability, but that criterion-related (predictive) validity has not been proven yet (see chapters 10 and 11).

Finally, many constraints refer to the practical usability and the implementation of the assessment system (R13, C1, C4, C5, C6, C7, C8). They are generally rated as 'largely achieved'. The users evaluated the system as sufficiently usable (see also chapter 12). The acceptability of the system is high despite of some technical problems with the web-based assessment tool. But this tool is going to be replaced by another system soon. The intensive collaboration with the users and a careful introduction prevented problems with the implementation of the system. Another reason is supported by R11: continuous assessment was kept in order to avoid redundant changes of the existing assessment system.

13.4 Conclusions

In sum, the results of the evaluation against the PoR with the main stakeholders indicate that the redesigned assessment system supports learning processes adequately, that it is reliable and practically usable, and that it has been designed consistently. More research on predictive validity is needed to prove that the output from training has increased and that pass-fail decisions have been improved.

14 General conclusions

The objective of the study was to design an assessment system that optimally supports the acquisition of air traffic controller competences in simulator and on-the-job training. In the previous chapters we explained how we designed the competence-based assessment system for ATC training at LVNL based on general design methodology, we described the main properties of the assessment system, and we discussed how we evaluated the new assessment system, addressing psychometric quality, learning processes, and user evaluation. In this chapter, we draw conclusions concerning the objective of the study.

14.1 Assessment systems for complex skill acquisition

Air traffic control is considered to be a complex cognitive skill (Schneider, 1990; Van Merriënboer, 1997) like other process control in transportation (e.g., aviation, shipping, railways) and in process industries (e.g., chemical, nuclear plants). Human operators are working in an environment in which any mistake may cause an unsafe situation. Therefore, the requirements regarding operators' competences are very high. This may result in a shortage of competent personnel since learnability appears to be limited. The pass rates in training often remain too low despite of high selection standards. The training of complex skills can be very time-consuming and expensive, especially when high-fidelity simulators are used. However, this is necessary to train trainees to the highest possible level before they enter the safety-critical working environment in the on-the-job training (OJT).

Under the assumption that task complexity will not diminish in the near future, this study proposed a new approach to this problem: the design of a reliable and valid assessment system for the training. A well-designed assessment system can contribute to increase the output of competent operators in several ways. First, assessment can support learning processes through feedback and adaptation of training to individuals' needs. Second, better founded and valid pass-fail decisions can be produced by restricting the number of false positives, i.e. prospectiveless trainees who unnecessarily occupy expensive training positions during later phases, and false negatives, i.e. failed trainees who could have had a chance to succeed training ultimately. Third, predictive validity of selection can be improved through the use of more reliable training criteria in validation studies.

Within the domain of aviation, this study has contributed to a scarcely investigated field of research. Most scientific research has concentrated on assessment in selection systems (job samples) in ATC (e.g., Ramos, Heil & Manning, 2001; Eissfeldt, Heil & Broach, 2002), and on performance assessment of aircrew (e.g., Flin & Martin, 2001; O'Conner et al., 2002). Research on the design of assessment systems in ATC training is virtually non-existent. There are only some publications that describe the practical use of assessment systems (e.g., Hopkins, 1995; EATCHIP, 1999).

This study has led to some general recommendations for designing an appropriate assessment system for complex skill acquisition in simulator and on-the-job training. The first step is to get a deep insight into the tasks that are performed by the operators and into the competences that are required for the job. Operators have become competent by learning and by experiences at the workplace, and it is not easy for them to explicate what they are doing in their work (Bereiter, 2002). They use many different strategies that are person-dependent and adaptive to task situations. Complex cognitive processes are usually involved, which have been automated and occur unconsciously. It explains why assessors recognize behaviours in trainees on the basis of their 'gut feeling', while they have problems with arguing why trainees perform above or below standards. Learning implies a development from controlled towards automated information processing. If the final product of learning is clear, it will be possible to define the steps towards this goal. This seems to be straightforward, but many difficulties are often experienced in this process due to the task complexity.

In this study, we proposed to start with a competence analysis to define the competences that have to be acquired at the end of the training. In our view, competence refers to the ability to effectively apply acquired knowledge, skills and attitudes while performing tasks in realistic settings. Published (cognitive) task analyses can be used for the competence analysis. Additionally, we recommend involving the operators as subject matter experts in the process. Their implicit knowledge has to be made explicit, and they should achieve a common understanding of the competences. The competences have to be defined in their own jargon to be able to use them in practice. Defining sets of performance criteria in observable behaviours (cf. 'behavioural markers'; Flin & Martin, 2001) that represent the competences helps to recognize them better. For the purpose of assessment design, we developed the so-called ATC Performance Model (Oprins, Burggraaff & Van Weerdenburg, 2006a). The model is based on general performance models, and therefore its philosophy can be applied for other complex skills as well.

The competences can serve as a starting point for designing a training and assessment system. It should be noticed that training and assessment are inherently connected to each other. Training objectives usually express what has to be learned within a certain training period, and assessment should ascertain to which extent these objectives have been achieved (Mager, 1962). We could say that training in simulators and OJT is always competence-based, because they are realistic learning environments and knowledge, skills and attitudes are merged by learning-by-doing. In order to support learning processes, it is important to divide training into successive phases that are clearly defined. Trainees should not only know what they would

have become at the end of training, but also how to get there. They must learn step-by-step. For simulator training this implies that a gradual sequence of task complexity in the exercises is needed. However, phases in OJT are not commonly applied yet (Jacobs & Jones, 1995). Process control tasks often have many dynamically changing variables, which cannot be planned in advance. This makes it difficult to structure OJT, but we consider this necessary for realizing a gradual learning curve. The phases can be flexible in length if we want to take individual differences in learning into account.

The next step is to develop the assessment system as part of the training system. Assessment should support learning through feedback on the development of the trainee's learning process. If training is divided into phases, it should be known what is expected from trainees at the end of each phase. This helps trainees to get insight into what they have to learn in a certain training period. Agreement between assessors is required to make assessments reliable. Assessment should identify the trainee's strengths and weaknesses so that training can be adapted to the trainee's needs (e.g., by remedial teaching, adaptive task selection, or specific coaching). Complex skills applied in safety-critical environments ask for an additional purpose of assessment. Future learning should be predicted as far as possible in order to avoid useless costs and incompetent personnel at the workplace. Pass-fail decisions must be predictive valid.

There exist many possibilities of designing an assessment system that fulfils these purposes. We chose for a competence-based assessment system in which the competences are assessed that were derived from the ATC Performance Model. The complex and dynamic character of ATC makes it impossible and useless to assess smaller pieces of knowledge or skills, because they are merged in simulator training and OJT. The same competences, represented by a set of performance criteria, can be assessed from the start till the end of the training due to their generic character. Because trainees do not perform at final competence level directly, we defined performance standards for each training phase against which the competences are assessed. These standards indicate the requirements on the assessment tasks to be executed, such as task complexity, in combination with specific norms, such as the number of mistakes allowed. We expressed them in exemplary behaviours, but this can be done in many ways. We refer to the various variations on behavioural anchored rating scales (BARS; Berk, 1986). Most important is that the assessors achieve consensus on required performance in each phase.

Such a competence-based assessment system has some general advantages, and can therefore be applied in related domains as well. First, assessing the same competences over time makes it possible to follow the trainee's progression intensively, based on various task situations and circumstances. Potential deficiencies of trainees, founded in competences, can be detected in an early phase. Assessors get familiarized with their meaning easily. Second, the emotional and social aspects are assessed in addition to the commonly applied technical and cognitive aspects, since they all belong to competence (Bereiter, 2002). We consider assessment of cognitive processes important in complex skill acquisition, because it provides diagnostic information on shortcomings in performance. This is needed for determining appropriate training interventions and for predicting future performance (Regian & Schneider,

1990). In this respect, trainability of competences should be taken into account (cf. consistent vs. non-consistent task components; Schneider, 1990). Decisions to stop training should be based on less trainable competences. Third, the competences can be used as criteria in validation studies to improve selection systems (Roe, 2005). The same competences can be assessed in job samples.

The design of a web-based assessment tool, in which assessments are filled in digitally and stored in a database, contributes to the aforementioned advantages. Authorized persons can generate several reports and overviews from different places, and training results can be used in reliability and validity studies. We proposed a new method to derive learning curves from assessment results. We explained how we made so-called 'recalibrated' learning curves, based on assessment against increasing performance standards that are constantly translated into the same rating scale. In this way, this study adds to other studies that examine learning curves in complex skills such as ATC (e.g., Kanfer & Ackerman, 1988; Lee & Anderson, 2001). Making learning curves helps to follow the trainee's learning processes better. Patterns and individual differences in learning, such as slow starters and learning plateaus, can be recognized. The development of specific competences can be followed. This can provide information about trainability of competences. Ultimately, cut-offs for pass-fail decisions can be made if there is sufficient evidence of predictive validity. However, one should be careful with relying on only quantitative measures in complex skills because of the underlying cognitive processes.

It could be beneficial to choose for automated measurement. Simulators and operational systems can log certain data, and this makes assessments more reliable. However, these data cannot capture the aforementioned diagnostic information. Therefore, we recommend that automated measurement should always be combined with subjective measures such as over-the-shoulder observations (Vreuls & Obermayer, 1985; Sanders, 1991). This relates to the distinction between performance tests and continuous assessment (Hopkins, 1995). We prefer a combination of both methods. Performance tests in simulators are more reliable because of well-balanced test scenarios and the absence of interference by a coach, while continuous assessment provides extra information about learning processes and is based on a variety of task situations. Coaches are continuously in interaction with trainees and can force them to verbalize their thoughts (Farmer et al., 1999). In this way, the trainees' learning potential is assessed. This refers to 'dynamic assessment' (Campione & Brown, 1990).

Finally, designers of assessment systems should pay attention to the role and quality of assessors. The involvement of multiple assessors enhances the reliability of assessments. Assessors should be trained in the practical use of the system, the meaning of criteria and standards (cf. frame-of-reference (FOR) training and calibration sessions), and the avoidance of rating errors (Murphy & Cleveland, 1995). It can be very critical to get a new system accepted by the assessors and other users such as trainees and managers. Therefore, the tools should be practically usable. Involvement of the user group in the design process with intensive collaboration and a careful introduction of the system (e.g., by starting with pilots and organising presentations) are needed to enhance its implementation.

14.2 The LVNL assessment system

The main purpose of the competence-based assessment system designed for LVNL was to make training more effective and efficient by supporting learning processes better and by taking more reliable and valid pass-fail decisions. This should contribute to a higher output from training. We conclude that we have achieved this main purpose to a certain extent, based on the evaluation of the assessment system (see chapters 9 – 12). This section summarizes the main findings and implications for LVNL.

Learning processes

The assessment system has resulted in a better support of learning processes through feedback. Assessors can identify the trainee's strengths and weaknesses more easily due to the focus on competences. In the competence workshops the 'gut feeling' of controllers was explicated. The jargon of controllers has been maintained in the performance criteria and standards. This has enhanced the recognizability of behaviours. Assessors can better point at potential causes for insufficient performance that are often founded in cognitive processes. The performance standards defined for successive training phases have made clear what is expected from trainees in each phase. This has been a large improvement especially in OJT, which was not structured before. Since the same competences are assessed over time, the assessors have been familiarized with their meaning. The web-based assessment tool helps to follow the trainee's learning processes. Analyses of learning curves, derived from assessment results, have shown that the assessment system is able to represent patterns and individual differences in learning (learning plateaus, slow starters) in conformance with general learning theory (Newell & Rosenbloom, 1981). The assessment system also differentiates in competence development between trainees adequately.

A practical implication of these findings is that training can be adapted to the trainee's needs. A choice could be made from various coaching techniques, flexible lengths of training phases, adaptive task selection, retraining, extra training, alternative training trajectories, remedial teaching, and counselling. Unfortunately, adaptive training is not applied as much as needed yet. Practical problems play a role such as the limited availability of coaches and simulators, operational and training schedules, and organizational regulations. Furthermore, it seems that while coaches and assessors are better able to detect deficiencies, they do not know as well how to improve specific competences. We could say that the focus is still on assessment as such than on its use as a tool for more effective learning processes. We are searching for the benefits of various training interventions in relation to the competences, for instance, by organizing tryouts with remedial teaching. Additionally, we are providing the coaches with techniques of coaching on specific competences in train-the-trainer sessions. Finally, we are trying to give trainees more responsibilities in guiding their own learning process, for instance, by stimulating self-reflection in self-assessments. In general, more research is needed to maximize the adaptivity of training.

Pass-fail decisions

Since the new assessment system was implemented, pass-fail decisions have been based on an extensive quantitative and qualitative analysis of the trainee's assessment results. We have been using the history of previous trainees to help predicting future performance better. However, we have not been able to analyse the predictive validity of pass-fail decisions, because the number of trainees who completed the whole training period is yet too low. False negatives cannot be verified because the trainees have left already, which produces a restriction-of-range problem. The tendency to avoid false negatives because of the shortage of controllers has resulted in a too high number of false positives in later training phases (OJT).

Nevertheless, we investigated some issues that refer to criterion-related validity. Learning curves in earlier phases seem to be comparable with learning curves in later phases, although this is based on a very low number of cases. Pass-fail decisions were correctly predicted by the mean competence ratings and progression. Progression relates to the trainability of competences. Various quantitative and qualitative analyses in this study pointed at the following main reasons for failing, respectively: *mental picture*, *workload management*, *attention management*, and *decisiveness*. In accordance with findings in the literature, they appear to be more critical and less trainable than others because they are founded in underlying, innate abilities and in personality of individuals. For instance, *mental picture* (cf. situational awareness; Endsley, 1995) is supposed to be strongly dependent on memory capacities, spatial orientation and speed of information processing (Isaac & Ruitenberg, 1999; Garland, Stein & Muller, 1999). *Workload management* requires personal properties such as stress resistance, vigilance, dominance and action orientation (Hettema, Leidelmeijer & Geenen, 2000; Kuhl, 1994).

The predictive value of learning curves and competence development should be analysed further to improve pass-fail decisions in the future. Ultimately, cut-offs can be made based on a (weighted) sum of competence ratings and progression. However, strict cut-offs are not likely to be applied soon. The assessments made by human assessors will never be completely reliable. Qualitative information (e.g., observations and comments) will remain indispensable for the argumentation behind the assessors' judgements. The presence of learning plateaus is also a problem. They can be final (i.e., asymptote of learning) or only intermediate, needed for assimilating new skills and knowledge with previous ones (Schneider, 1990). Furthermore, the large amount of individual differences in learning makes it difficult to predict future learning and to adapt learning to individual needs.

Reliability

Major efforts were made to maximize reliability such as defining clear performance criteria and standards, involving multiple assessors, and organizing assessor training. We investigated the reliability of the assessment system in various ways. We examined interrater agreement across assessors in terms of level, dispersion and shape indices (Church, 1997; Youngstrom, Loeber & Stouthamer-Loeber, 2000). The results have shown that the assessors' ratings are interchangeable for the overall

performance level, but not for singular competences. Apparently, the assessors have difficulties with pointing at the same deficiencies of trainees. The assessors make some rating errors (Saal, Downey & Lahey, 1980; Murphy & Cleveland, 1990). Assessors show range restriction but not central tendency. They do not differ in their distribution of ratings, nor in severity or leniency. Assessors tend to give ratings higher than the scale midpoint, but this is not caused by severity or leniency errors; it is the result of assessment against standards. The halo-error is only present to a certain extent: some competence ratings should be intercorrelated because they are conceptually related. We recognized the ATC Performance Model in the results. Analyses of internal consistency confirmed that the classification into competences and performance criteria has been well-chosen. The reliability of simulator tests is higher than for progression measurements in general as expected.

We conclude that the role of the assessor should be improved to get more reliable assessments. The involvement of multiple assessors has to be maintained. We have been organizing assessor training to pay attention to the avoidance of rating errors, a common understanding of competences and standards, and a better separation of the coach and assessor role. Furthermore, we could say that the technical possibilities of the simulators are not completely exploited. Automated measurement could be applied in performance tests, although this should always be complemented by human judgments as argued before.

Output of competent controllers

The aforementioned measures should contribute to an increased output of competent controllers from training in order to solve the shortage of controllers. Unfortunately, we do not have convincing evidence yet, although the pass rate has been increasing since 2006. We should take into consideration that many factors are involved, not only training and assessment.

First, we designed the Dutch Air Traffic Controller Selection System (DAT-COSS; Oprins et al., 2006, 2007) simultaneously, which was implemented in 2003. Long-term validation research is needed to adjust selection and training requirements to each other on a continuous basis, and to maximize the pass rate in training ultimately. Because detailed training results are stored in the web-based assessment tool, validation research can be based on more criteria than only the pass-fail criterion. A feedback loop is not only required from training to selection but also vice versa. Possible deficiencies of trainees that are detected in the stage of selection can be the base for adapting training to individual needs. This would enhance the chance of success in training. In this way, selection and training can be considered as one integrated system, which is also dependent on recruitment. The influx of candidates in selection has been too low in the last three years; thus, it was not possible to make a top selection of excellent candidates entering training. This may have influenced the output from training as well.

Another influencing factor is the design of the ATM system. The selection and training system is not a static system but should be adapted to changes in the operational working environment. Vice versa, evidence from selection and training can guide some changes in the ATM system. From this perspective, the requirements on

the controller's competences should not become higher, but the task complexity should be reduced. There is much attention for this subject now, for instance, in European projects such as SESAR (Single European Sky ATM research) and FAB EC (Functional Airspace Blocks in Europe Central). One of the reasons for concern is that the amount of air traffic keeps growing while this cannot be managed anymore with the current ATM systems and human resources. A reduction of the complexity of the ATM system might be achieved by, for instance, the organization of work (e.g., splitting up sectors), standardization in working methods and procedures (e.g., fixed routes), and more automation by means of advanced technology (e.g., decision support tools). Insight in learning processes and trainability of competences might help to propose a certain direction of technological developments, for instance, by defining aspects that can be automated easily and which aspects cannot. One should take into account the limitations of the cognitive abilities of human beings. While automation can reduce the task complexity, automation should not result in a loss of attention due to a change towards more routine tasks. A possible next step in this study can be to analyse the relationship between the competences and the design of ATM systems, based on experienced difficulties in the acquisition of competences in selection and training.

14.3 The design process

In our study, we combined two design methodologies that seem to be most suitable for the purpose of assessment design in this study: the design cycle of Roozenburg and Eekels (1991) and the training design methodology of Moonen (2000). This resulted in three main phases: (1) *analysis*, (2) *design and implementation*, (3) *evaluation*. The analysis (phase 1) comprised three parts: a literature review, an ATC competence analysis, and an analysis of the practical problem at LVNL by constructing a Program of Requirements. This PoR has guided the actual design and development of the assessment system, alternated by its gradual implementation (phase 2). Finally, we evaluated the quality of the assessment system against the PoR in the phase of evaluation (phase 3). We evaluated its psychometric quality, its representativeness of learning processes, and its practical usability. The phases were not subdivided because the processes within each phase were very iterative, even across phases.

The design process followed in this study has shown that the segmentation of the design process in phases is practically usable. However, we have experienced that it is very difficult to pre-structure this process, especially when the product should be used in practice. While scientific methods are usually formalized and quantitative, methods in practical research often tend to be informal and qualitative (e.g., informal conversations, observations, document analyses). The use of these methods is very valuable for improving the system and getting it implemented, but it cannot be planned, structured and summarized easily. The construction of a PoR helps to guide the process by defining what has to be designed and by evaluating to which extent the designed product fulfils the original requirements. Finally, intensive

collaboration of scientists with the user group (air traffic controllers) has been a key success factor in designing a really innovative product that has been completely accepted for practical use.

14.4 Implications for further research

The research described in this study, conducted at LVNL, has resulted in a valuable product. However, we should notice that it has its limitations. First, the fact that design methodology was not maintained strictly may have influenced the scientific quality of the assessment system. Second, the data used in the various analyses were collected in 'real' training settings. This has led to some incompleteness and unreliability of data sets as we found back in the results. Third, only a small sample of subjects in one domain, and even in only one course (ACC), was involved. These limitations imply that some issues need further research.

In the first place, more research on the design of assessment systems for complex skill acquisition in simulator and on-the-job training should be done. We argued that this is one of the first scientific studies on this subject in ATC. Several properties of our assessment system might be applicable in related domains. For instance, we proposed a specific method of assessing the same competences against increasing performance standards formulated in exemplary behaviours. It should be investigated whether this method is also valuable for the acquisition of process control in other safety-critical environments (e.g., nuclear plants, railways or shipping). The largest benefits are expected in OJT, as we mentioned before. Special attention should be paid on the improvement of the reliability and validity of performance assessment systems. We argued that the cognitive processes involved in complex skills make this difficult to attain, since human assessors are dependent on observable behaviours. It remains a challenge to maximize the quality of the assessors' ratings, for instance, by providing an effective assessor training.

In addition, more insight in complex skill acquisition is needed in order to achieve that assessment can support learning optimally. Therefore, the assessment results should represent learning processes as expected from learning theory. We explained how we derived learning curves from these results, using a new method. However, a uniform theory of complex skill acquisition does not exist yet (e.g., Anderson & Lee, 2001; Taatgen & Lee, 2003). Analysing learning curves can be used for modelling learning processes as being done in ATC tasks (e.g., Kanfer & Ackerman, 1988). These studies should be continued with the emphasis on the acquisition of complex skills in realistic settings, because this seems to be different from (simple) skill acquisition in laboratory situations. Our method of making learning curves might contribute to this field of research if it is further developed. Transfer to other domains should be possible, because recalibrated scales are applied in many educational and training systems.

Furthermore, trainability of competences in ATC and in related domains requires more research. Its relation with innate cognitive abilities and personality characteristics should be investigated further in order to improve the predictive validity of

selection and training. Effective training methods should be developed for the acquisition of specific competences that are trainable to a certain extent. For instance, *situational awareness* is assumed to improve by applying attention management strategies (Shebilske, Goette & Garland, 2000) or chunking methods (Garland, Stein & Muller, 1999). Another critical competence is supposed to be *workload management*, which relates to strategies. The acquisition of strategies is a challenge that is often mentioned as a key success factor in ATC training (Nunes & Mogford, 1993; Schneider, 1990). Strategies are very important in process control tasks: operators constantly adapt their strategies to changing task demands and to their personal states to keep in control (Sperandio, 1978; Hockey, 1997). However, it is not well known which training methods (cf. task selection, coaching techniques, learning activities) can facilitate strategy learning in an optimal way.

Moreover, in this study we argued that training would be maximally effective and efficient if it were completely adaptive to the trainee's needs. In this way, assessment would really support learning processes. Various studies have been done to adaptive training such as aptitude-treatment interaction (Snow, 1989) and dynamic task selection in ATC (Salden, 2005; Salden et al., 2004, 2006). Self-directed learning skills and self-assessment methods could be withdrawn in the choice of the learning tasks, and an optimal balance is needed between part-task and whole-task training (Van Merriënboer & Kirschner, 2007). However, we explained that adaptive training is difficult to realize in practice due to the limited availability of coaches and simulators. The existence of many individual differences in learning, such as learning styles and other personal features, adds to this problem. Studies like those of Salden et al. (2004, 2006) are executed with simplified simulated ATC tasks with novices as trainees, while in practice, especially in OJT, more variables are involved that are difficult to control. More research is needed on the relationship between adaptive training and deficiencies in competences with the focus on task selection in ATC simulator training. Various individual differences in learning as well as self-directed learning skills should be taken into account.

Finally, more advanced technology can help to improve assessment systems in training in various ways. Automated measurement may lead to more reliable assessments by logging and scoring the aspects that can be measured objectively (Sanders, 1991; Vreuls & Obermayer, 1985). As argued, this should be complemented by subjective judgments of human assessors. Additionally, more sophisticated assessment tools should be developed to fill in assessments, to store the results in a database and to generate various reports and overviews. We experienced that these tools cannot be bought yet, but existing systems should be adjusted to our needs.

In conclusion, this thesis has gained more insight into the design of reliable and valid assessment systems for complex skill acquisition in simulator and on-the-job training with the focus on competence-based assessment. Future research should not only focus on assessment design but also on its relationship with learning processes. Assessment contributes to improve the effectiveness and efficiency of training. Finally, this may help to solve the problem of a shortage of competent personnel.

References

- Ackerman, P.L. (1988). Determinants of individual differences during skill acquisition: cognitive abilities and information processing. *Journal of experimental psychology: general*, 117, 288-318.
- Ackerman, P.L. (1989). Individual differences and skill acquisition. In P.L. Ackerman, R.J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences: advances in theory and research* (pp. 165-217). New York: Freeman and company.
- Ackerman, P.L. (1990). A correlational analysis of skill specificity: learning, abilities and individual differences. *Journal of experimental psychology: learning, memory and cognition*, 16, 883-901.
- Adrales, G.L., Park, A.E., Chu, U.B., Witzke, D.B., Donnelly, M.B., Hoskins, J.D., Mastrangelo, M.J., & Gandsas, A. (2003). A valid method of laparoscopic simulation training and competence assessment. *Journal of surgical research*, 144, 156-162.
- Aiken, L.R. (1996) *Rating scales and checklists. Evaluating behavior, personality and attitudes*. New York: John Wiley & Sons.
- Anderson, J.R. (1982). Acquisition of cognitive skill. *Psychological review*, 89, 369-406.
- Anderson, J.R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard university press.
- Anderson, J.R. (1995). *Cognitive psychology and its implications* (4th ed.). New York: Freeman and company.
- Antersijn, P.A. & Verhoef, M.C. (1995). Assessment of non-technical skills: is it possible? In N. McDonald, N. Johnston, & R. Fuller (Eds.), *Applications of psychology to the aviation system* (pp. 243-250). Hants, England: Avebury.
- Atkinson, R.C., & Shiffrin, R.M. (1968). Human memory: a proposed system and its control processes. In Spence, K., & Spence, J. (Eds.). *The psychology of learning and motivation* (Vol. 2, pp. 742-775). New York: Academic Press.
- Avermaete, J. van, & Kruijsen, E. (1998). *NOTECHS: The evaluation of non-technical skills of multi-pilot aircrew in relation to the JAR-FCL*. Amsterdam: NLR. (Report No. NLR-CR-98443).
- Averty, P., Collet, C., Dittmar, A., Athènes, S., & Vernet-Maury, E. (2004). Mental workload in air traffic control: an index constructed from field tests. *Aviation, space, and environmental medicine*, 75(4), 333-341.
- Balzer, W.K., & Sulsky, L.M. (1992). Halo and performance appraisal research: a critical examination. *Journal of applied psychology*, 77(6), 975-985.
- Bakel, A. van (1995). *Styles of architectural designing*. Doctoral dissertation, Technical University, Eindhoven.
- Bartram, D., & Roe, R.A. (forthcoming). Individual and organizational factors in competence acquisition. In W.J. Nijhof, P.R.J. Simons, & A.F. Nieuwenhuis (Eds.), *The Learning Potential of the Workplace*. Rotterdam: Sense Publishers.
- Beaubien, J.M, Baker, D.P., & Salvaggio, A.M. (2004). Improving the construct validity of line operational simulation ratings: lessons learned from the assessment center. *The international journal of psychology*, 14(1), 1-18.
- Bereiter, C. (2002). *Education and mind in the knowledge age*. Mahwah, NJ: Erlbaum.
- Berk, R.A. (1986). *Performance assessment: methods and applications*. Baltimore: The Johns Hopkins University.
- Bernardin, H.J., & Buckley, M.R. (1981). A consideration of strategies in rater training. *Academy of management review*, 6, 205-212.

- Biber, L. (1999). *Het wervings-, selectie en opleidingstraject voor luchtverkeersleiders* [The recruitment, selection and training system for air traffic controllers]. Soesterberg: Aeromedisch instituut (Report No. 1999-I2).
- Birenbaum, M. (1996). Assessment 2000: towards a pluralistic approach to assessment. In M. Birenbaum & F. Dochy (Eds.), *Alternatives in assessment of achievements, learning process and prior knowledge* (pp. 3-29). Boston: Kluwer Academic Publishers.
- Black, P., & Wiliam, D. (1998). Inside the black box: raising standards through classroom assessment. *Phi Delta Kappa*, 80(2), 139-148.
- Boekaerts, M., & Simons, P.R.J. (1993). *Leren en instructie: Psychologie van de leerling en het leerproces* [Learning and instruction: Psychology of the learner and the learning process]. Assen: Dekker & v.d. Vegt.
- Bolhuis, S., & Simons, R.J. (1999). *Leren en werken* [Learning and working]. Deventer: Kluwer.
- Boone, J.O. (1976). *Automated assessment of student progress and performance in radar air traffic controller training*. Paper presented at the Annual Meeting of the Military Psychological Association, San Diego, California.
- Borman, W.C. (1986). Behavior-based rating scales. In R.A. Berk (Ed.), *Performance assessment: methods and applications* (pp. 100-120). Baltimore: The Johns Hopkins University.
- Brannick, M.T., Prince, C., & Salas, E. (2002). The reliability of instructors evaluations of crew performance: good news and not so good news. *International journal of aviation psychology*, 12(3), 241-261.
- Brown, J.S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational researcher*, 18(1), 32-42.
- Brown, M.W., & Du Toit, S.H.C. (1991). Models for learning data. In L.M. Collins & J.L. Horn (Eds.), *Best methods for the analysis of change* (pp. 47-69). American Psychological Association: Washington.
- Boston, C. (2002). The concept of formative assessment. *Practical Assessment, Research & Evaluation*, 8(9). Retrieved April 15, 2006, from <http://PAREonline.net/getvn.asp?v=8&n=9>.
- Buckley, E.B., & Rood, R. (1977). *CPM Probe experiment on performance information feedback*. Atlantic City, NJ: Federal Aviation Administration. (NAFEC Technical Report NA-77-18-LR)
- Campione, J.C., & Brown, A.L. (1990). Guided learning and transfer: implications for approaches to assessment. In N. Frederiksen, R. Glaser, A. Lesgold, & M.G. Shafto (1990). *Diagnostic monitoring of skill and knowledge acquisition* (pp. 141-172). Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Card, S.K., Moran, T.P., & Newell, A. (1986). The model of human processor: an engineering model of human performance. In K.R. Boff, L. Kaufman, & J.P. Thomas (Eds.), *Handbook of perception and human performance: volume II, cognitive processes and performance* (pp. 1-35). New York: Wiley-Interscience.
- Cascio, W.F. (1986). Technical and mechanical job performance. In R.A. Berk (Ed.), *Performance assessment: methods and applications* (pp. 361-375). Baltimore: The Johns Hopkins University.
- CAST (1999). *Consequences of future ATM systems for air traffic controller selection and training. WP 3 future ATCO selection and training* (NLR-TR-99315-VOL-3). Amsterdam: National Aerospace Laboratory NLR.
- Chi, M.T.H., Feltovich, P.J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive science*, 5, 121-152.
- Church, A.H. (1997). Managerial self-awareness in high-performing individuals in organizations. *Journal of applied psychology*, 82(2), 281-292.
- Clark, L.A., & Watson, D. (1995). Constructing validity: basic issues in objective scale development. *Psychological assessment*, 7(3), 309-319.
- Cnossen, F., Meijman, T., & Rothengatter, T. (2004). Adaptive strategy changes as a function of task demands: a study of car drivers. *Ergonomics*, 47(2), 218-236.
- Cole, M., & Engeström, Y. (1993). A cultural-historical approach to distributed cognition. In G. Salomon (Ed.), *Distributed cognitions: psychological and educational considerations* (pp. 1-46). Cambridge: University Press.

- Collet, C., Averty, P., Delhomme, G., Dittmar, A., & Vernet-Maury, E. (2003). Subjective aspects of mental workload in air traffic control. In G.R. Hockey, A.W.K. Gaillard, & O. Burov (Eds.). *Operator functional state. The assessment and prediction of human performance degradation in complex tasks* (pp. 291-302). Amsterdam: IOS Press.
- Collins, A., Brown, J.S., & Newman, S.E. (1989). Cognitive apprenticeship: teaching the crafts of reading, writing and mathematics. In L.B. Resnick (Ed.). *Knowing, learning and instruction. Essays in honor of Robert Glaser* (pp. 453-494). Hillsdale, NJ: Erlbaum.
- Corey, D.M., Dunlap, W.P., & Burke, M.J. (1998). Averaging correlations: expected values and bias in combined pearson r s and Fisher's z transformations. *The journal of general psychology*, 125(3), 245-261.
- Cortina, J.M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of applied psychology*, 78(1), 98-104.
- Costa, G. (1993). Evaluation of workload in air traffic controllers. *Ergonomics*, 36(9), 1111-1120.
- Cronin, P. (1997). *Learning and assessment of instruction*. Unpublished manuscript. Edinburgh: University of Edinburgh Centre for Cognitive Science.
- Davison, H.J., & Hansman, R.J. (2003). Use of structure as basis for abstraction in air traffic control. In R. Jensen (Ed.). *Proceedings of the 12th International Symposium on Aviation Psychology*.
- Dijk, J. van, Goede, M. de, Hart, H., & Teunissen, J. (1995). *Onderzoeken en veranderen: methoden van praktijkonderzoek* [Research and change: methods of practical research]. Houten: Stenfert Kroese.
- Drenth, P.J.D., & Sijtsma, K. (1990). *Testtheorie: Inleiding in de theorie van de psychologische test en zijn toepassingen* [Test theory: Introduction in the theory of the psychological test and its applications]. Houten/Antwerpen: Bohn Stafleu Van Loghum.
- Duffy, T., & Jonassen, D.H. (Eds.) (1992). *Constructivism and the technology of instruction: a conversation*. Hillsdale, N.J, London: Erlbaum.
- Dunlap, W.P., Burke, M.J., Smith-Crowe, K. (2003). Accurate tests of statistical significance for r_{wg} and average deviation interrater agreement indexes. *Journal of applied psychology*, 88(2), 356-362.
- EATCHIP (1997). *Model of the cognitive aspects of air traffic control* (HUM.ET1.ST01.1000-REP-02). Brussels: Eurocontrol.
- EATCHIP (1999). *Integrated task and job analysis of air traffic controllers phase 2: task analysis of en-route controllers* (HUM.ET1.ST01.1000-REP-04). Brussels: Eurocontrol.
- EATCHIP (2000). *Air traffic controller training at operational units*. (HUM.ET.ST05.4000-GUI-01). Brussels: Eurocontrol.
- EATCHIP (2002). *Short report on human performance models and taxonomies of human error in ATM* (HERA). (HRS/HSP-002-REP-02). Brussels: Eurocontrol.
- Eissfeldt, H., Heil, M.C., & Broach, D. (2002). *Staffing the ATM system: the selection of air traffic controllers*. Aldershot: Ashgate Publishing Company.
- Eissfeldt, H. (2002). ATC-specific tests. In H. Eissfeldt, M.C. Heil, & D. Broach (Eds.). *Staffing the ATM system: the selection of air traffic controllers* (pp. 85-96). Aldershot: Ashgate Publishing Company.
- Eissfeldt, H. (2002). *Der Einsatz einer computergestützten Arbeitsproben nach dem Lerntestkonzept zur Vorhersage des Erfolges in berufspraktischen Abschnitten der Fluglotsenausbildung* [Using a computer-based work sample following the Lerntest approach to predict the success in practical phases of air traffic controller training]. Doctoral dissertation. Hamburg: Deutsches Zentrum für Luft- und Raumfahrt (DLR).
- Eissfeldt, H. (2003). A work sample test in a Lerntest design – 10 years with the dynamic air traffic control test (DAC). In R. Jensen (Ed.). *Proceedings of the 13th Symposium on Aviation Psychology*. Dayton: Wright State University.
- Endsley, M. (1995). Towards a theory of situational awareness in dynamic systems. *Human factors*, 37(1), 32-64.
- ESARR (2000). *Eurocontrol Safety Regulatory Requirement No 5 - ESARR 5: ATM Services' Personnel*. Brussels: Eurocontrol.
- Eyring, J.D., Steel Johnson, D., & Francis, D.J. (1993). A cross-level units-of-analysis approach to individual differences in skill acquisition. *Journal of applied psychology*, 38, 805-814.

- Farmer, E., Rooij, J. van, Riemersma, J., Jorna, P., & Moraal, J. (1999). *Handbook of simulator-based training*. Aldershot: Ashgate.
- Fitts, P.M. (1965). Factors in complex skill training. In R. Glaser (Ed.), *Training, research and education* (pp. 177-197). New York: Wiley.
- Fitts, P.M., & Posner, M.I. (1967). *Human performance*. Belmont, CA: Brooks Cole.
- Flin, R., & Martin, L. (2001). Behavioral Markers for Crew Resource Management: A Review of Current Practice. *The international journal of aviation psychology*, 11, 95-118.
- Flin, R., Goeters, M.L., Hörmann, H.J., Amalberti, R., Valot, C., & Nijhuis, H. (2003). Development of the NOTECHS (non-technical skills) system for assessing pilot's CRM skills. *Human factors and aerospace safety*, 3, 95-117.
- Flin, R. (2004). The NOTECHS system. In K. M. Goeters (Ed.), *Aviation psychology: Practice and research* (pp. 245-249). Hampshire: Ashgate.
- Gagné, R.M. (1985). *The conditions of learning* (4th ed.). New York: Holt, Rinehart & Winston.
- Garland, D.J., Stein, E.S., & Muller, J.K. (1999). Air traffic controller memory: capabilities, limitations and volatility. In D.J. Garland, J.A. Wise, & V.D. Hopkin (Eds.), *Handbook of aviation human factors* (pp. 455-496). Mahwah, NJ: Erlbaum.
- Gipps, C.V. (1994). *Beyond testing*. London: The Falmer Press.
- Gitomer, D.H. (1993). Performance assessment and educational measurement. In R.E. Bennett & W.C. Ward (Eds.), *Construction versus choice in cognitive measurement: issues in constructed response, performance testing and portfolio assessment* (pp. 241-263). London: Lawrence Erlbaum Associates.
- Gagné, R.M. (1989). Some reflections on learning and individual differences. In P.L. Ackerman, R.J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences: advances in theory and research* (pp. 1-11). New York: Freeman.
- Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: some questions. In R.R. Hoffman & D.S. Palermo (Eds.), *Cognition and the symbolic processes: applied and ecological perspectives* (pp. 519-521). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Goldsmith, T.E., & Johnson, P.J. (2002). Assessing and improving evaluation of aircrew performance. *Human Factors*, 12(3), 223-240.
- Groot, A.D. de (1965). *Thought and choice in chess*. The Hague: Mouton.
- Guion, R.M. (1986). Personnel evaluation. In R.A. Berk (Ed.), *Performance assessment: methods and applications* (pp. 345-360). Baltimore: The Johns Hopkins University.
- Guion, R.M. (1998). *Assessment, measurement and prediction for personnel decisions*. Mahwah NJ: Lawrence Erlbaum Associates.
- Hacker, W. (1998). *Allgemeine arbeitspsychologie. Psychische regulation von arbeitstätigen* [Work psychology. Psychic regulation of actions at work]. Bern: Verlag Hans Huber.
- Hadley, G., Guttman, J., Stringer, P. (1999). *Air traffic control specialist performance measurement database*. Washington, DC: Federal Aviation Administration (DOT/FAA/CT-TN99/17).
- Hamman, W.R., Beaubien, J.M., & Holt, R.W. (1999). Evaluating instructor/evaluator interrater reliability from performance database information. In R. Jensen (Ed.), *Proceedings of the tenth international symposium on aviation psychology*. Columbus, OH: The Ohio State University Press.
- Handyside, J.D. (1989). On ratings and rating scales. In P. Herriot (Ed.), *Assessment and selection in organizations: methods and practice for recruitment and appraisal*. Chichester: John, Wiley & Sons.
- Heijden, B. van der (1998). *The measurement and development of professional expertise throughout the career*. Doctoral dissertation, Technical University, Enschede.
- Helmreich, R.L., & Davies, J.M. (1997). Anaesthetic simulation and lessons to be learned from aviation. *Canadian journal of anaesthesia*, 44(9), 907-912.
- Hettema, P.J. (1979). *Personality and adaptation*. Amsterdam: North-Holland Publishing Company.
- Hettema, P.J. (Ed.) (1989). *Personality and environment: assessment of human adaptation*. Chichester: Wiley.
- Hettema, P.J., Leidelmeijer, K.C., & Geenen, R. (2000). Dimensions of information processing: physiological reactions to motion pictures. *European journal of personality*, 14, 39-63.
- Hettema, P. J. (2001). *Persoonlijkheid van top tot teen* [Personality all over]. Assen: Van Gorcum.

- Hockey, G.R. (1997). Compensatory control in the regulation of human performance under stress and high workload: a cognitive-energetical framework. *Biological psychology*, 45, 73-93.
- Hockey, G.R., Gaillard, A.W., & Coles, M.G. (Eds.) (1986). *Energetics and human information processing*. Dordrecht: Nijhoff.
- Hockey, G.R. (2003). Operator functional state as a framework for the assessment of performance degradation. In G.R. Hockey, A.W.K. Gaillard, & O. Burov (Eds.), *Operator functional state. The assessment and prediction of human performance degradation in complex tasks* (pp. 8-23). Amsterdam: IOS Press.
- Hoekstra, H.A., & Sluijs, E. van (2000). *Management van competenties. Het realiseren van HRM* [Management of competences. The realization of HRM]. Assen: Van Gorcum.
- Holt, R.W., Hansberger, J.T., & Boehm-Davis, D.A. (2002). Improving rater calibration in aviation: a case study. *The international journal of aviation psychology*, 12(3), 305-330.
- Hopkins, D. (1995). *Human factors in air traffic control*. London: Taylor & Francis.
- Hox J. (2002). *Multilevel analysis: techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.
- International Civil Aviation Organization (2004). *Air traffic management*. Montreal: International Civil Aviation Organization.
- Isaac, A.R., & Ruitenbergh, B. (1999). *Air traffic control: human performance factors*. Aldershot: Ashgate.
- Jacobs, R.L., & Jones, M.J. (1995). *Structured on-the-job training. Unleashing employee expertise in the workplace*. San Francisco: Berrett-Koehler Publishers.
- James, L.R., Demaree, R.G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of applied psychology*, 69(1), 85-98.
- John, B.E., & Lallemand, Y. (1997). Strategy use while learning to perform the Kanfer-Ackerman air traffic controller task. In M.C. Shafto & P. Langley (Eds.), *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 337-342).
- Jonassen, D.H. (1992). Evaluating constructivistic learning. In T. Duffy & D.H. Jonassen (Eds.) (1992), *Constructivism and the technology of instruction: a conversation* (pp. 137-148). Hillsdale, N.J, London: Erlbaum.
- Jonassen, D.H., & Grabowski, B.L. (1993). *Handboek of individual differences, learning and instruction*. Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Jong, T. de, & Joolingen, W.R. van (1996). *Discovery learning with computer simulations of conceptual domains*. Enschede: Universiteit Twente. (IST memorandum, vakgroep instructietechnologie, 96-02)
- Kane, J.S. (1986). Performance distribution assessment. In R.A. Berk (Ed.), *Performance assessment: methods and applications* (pp. 237-274). Baltimore: The Johns Hopkins University
- Kane, J.S., Bernardin, H.J., Villanova, P., & Peyrefitte, J. (1995). Stability of rater leniency: three studies. *Academy of management journal*, 38(4), 1036-1051.
- Kanfer, R., Ackerman, P.L. (1989). Motivation and cognitive abilities. An integrative aptitude treatment approach to skill acquisition. *Journal of applied Psychology*, 74, 657-690.
- Kessels, J. (1999). Het verwerven van competenties: kennis als bekwaamheid [The acquisition of competences: Knowledge as capability]. *Opleiding en ontwikkeling*, 12, 7-11.
- Kessels, J. & Smit (1999). *Opleidingskunde*. Amsterdam: Wolters Kluwer.
- King, R.E., & Dattel, A.R. (2005). The Air Traffic Control selection and training battery: what it is and isn't (and how it has changed and hasn't). In R. Jensen (Ed.), *Proceedings of the 13th Symposium on Aviation Psychology*. Dayton: Wright State University.
- Klarus, R. (2000). Beoordeling en toetsing in het nieuwe onderwijsconcept [Assessment and testing in the new education concept]. In J. Onstenk (Ed.), *Op zoek naar een krachtige beroepsgerichte leeromgeving* (pp. 167-204). Den Bosch: CINOP.
- Kojevnikov, A. (2001). Spotlight on engine rooms. *Shipping world and shipbuilder*, 202, 54-58.
- Kolb, D.A. (1984). *Experiential learning*. Englewood Cliffs, NJ: Prentice Hall.
- Kuhl, J. (1994). A theory of action and state orientation. In J. Kuhl, & J. Beckman (Eds.), *Volition and personality: action versus state orientation* (pp. 9-46). Seattle: Hogrefe & Huber.
- Kommers, P. & Zhiming, Z. (1999). *Virtual reality for education*. Retrieved April 15, 2006, from University of Twente, Enschede, website: <http://projects.edte.utwente.nl/proo/kommers.htm>.

- Lamoureux, T.M., & Cox, M.J. (1999). *Cognitive task analysis of CATC radar skills course*. Bournemouth: National Air Traffic Services (8RD/14/16/26/1014/HF).
- Landy, F.L., & Farr, J.L. (1983). *The measurement of work performance. Methods, theory and applications*. New York: Academic Press.
- Landy, F., Zedeck, Z., & Cleveland, J. (Eds.) (1983). *Performance measurement and theory*. Hillsdale NJ: Lawrence Erlbaum Associate Publishers.
- Larkin, J.H., McDermott, J., Simon, D.P., & Simon, H.A. (1980). Expert and novice performance in solving physics problems. *Science*, 208, 1335-1342.
- Laurillard, D. (1993). *Rethinking University Teaching: a framework for the effective use of educational technology*. London: Routledge.
- Lave, J., & Wenger, E. (1991). *Situated learning. Legitimate peripheral participation*. Cambridge: Cambridge University Press.
- Lee, F. & Anderson, J.R. (2001). Does learning a complex task have to be complex? A study in learning composition. *Cognitive psychology*, 42, 267-316.
- Linn, R.L., Baker, E.L., & Dunbar, S.B. (1991). *Complex performance-based assessment: expectations and validation criteria*. Los Angeles, CA: National Center for Research on Evaluation, Standards and Student Testing (CRESST, SCE report 331).
- Linn, R.L., & Baker, E.L. (1996). Can performance-based student assessments be psychometrically sound? In J.B. Baron & D.P. Wolf (Eds.), *Performance-based student assessment: challenges and possibilities* (pp. 84-103). Chicago, Illinois: The university of Chicago press.
- Lowyck, J., & Verloop, N. (1995). *Onderwijskunde. Een kennisbasis voor professionals* [Educational science. A knowledge base for professionals]. Groningen: Wolters-Noordhoff.
- Mager, R.F. (1962). *Preparing instructional objectives*. Belmont, CA: Fearon.
- Manning, C.A. (2000). *Measuring air traffic controller performance in a high-fidelity simulation*. Oklahoma City: Federal Aviation Administration (DOT/FAA/AM-00/2).
- Manning, C., & Stein, E. (2005). Measuring air traffic controller performance in the 21st century. In B. Kirwan, M. Rodgers, & D. Schäfer (Eds.), *Human factors impacts in air traffic management* (pp. 283-316). Hampshire: Ashgate.
- Mayer, R. (1987). *Educational psychology*. California: Harper Collins.
- Merriënboer, J.J.G. van (1997). *Training complex skills: a four component instructional design model for technical training*. Englewood Cliffs, NJ: Educational Technology Publications.
- Merriënboer, J.J.G. van, & Kirschner, P.A. (2007). *Ten steps to complex learning*. Mahwah, NJ: Erlbaum / Taylor and Francis.
- Merrill, D. (1994). *Instructional design theory*. Hillsdale, New Jersey: Educational technology Publications.
- Messick, S. (1993). Validity. In R.L. Linn (Ed.), *Educational measurement* (3th ed., pp. 13-103). Phoenix, AZ: Oryx Press.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13-23.
- Moertl, P.M., Canning, J.M., Gronlund, S.D., Dougherty, M.R.P., Johansson, J., & Mills, S.H. (2002). Aiding planning in air traffic control: an experimental investigation of the effects of perceptual information integration. *Human Factors*, 44(3), 404-412.
- Moonen, J. (2000). Design methodology. In H. Adelsberger, B. Collis, & J. Pawlowski (Eds.), *Handbook on information technologies for education and training* (pp. 153-180). Berlin: Springer-Verlag.
- Morgan, P.J., & Guest, C.B. (2001). A comparison of global ratings and checklist scores from an undergraduate assessment using an anaesthesia simulator. *Academic medicine*, 76(10), 1053-1055.
- MRU (1993). *Recruitment, selection and training of air traffic controller students* (Report No. 10). Norrköping: Swedisch Civil Aviation Administration, ANS department.
- Mulder, G. (1986). The concept and measurement of mental effort. In G.R. Hockey, A.W. Gaillard, & M.G. Coles (Eds.), *Energetics and human information processing* (pp. 175-198). Dordrecht: Nijhoff.

- Mulqueen, C., Baker, D.P., & Dismukes, R.K. (2002). Pilot instructor rater training: the utility of the multifaceted item response theory model. *International journal of aviation psychology*, 12(3), 287-303.
- Murphy, K.R., & Cleveland, J.N. (1995). *Understanding performance appraisal: social, organizational and goal-based perspectives*. London: Sage publications.
- Murphy, K.R., Jako, R.A., & Anhalt, R.L. (1993). Nature and consequence of halo error: a critical analysis. *Journal of applied psychology*, 78(2), 218-225.
- Murphy, K.R., & Balzer, W.K. (1989). Rater errors and rating accuracy. *Journal of applied psychology*, 74(4), 619-624.
- Newell, A., & Simon, H. (1972). *Human problem solving*. Englewood cliffs, NJ: Prentice-Hall.
- Newell, A., & Rosenbloom, P.S. (1981). Mechanisms of skill acquisition and the law of practice. In J.R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1-55). Hillsdale, NJ: Erlbaum.
- Niessen, C. (1999). Modeling cognitive processes of experienced air traffic controllers. *Ergonomics*, 42(11), 1507-1520.
- Nolan, M.S. (1999). Air traffic control. In D.J. Garland, J.A. Wise, & V.D. Hopkin (Eds.) *Handbook of aviation human factors* (pp. 431-454). Mahwah, NJ: Erlbaum.
- Nonaka, I., & Takeuchi, H. (1995). *The knowledge-Creating Company*. New York: Oxford University Press.
- Nunes, A., & Mogford, R.H. (2003). Identifying controller strategies that support the 'picture'. *Proceedings of the 47th Annual Meeting of the Human Factors and Ergonomics Society*. Santa Monica: Human Factors and Ergonomics Society.
- Nunnally, J.C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
- O'Connor, P., Hormann, H., Flin, R., Lodge, M., & Goeters, K. (2002). Developing a method for evaluating crew resource management skills: a European perspective. *The international journal of aviation psychology*, 12(3), 263-285.
- O'Connor, P. (2004). JAR-TEL results: Interrater reliabilities, sensitivity, and acceptability of the NOTECHS method. In K. M. Goeters (Ed.), *Aviation psychology: Practice and research* (pp. 257-271). Hampshire: Ashgate.
- Onstenk, J. (1997). *Lerend leren werken. Brede vakbekwaamheid en de integratie van leren, werken en innoveren* [Learning while working: broad competences and the integration of learning, working and innovation]. Doctoral dissertation. Delft: Eduron.
- Onstenk, J. (2000). Brede vakbekwaamheid als leerdoel [Broad competences as learning objective]. In J. Onstenk, J. (Ed.), *Op zoek naar een krachtige beroepsgerichte leeromgeving* (pp. 41-62). Den Bosch: Cinop.
- Oprins, E., & Dankbaar, M. (2002). *Kennismanagement: methoden om leerprocessen in werkomgevingen te managen* [Knowledge management: Methods to manage learning processes in work environments]. Den Bosch: CINOP.
- Oprins, E., & Schuwer-van Blanken, M. (2003). Competentiegericht opleiden en beoordelen bij LVNL [Competence-based training and assessment at LVNL]. *HUFAG Nieuwsbrief*, 6, 2-4.
- Oprins, E., Burggraaff, E. & Weerdenburg, H. van (2005). Competence-based assessment design for air traffic control training. In R. Jensen (Ed.), *Proceedings of the 13th Symposium on Aviation Psychology*. Dayton: Wright State University.
- Oprins, E. & Burggraaff, E. (2006). Learnability of situational awareness and workload management in air traffic control training. *Proceedings of the 16th World Congress on Ergonomics*.
- Oprins, E., Burggraaff, E. & Weerdenburg, H. van (2006a). Design of a competence-based assessment system for ATC training. *The international journal of aviation psychology*, 16(3), 297-320.
- Oprins, E., Burggraaff, E. & Weerdenburg, H. van (2006b). Reliability and validity of a competence-based assessment system for air traffic control training. *Proceedings of the 27th conference of the European Association for Aviation Psychology*.
- Oprins, E., Burggraaff, E. & Weerdenburg, H. van (2007a). Analysis of learning curves: evaluation of assessment in air traffic control training. In R. Jensen (Ed.), *Proceedings of the 13th Symposium on Aviation Psychology*. Dayton: Wright State University.
- Oprins, E., Burggraaff, E. & Weerdenburg, H. van (2008). Reliability of assessors' ratings in competence-based air traffic control training. *Human factors and aerospace safety*, 6 (4), 305-322.

- Oprins, E., Geven, E., Veldhuijzen, E., & Roe, R. (2006). Development of a new selection system for air traffic controllers: Design, implementation and initial validity evidence. *Proceedings of the 27th conference of the European Association for Aviation Psychology*.
- Oprins, E., Geven, E., Veldhuijzen, E., & Roe, R. (2007). Procedural control in air traffic control selection tests to predict situational awareness. In R. Jensen (Ed.), *Proceedings of the 14th Symposium on Aviation Psychology*. Dayton: Wright State University.
- Pellegrino, J.M., Chudowski, N., & Glaser, R. (2001). *Knowing what students know. The science and design of educational assessment*. Washington: National Academic Press.
- Petraglia, J. (1998). *Reality by design: the rhetoric and technology of authenticity in education*. Mahwah, NJ: Erlbaum.
- Polanyi, M. (1967). *The tacit dimension*. New York: Doubleday & Co.
- Pugh, C.M., & Youngblood, P. (2002). Development and validation of assessment measures for a newly developed physical examination simulator. *Journal of the American medical informatics association*, 9(5), 448-460.
- Ramos, R.A., Heil, M.C., & Manning, C.A. (2001). *Documentation of validity for the AT-SAT Computerized Test Battery* (Vol. 1&2). Oklahoma City: Federal Aviation Administration (DOT/FAA/AM-01/6).
- Rasmussen, J. (1986). *Information processing and human-machine interaction: an approach to cognitive engineering*. New York: North Holland Series.
- Redding, R.E., & Seamster, T.L. (1994). Cognitive task analysis in air traffic controller and aviation crew training. In N. Johnston, N. McDonald, & R. Fuller (Eds.), *Aviation psychology in practice* (pp. 190-122). Hants, England: Avebury Technical.
- Regian, J.W., & Schneider, W. (1990). Assessment procedures for predicting and optimizing skill acquisition after extensive practice. In N. Frederiksen, R. Glaser, A. Lesgold, & M.G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 297-323). Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Resnick, L.B., & Resnick, P.D. (1992). Assessing the thinking curriculum: new tools for educational reform. In B.R. Gifford & M.C. O'Connor (Eds.), *Changing assessments: alternative views of aptitude, achievement and instruction* (pp. 37-75). Boston: Kluwer Academic Publishers.
- Ridderbos, A. (1992). *Selection by simulation*. Doctoral dissertation, Technical University, Eindhoven.
- Roe, R.A. (1989). Designing selection procedures. In P. Herriot (Ed.), *Assessment and selection in organizations: methods and practice for recruitment and appraisal* (pp. 127-142). Chichester: John Wiley & Sons.
- Roe, R.A. (1999). Work performance: a multiple regulation perspective. In C.L. Cooper, & I.T. Robertson (Eds.), *International review of industrial and organizational psychology* (Vol. 14, pp. 231-335). Chichester: John Wiley & Sons.
- Roe, R.A., & Daniels, M.J.M. (1994). *Personeelbeoordeling: achtergrond en toepassing* [Personnel appraisal. Background and application] (3th ed.). Assen: Van Gorcum.
- Roe, R.A. (2002). Competenties: een sleutel tot integratie in theorie en praktijk van de A&O psychologie [Competences – A key towards integration in theory and practice in W&O Psychology]. *Gedrag en organisatie*, 15(4), 203 – 224.
- Roe, R.A. (2003). *Programma van Eisen selectiesysteem LVNL* [Program of Requirements selection system LVNL]. Internal report. Amsterdam: Air Traffic Control the Netherlands.
- Roe, R.A. (2005). The design of selection systems: contexts, principles, issues. In A. Evers, O. Smit, & N. Anderson (Eds.), *Handbook of personnel selection* (pp. 73-97). Oxford: Blackwell.
- Roe, R.A. & Hermans, P. (2006). Psychological factors in crew selection. In Bor, R. & Hubbard, T. (Eds.), *Aviation mental health* (pp. 161-193). Aldershot: Ashgate Publishers.
- Roessingh, J.J.M. (2005). Transfer of manual flying skills from PC-based simulation to actual flight – comparison of in-flight measured data and instructor ratings. *The international journal of aviation psychology*, 15(1), 67-90.
- Roessingh, J.J.M. (2002). *The acquisition of complex skills*. Doctoral dissertation, University of Utrecht, Utrecht.
- Romiszowski, A. (1988). *Designing instructional systems. Decision making in course planning and curriculum design*. Londen: Kogan page.

- Roozenburg, N.F.M., & Eekels, J. (1991). *Productontwerpen, structuur en methoden*. [Product Design: Fundamentals and Methods]. Utrecht: Lemma BV.
- Rose, F.D. (2000). Training in virtual environments: transfer to real world tasks and equivalence to real task training. *Ergonomics*, 43(4), 494-511.
- Rothwell, W.J., & Kazanas, H.C. (1994). *Improving on-the-job training: how to establish and operate a comprehensive OJT program*. San Francisco: Jossey-Bass Publishers.
- Rasmussen, J. (1986). *Information processing and human-machine interaction: an approach to cognitive engineering*. Amsterdam: Elsevier.
- Saal, F.E., Downey, R.G., & Lahey, M.A. (1980). Rating the ratings: assessing the psychometric quality of rating data. *Psychological bulletin*, 88(2), 413-428.
- Salden, R., Paas, F., Broers, N., & Merriënboer, J. van (2004). Mental effort and performance as determinants for the dynamic selection of learning tasks in air traffic control training. *Instructional Science*, 32, 153-172.
- Salden, R. (2005). *Dynamic task selection in aviation training*. Doctoral dissertation, Open University the Netherlands, Maastricht.
- Salden, R., Paas, F., Pal, J. van der, & Merriënboer, J. van (2006). Dynamic task selection in flight management system training. *The international journal of aviation psychology*, 16(2), 157-174.
- Sanders, A.F. (1991). Simulation as a tool in the measurement of human performance. *Ergonomics*, 34(8), 995-1025.
- Sanders, A.F. (1986). Energetical states underlying task performance. In G.R. Hockey, A.W. Gaillard, & M.G. Coles (Eds.). *Energetics and human information processing* (pp. 139-153). Dordrecht: Nijhoff.
- Sauer J., Hockey G.R.J., & Wastell D.G. (2000). Effects of training on short- and long-term skill retention in a complex multiple-task environment. *Ergonomics*, 43 (12), 2043-2064.
- Schaafstal, A.M. (1991). *Diagnostic skill in process operation: a comparison between experts and novices*. Doctoral dissertation, University of Groningen.
- Schijven, M.P., & Jakimowicz, J. (2004). The learning curve on the Xitact LS 500 laparoscopy simulator: profiles of performance. *Surgical endoscopy*, 18, 121-127.
- Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological assessment*, 8(4). 350-352.
- Schneider, W. (1990). Training high-performance skills: fallacies and guidelines. In M. Venturino (Ed.). *Selected readings in human factors* (pp. 297-311). Santa Monica, CA: The human factors society.
- Schneider Lidz, C. (1987) *Dynamic assessment. An interactional approach to evaluating learning potential*. New York: The Guilford Press.
- Schneider, W., & Shiffrin, R.M. (1977). Controlled and automatic human information processing: I. Detection, search and attention. *Psychological review*, 84, 1-66.
- Schunn, C.D., & Reder, L.M. (1998). Strategy adaptivity and individual differences. In D.L. Medin (Ed.), *The psychology of learning and motivation*, vol. 38 (pp. 115-154). New York: Academic Press.
- Seamster, T., Redding, E., & Kaempff, G. (1997). *Applied cognitive task analysis in aviation*. Aldershot: Avebury Aviation.
- SESAR (2007). *The ATM target concept. SESAR definition phase – deliverable 3 (DLM-0612-001-02-00)*. Brussels: Eurocontrol.
- Shavelson, R.J., Gao, X., & Baxter, G.P. (1996). On the content validity of performance assessments: centrality of domain specification. In M. Birenbaum & F. Dochy (Eds.). *Alternatives in assessment of achievements, learning process and prior knowledge* (pp. 131-141). Boston: Kluwer Academic Publishers.
- Shebilske, W.L., Goettl, B.P., & Garland, D.J. (2000). Situation awareness, automation and training. In M.R. Endsley & D.J. Garland (Eds.), *Situation awareness: analysis and measurement* (pp. 303-324). Mahwah, NJ: Lawrence Erlbaum Associates.
- Shiffrin, R.M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological review*, 84, 127-190.

- Silver, N.C., & Dunlap, W.P. (1987). Averaging correlation coefficients: Should Fisher's z transformation be used? *Journal of applied psychology*, 72(1), 146-148.
- Skinner, B.F. (1938). *The Behavior of organisms: an experimental analysis*. New York: Appleton-Century-Crofts.
- Smith, P., & Kendall, L.M. (1963). Retranslation of expectations: an approach to the construction of unambiguous anchors for rating scales. *Journal of applied psychology*, 47, 149-155.
- Snow, R.E. (1989). Aptitude Treatment Interactions. In: P.J. Ackerman, R.J. Sternberg, & R. Glaser (Eds.) *Learning and individual differences: advanced in theory and research* (pp. 13-59). New York: W.H. Freeman & company.
- Spencer, L.M., & Spencer, S.M. (1993). *Competence at work: models for superior performance*. New York: Wiley.
- Sperandio, A. (1978). The regulation of working methods as a function of workload among air traffic controllers. *Ergonomics*, 21, 367-390.
- Spiro, R.J., Feltovich, P.J., Jacobson, M.J., & Coulson, R.L. (1992). Cognitive flexibility, constructivism, and hypertext: random access instruction for advanced knowledge acquisition in ill-structured domains. In T. Duffy & D.H. Jonassen (Eds.), *Constructivism and the technology of instruction: a conversation* (pp. 24-33). Hillsdale, N.J. London: Erlbaum.
- Sauer, J. Hockey, G.R.J., & Wastell, D.G. (2000). Effects on training on short- and long-term skill retention in a complex multiple-task environment. *Ergonomics*, 43, 2043-2064.
- Taatgen, N.A. (2001). A model of individual differences in skill acquisition in the Kanfer-Ackerman air traffic control task. *Cognitive system research*, 3(1), 103-112.
- Taatgen, N.A., & Lee, F.L. (2003). Production compilation: a simple mechanism to model complex skill acquisition. *Human Factors*, 45(1), 61-76.
- Thijssen, J. (1998). Hindernissen voor competentie management [Barriers for competence management]. *Opleiding & ontwikkeling*, 11(10), 17-23.
- Thomas, E.J., Sexton, J.B., & Helmreich, R.L. (2004). Translating teamwork behaviours from aviation to healthcare: development of behavioural markers for neonatal resuscitation. *Quality and safety in health care*, 13(1), i57-i64.
- Thorndike, E.L. (1911). *Animal intelligence*. New York: MacMillan.
- Thorndike, E.L. (1920). A constant error in psychological ratings. *Journal of applied psychology*, 4, 25-29.
- Tillema, H.H. (Ed.) (1996). *Development centers. Ontwikkelen van competenties in organisaties*. Groningen: Kluwer bedrijfswetenschappen.
- Vreuls, D., & Obermayer, R.W. (1985). Human-system performance measurement in training simulators. *Human Factors*, 27(3), 241-250.
- Wickens, C.D. (1992). *Engineering psychology and human performance* (2nd ed.). New York: HarperCollins Publishers.
- Wickens, C.D., Mavor, A.S., & McGee, J.P. (Eds.) (1997). *Flight to the future: human factors in air traffic control*. Washington: National Academy Press.
- Wigdor, A.K. & Green, B.F. (1991). *Performance assessment for the workplace (volume 1 and 2)*. Washington: National Academic Press.
- Wiggins, G. (1990). The case for authentic assessment. *Practical Assessment, Research & Evaluation*, 2(2). Retrieved April 15, 2006, from <http://PAREonline.net/getvn.asp?v=2&n=2>.
- Wolf, D.P. (1993). Assessment as an episode of learning. In R.E. Bennett, & W.C. Ward (Eds.), *Construction versus choice in cognitive measurement: issues in constructed response, performance testing and portfolio assessment* (pp. 213-240). London: Lawrence Erlbaum Associates.
- Yerkes, R.M., & Dodson, J.D. (1908). The relation of strengths of stimulus to rapidity of habit formation. *Journal of comparative neurology and psychology*, 18, 459-482.
- Youngstrom, E., Loeber, R., & Stouthamer-Loeber, M. (2000). Patterns and correlates of agreement between parent, teacher, and male adolescent ratings of externalizing and internalizing problems. *Journal of consulting and clinical psychology*, 69(6), 1038-1050.
- Zijlstra, F.R.H. (1993). *Efficiency in work behaviour: a design approach for modern tools*. Doctoral dissertation, University of Delft.

Summary

Like other process control tasks in transportation (aviation, shipping, railways) or process industries (e.g., chemical and nuclear plants), the ATC task is considered a complex cognitive skill. Learnability seems to be limited. The strict safety requirements make the performance standards even higher because any error caused by incompetence of operators must be avoided. Consequently, the outcome from training is often too low despite of high selection standards. This may result in a shortage of competent personnel. Low pass rates are also undesirable because the training is usually very time-consuming and expensive. High-fidelity simulators are used to train trainees to the highest possible level before they enter the safety-critical working environment in on-the-job training (OJT).

A well-designed assessment system can contribute to solve this problem in several ways. First, assessment can support learning processes through feedback and adaptation of training to individuals' needs. This increases the chance for trainees to successfully complete the training. Second, reliable and valid pass-fail decisions can be produced by restricting the number of false positives, i.e. prospective-less trainees who unnecessarily occupy expensive training positions during later phases, and false negatives, i.e. failed trainees who could have had a chance to succeed training ultimately. Third, predictive validity of selection can be improved using more reliable training criteria in validation studies.

This study is the result of a project internally performed at Air Traffic Control the Netherlands (LVNL). Its objective was to design an assessment system that optimally supports the acquisition of air traffic controller competences in simulator and on-the-job training (OJT).

Design methodology

We combined two types of design methodologies, mainly based on the design cycle of Roozenburg and Eekels (1991) and the training design methodology of Moonen (2000). This resulted in a cyclic design process of three phases: (1) *analysis*; (2) *design and implementation*; (3) *evaluation*. The thesis is divided into three parts that correspond with these design phases. The first phase, *analysis*, comprises a literature review, an ATC competence analysis, and the construction of a Program of Requirements (PoR) with stakeholders. The second phase, *design and implementation*, refers to the design and development of the assessment system itself, and the introduction of the system in the training. The third phase, *evaluation*, involves the evaluation of the assessment system in three parts: (1) psychometric quality; (2)

learning processes (analysis of learning curves and competence development); (3) user evaluation. Finally, an overall evaluation against the Program of Requirements was held. We have learned that the design phases occur iteratively. A key success factor was the intensive collaboration with the user group (air traffic controllers) in designing an innovative product that was completely accepted for practical use.

Analysis

We started with a literature review that covers: (1) general theories about performance, competences and learning; (2) domain-specific literature about the nature of the ATC task and learning processes in ATC; (3) relevant literature about assessment that could be used in the context of ATC training. The findings were used in the design of the new assessment system.

Next, we did a competence analysis by organizing two workshops. Controllers were involved as subject matter experts (SME's) to explicate their implicit knowledge. Thirteen competences, represented by a set of behavioural markers, were formulated in their own jargon. This should enhance recognizability and common understanding for practical use. The next step was to operationalize the competences. We compared the list with published (cognitive) task analyses looking for additional aspects that might have been forgotten. This process resulted in the ATC Performance Model, which has served as a framework for the assessment design at LVNL (Oprins, Burggraaff & Van Weerdenburg, 2006).

Finally, a Program of Requirements (PoR), a set of requirements and constraints on the new assessment system, was constructed together with a group of stakeholders. We analysed the shortcomings of the previous system and the purposes of the redesign. The PoR has primarily guided the design of the new assessment system.

Design and implementation

The design process occurred in strong collaboration with the user group (managers, coaches and trainees) as part of the design teams. This has enhanced the implementation of the assessment system. We made choices in the design of the system to fulfill its purposes, and this has led to some specific properties.

The assessment system is competence-based. This implies that competences are assessed: the successful integration of knowledge, skills and attitudes and their application in realistic environments. Competences optimally reflect the individual abilities to perform effectively without paying attention to situational factors that may influence performance. All aspects that belong to competences are assessed (technical, cognitive, emotional, social) to get a complete picture. The competences were directly derived from the ATC Performance Model. This has resulted in the following set of competences for ACC: *safety, efficiency, verbal expression, listening, coordination, equipment operation, strip/label management, mental picture, attention management, planning, decisiveness, workload management, attitude, and teamwork ability*. Since they are not directly visible, each competence is represented by a set of *performance criteria*, formulated as 'behavioural markers' and rated at a six

points rating scale. During training, the same competences are assessed in order to follow the trainee's progression on each competence over a training period, based on various task situations and circumstances. The generic character of competences makes this possible, for instance, *planning* is relevant in any ATC task execution whether simple or complex. Progression provides an important indicator of whether a trainee is still learning or has reached a learning plateau. Potential deficiencies of trainees can be detected in an early stage so that training can be maximally adapted to the trainee's needs.

The competences are assessed against augmenting *performance standards* during training. Simulator training (preOJT) and OJT comprise successive phases with specific standards for each phase (cf. 'norms'). The standards were formulated as exemplary behaviours, a variant on behaviourally anchored rating scales (BARS). They do not specify scale positions but standards to be achieved at the end of each phase. In this way, assessors have more agreement in what is expected from trainees in intermediate phases. For trainees it is clearer which competences they have to develop further in a specific phase. The standards in preOJT are mainly defined by the sequence of simulator exercises. The classification into phases is based on two main variables: *traffic handling* (degree of safety and efficiency required) and *traffic complexity* (degree of task difficulty). Structuring OJT is more difficult due to the ongoing live traffic and was therefore most innovative. The division in phases in OJT, four phases in total, was based on three main principles: *traffic handling*, *traffic complexity*, and *aid of the coach*. The length of each phase is flexible, dependent on the trainee's progression.

Continuous assessment is applied to get a complete picture of the trainee's performance and to measure progression over time. Coaches are continuously in interaction with trainees and can force them to verbalize their thoughts. Assessment of cognitive processes is required to obtain diagnostic information on performance and especially on performance shortcomings. *Progression reports* are filled in after one or two weeks of training in preOJT and OJT. The time intervals vary because of the different operational schedules. The coach has a double role as an assessor and a coach, and multiple assessors are involved. In preOJT, additional *simulator tests* are applied to measure trainees' performance objectively at a certain moment of time and in well-balanced test scenarios without interference by the coach. Pass-fail decisions are made at the end of each phase in preOJT and OJT. There does not exist a fixed cut-off, but decisions are based on an extensive quantitative and qualitative analysis of simulator tests and progression reports.

A web-based assessment tool is designed to fill in progression reports, to store the results in a database, and to generate several overviews of trainee performance. In this way, interested persons who have access from several places can better follow the trainee's progression over time. Training results can easily be used for reliability and validity studies. The first version was made in the assessment tool Questionmark Perception. A project has started to design a new tool because of some technical problems.

Much effort was put in the implementation of the assessment system, as this was considered very important for its correct use and acceptance. The close collabora-

tion with the users made that the system was accepted rather easily. Furthermore, we introduced the system carefully by starting with small pilots and by organizing presentations. We paid extra attention to assessment in coach and assessor training. Intermediate evaluations resulted in small changes before the final version was launched.

Evaluation

The requirements and constraints in the PoR guided the evaluation that comprised three parts. The data were from 34 trainees in ACC preOJT (188 progression reports; 36 simulator tests), and from 27 trainees in ACC OJT (407 progression reports), collected during the period January 2001 until December 2006.

Evaluation of psychometric quality

We investigated the following reliability issues: (1) *interrater agreement*; (2) *rating errors* (halo error, leniency and severity error, range restriction and central tendency error); (3) *test reliability* (internal consistency, split-half reliability). Predictive validity was part of the evaluation of learning processes (see next paragraph) because this involves analyses over time.

Interrater agreement was investigated by using simulator tests in which two assessors filled in test reports independently of each other. Their ratings should be interchangeable since assessments may not be dependent on the person who assesses. We examined three types of indices in interrater agreement: shape, dispersion, and level. The results have shown that the interrater agreement between assessors is sufficiently high for the overall performance level (weighted sum of competence ratings), but only moderate with respect to profile similarity (shape, dispersion). This implies that some assessors give low ratings on some competences, while others give high ratings on the same competences and vice versa. Apparently, assessors find it difficult to point at specific deficiencies of trainees.

Next, the presence of the most important rating errors was examined. Analyses of *leniency* and *severity errors* showed that assessors tend to leniency as expected: the mean competence ratings are above the scale midpoint. We found a few systematic differences in leniency between assessors: there exist some 'Santa Claus' (high means) and 'Axeman' (low means) assessors. Furthermore, assessors generally show *range restriction*. They do not equally distribute their ratings over the 6-points rating scale, but tend to give positive ratings and avoid the extreme scale points; thus, *central tendency* does not occur. Differences between assessors in their distribution of ratings, expressed in congruency indexes, were not found. Finally, we examined the occurrence of *halo errors*. We found high intercorrelations between some competence ratings. We recognized the classification of the ATC Performance Model in the results. Specific competences should be intercorrelated because they are conceptually related to each other. Therefore, we concluded that the presence of halo errors is not severe. In general, rating errors were found more often in progression reports. This confirms that simulator tests are more reliable.

Finally, we examined *internal consistency* and *split-half reliability*. Calculations of the Cronbach's alphas and item-total correlations per competence have shown

that the internal consistency is very high. Only a small number of criteria should be deleted or changed. Split-half reliability was estimated for simulator tests with the Spearman-Brown formula. We found very high reliability coefficients. Therefore, we considered test reliability of the assessment system as sufficient.

In sum, the findings have shown that the reliability of the assessment system is sufficient to a certain extent. The system is well designed in terms of the classification in competences, performance criteria and performance standards. However, the role of the assessor can be improved. Assessor training is organized to pay attention to the avoidance of rating errors and to achieve a common understanding of the competences.

Evaluation of learning processes

A well-designed assessment system works adequately if the assessment results represents learning processes adequately. Patterns and individual differences in learning (e.g., slow starters, learning plateaus) and in performance (strengths and weaknesses) should be clearly distinguished to provide a basis for adequate feedback and interventions. Learning curves can be derived from the assessment results based on a sequence of performance measures over time. These learning curves should be sufficiently representative for learning. Then the training can be maximally adaptive to the trainee's needs. Pass-fail decisions will be more valid if they are based on the predictability of patterns in learning processes.

Learning curves are usually presented as growth curves based on repeated executions of the same task at successive moments of time, but this is not possible with our assessment system. The trainee's performance is assessed against augmenting performance standards, which are constantly translated into the same 6-points rating scale. Therefore, the learning curves produced by our assessment system could rather be referred to as recalibrated learning curves. We compared the learning curves derived from progression reports in preOJT and OJT with prototypical learning curves that are expected from general learning theory. We defined three groups: (1) *high performers* (passed without problems); (2) *moderate performers* (passed with difficulties); (3) *low performers* (failed). Three training managers classified the trainees into the groups based on expert judgment, serving as an external criterion for trainee success.

We distinguished two quantitative variables that define the learning curves, subdivided into four measures: (1) *performance*: mean performance level; occurrences of insufficient performance; (2) *progression*: growth; rate of growth. We examined the distinction between the three groups visually and quantitatively. The results showed that the learning curves of actual trainees are sufficiently representative for learning processes, because they reflect the prototypical learning curves adequately. Although the number of trainees is too low yet for getting convincing evidence, we also explored the predictability of learning curves over time. The findings suggested that learning curves are predictive for next phases to a certain extent, but it is too early for defining strict cut-offs.

A well-designed competence-based assessment system should not only be able to represent general learning processes as expressed in the learning curves, but also the development of specific competences. This is important for identifying

possible deficiencies. We used the same classification into the three groups of performers. The findings were in conformance with the literature with respect to trainability of competences. As expected, we found significant differences between individual trainees for more critical and less trainable competences such as *mental picture* and *workload management*, than for competences that are less critical and more trainable such as *label management* and *equipment operation*. The main reasons for failing are respectively: *mental picture*, *workload management*, *attention management* and *decisiveness*. We concluded that the assessment system sufficiently represents individual differences in learning processes. More longitudinal research is needed to get evidence of predictive validity.

User evaluation

Multiple methods were used to evaluate the assessment system with the users (trainees, coaches, training managers). The findings have shown that the system is practically usable and has certainly improved learning processes. The tools, procedures and assessment reports are clear.

Conclusions and further research

The purpose of designing an assessment system for LVNL that makes training more efficient and effective was achieved. Learning processes are better supported and pass-fail decisions are more reliable and valid. The majority of the requirements and constraints in the PoR were fulfilled. Unfortunately, we do not have convincing evidence for an increased outcome of competent controllers from training yet.

More generally, we gained insight into assessment design for complex skill acquisition, based on the application in the ATC domain. However, a limitation of this thesis is that the findings are only based on a small sample. Some issues need further research. More research on the design of assessment systems for complex skill acquisition in simulator and on-the-job training should be done with special attention for the reliability and validity of human assessors' ratings. Furthermore, more insight into complex skill acquisition is needed to achieve that assessment can support learning optimally. Our new method of making learning curves may contribute to the field of modeling learning processes. In addition, trainability of competences in ATC and in related domains should be further examined. Its relationship with innate cognitive abilities and personality characteristics should be investigated in order to improve the predictive validity of selection and training. Moreover, further research is required on the relationship between adaptive training and deficiencies in competences. We experienced many difficulties with adapting training to the trainee's needs, while this seems to be a key success factor in completing the training. Possible interventions are: (dynamic) task selection, specific coaching, remedial teaching and counseling, alternative training trajectories, lengthening training, retraining, etcetera. Finally, advanced technology and automated measurements can help to improve assessment systems.

Samenvatting

Zoals andere proces control taken in transport (luchtvaart, scheepvaart, spoorwegen) of procesindustrie (bijv. chemische fabrieken en kerncentrales), wordt ook luchtverkeersleiding beschouwd als een complexe cognitieve taak. De leerbaarheid blijkt beperkt te zijn. De strenge eisen aan de veiligheid maken de vereiste prestaties nog hoger omdat elke fout, veroorzaakt door incompetentie van de operators, voorkomen moet worden. Ten gevolge hiervan is het aantal mensen dat de opleiding haalt vaak te laag, ondanks een strenge selectie. Dit kan leiden tot een tekort aan vakbewaam personeel. Lage slagingspercentages zijn bovendien onwenselijk omdat de opleiding doorgaans erg veel tijd en geld kost. Simulators van hoge kwaliteit worden gebruikt om leerlingen op te leiden tot het hoogst haalbare niveau, voordat ze de echte werkomgeving betreden waar de veiligheid kritisch is.

Een goed ontworpen beoordelingssysteem kan bijdragen tot de oplossing van dit probleem in meerdere opzichten. Ten eerste, beoordeling kan leerprocessen ondersteunen door middel van feedback en het aanpassen van de opleiding aan de behoeftes van individuele leerlingen. Dit verhoogt de kans dat de leerlingen de opleiding succesvol afronden. Ten tweede, zo'n beoordelingssysteem maakt het mogelijk om betrouwbare en valide pass-fail beslissingen te nemen. Het aantal 'false positives' moet zo beperkt mogelijk blijven, d.w.z. kansloze leerlingen die onnodig dure opleidingsplekken bezet houden tijdens latere fasen in de opleiding. Dit geldt ook voor het aantal 'false negatives', d.w.z. gezakte leerlingen die de opleiding uiteindelijk wel hadden kunnen halen. Ten derde, de predictieve validiteit van de selectie kan verbeterd worden door betrouwbare criteria uit de opleiding te gebruiken voor validatieonderzoek.

Deze studie is het resultaat van een project dat intern is uitgevoerd bij Luchtverkeersleiding Nederland (LVNL). Het doel was om een beoordelingssysteem te ontwerpen dat het aanleren van de competenties voor luchtverkeersleiding optimaal ondersteunt in de simulatoropleiding en in de werkplekopleiding.

Ontwerpmethodologie

We hebben twee ontwerpmethodologieën gecombineerd die hoofdzakelijk gebaseerd zijn op de ontwerpcyclus van Roozenburg en Eekels (1991) en de methodologie voor opleidingsontwerp van Moonen (2000). Dit heeft geleid tot een cyclisch ontwerpproces, bestaande uit drie fasen: (1) *analyse*; (2) *ontwerp en implementatie*; (3) *evaluatie*. De dissertatie is onderverdeeld in drie delen die overeenkomen met deze drie fasen. De eerste fase, *analyse*, bestaat uit literatuuronderzoek, een competen-

tieanalyse van luchtverkeersleiding, en het opstellen van een Programma van Eisen (PvE) samen met stakeholders. De tweede fase, *ontwerp en implementatie*, betreft het ontwerp en de ontwikkeling van het beoordelingssysteem zelf, en de introductie van het systeem in de opleiding. De derde fase, *evaluatie*, bevat de evaluatie van het beoordelingssysteem in drie delen: (1) psychometrische kwaliteit; (2) leerprocessen (analyse van leercurves en competentieontwikkeling); (3) gebruikersevaluatie. Tenslotte is een algehele evaluatie ten opzichte van het Programma van Eisen gehouden.

We hebben ondervonden dat de ontwerpfasen iteratief zijn. Een sleutel tot succes is de intensieve samenwerking met de gebruikers (luchtverkeersleiders) geweest in het ontwerpen van een innovatief product dat volledig geaccepteerd is voor toepassing in de praktijk.

Analyse

We zijn gestart met literatuuronderzoek, bestaande uit: (1) algemene theorieën over prestaties, competenties en leerprocessen; (2) domeinspecifieke literatuur over de aard van de taak van luchtverkeersleiders en over leerprocessen in luchtverkeersleiding; (3) relevante literatuur over beoordeling die toepasbaar is in de context van luchtverkeersleiding. De bevindingen zijn gebruikt in het ontwerp van het nieuwe beoordelingssysteem.

Vervolgens hebben we een competentieanalyse uitgevoerd door middel van het organiseren van twee workshops. Luchtverkeersleiders namen hieraan deel. Zij waren de inhoudsdeskundigen die hun impliciete kennis expliciet hebben gemaakt. Dertien competenties, elk ondersteund met een set van gedragscriteria, werden geformuleerd in hun eigen jargon. Dit zou de herkenbaarheid en algemeen begrip moeten stimuleren ten behoeve van het praktisch gebruik. De volgende stap was het operationaliseren van de competenties. We hebben de lijst van competenties met gepubliceerde (cognitieve) taakanalyses vergeleken om te onderzoeken of we wellicht aspecten vergeten waren. Dit proces leidde tot het zgn. 'ATC Performance Model'. Dit model heeft gediend als algemeen kader voor het ontwerp van het beoordelingssysteem bij LVNL (Oprins, Burggraaff & Van Weerdenburg, 2006a).

Tenslotte is er samen met een groep stakeholders een Programma van Eisen (PvE) opgesteld, bestaande uit een set van eisen en randvoorwaarden aan het nieuwe beoordelingssysteem. We hebben de tekortkomingen van het voorafgaande beoordelingssysteem geanalyseerd en de doelen voor het herontwerp afgeleid. Het PvE heeft richting gegeven aan het ontwerp van het nieuwe beoordelingssysteem.

Ontwerp en implementatie

Het ontwerpproces vond plaats in nauwe samenwerking met de gebruikers (managers, coaches en leerlingen) die onderdeel uitmaakten van de ontwerpteams. Dit heeft de implementatie van het systeem bevorderd. We hebben keuzes gemaakt in het ontwerp van het systeem om te kunnen voldoen aan de doelstellingen, en dit heeft geleid tot een aantal specifieke eigenschappen.

Het beoordelingssysteem is competentiegericht. Dit houdt in dat er wordt beoordeeld op competenties: de succesvolle integratie van kennis, vaardigheden en houdingen en hun toepassing in realistische omgevingen. Competenties hebben betrekking op de individuele capaciteiten om effectief te kunnen presteren, zonder dat er aandacht wordt besteed aan omgevingsfactoren die de prestaties kunnen beïnvloeden. Alle aspecten die behoren tot competenties worden beoordeeld (technisch, cognitief, emotioneel, sociaal) om een compleet beeld te krijgen. De competenties zijn direct afgeleid van het ATC Performance Model. Dit heeft geresulteerd in de volgende set van competenties voor ACC: *veiligheid, efficiëntie, uitdrukkingsvaardigheid, luistervaardigheid, coordinatie, omgaan met apparatuur, strip- en labelbehandeling, mentale beeldvorming, aandachtsverdeling, planning, besluitvaardigheid, omgaan met werkdruk, houding, en samenwerking.*

Omdat competenties niet direct zichtbaar zijn, wordt elke competentie ondersteund door een set van *gedragscriteria*, geformuleerd in waarneembaar gedrag, waarbij een zespunts ratingschaal wordt gehanteerd. Tijdens de opleiding wordt beoordeeld op dezelfde competenties om de voortgang van leerlingen op elke competentie te kunnen volgen gedurende een opleidingsperiode, gebaseerd op wisselende taaksituaties en omstandigheden. Het generieke karakter van competenties maakt dit mogelijk; zo is *planning* relevant in de uitvoering van elke taak bij luchtverkeersleiding, zowel bij een eenvoudige als complexe taak. *Progressie* is een belangrijke indicator of een leerling nog aan het leren is, of dat hij/zij een leerplateau heeft bereikt. Mogelijke problemen van leerlingen kunnen in een vroeg stadium worden ontdekt zodat de opleiding zo goed mogelijk kan worden aangepast aan de behoeftes van de leerling.

De competenties worden beoordeeld ten opzichte van oplopende *standaarden* tijdens de opleiding. De simulatoropleiding (preOJT) en werkplekopleiding (OJT) zijn opgebouwd uit fasen die elk specifieke standaarden hebben (vgl. 'normering'). De standaarden zijn geformuleerd in voorbeeldgedrag, een variant op de geankerde gedragsschalen (vgl. 'BARS'). Ze geven niet de schaalankers weer, maar het niveau dat bereikt moet zijn aan het eind van elke fase. Op deze manier komen de beoordelaars makkelijker tot overeenstemming over wat er van leerlingen wordt verwacht in tussenliggende fasen. Voor leerlingen is het duidelijker welke competenties zij verder moeten ontwikkelen in een bepaalde fase. De standaarden in de preOJT worden hoofdzakelijk bepaald door de opbouw van de simulatoroefeningen. De indeling in fasen is voornamelijk gebaseerd op twee variabelen: *verkeersafhandeling* (mate van veiligheid en efficiëntie) en *verkeerscomplexiteit* (mate van de moeilijkheid van de taak). Het structureren van de OJT is lastiger vanwege het voortdurende echte verkeer, en was daarom het meest innovatief. De indeling in fasen voor de OJT, vier fasen in totaal, is voornamelijk gebaseerd op drie principes: *verkeersafhandeling, verkeerscomplexiteit, en hulp door de coach*. De lengte van elke fase is flexibel en hangt af van de voortgang van de leerling.

Continuous assessment wordt toegepast om een compleet beeld te krijgen van de prestaties van de leerling en om progressie over een bepaalde periode te kunnen meten. Coaches zijn voortdurend in interactie met de leerlingen en kunnen hen stimuleren om hun gedachten te verwoorden. Het beoordelen van cognitieve processen is nodig om diagnostische informatie te verkrijgen over de prestaties en tekort-

komingen in de prestaties in het bijzonder. *Progressierapporten* worden elke week of elke twee weken ingevuld tijdens de opleiding in de preOJT en OJT. De tijdsintervallen zijn verschillend vanwege de diversiteit in de operationele roosters. De coach heeft een dubbelrol als coach en beoordelaar, en meerdere beoordelaars worden ingezet. In de preOJT worden bovendien *simulator tests* gebruikt om de prestaties van de leerling op een objectieve manier te meten, op een vastgesteld moment en in uitgebalanceerde testscenarios zonder tussenkomst van de coach. Pass-fail beslissingen worden genomen aan het eind van elke fase in de preOJT en OJT. Er bestaat geen vaste afgestegrens, maar de beslissingen worden gebaseerd op een uitgebreide kwantitatieve en kwalitatieve analyse van simulator tests en progressierapporten.

Een web-based leerlingvolgsysteem is ontworpen om progressierapporten in te vullen, om de resultaten te kunnen opslaan in een database, en om diverse rapporten te kunnen genereren van de prestaties van leerlingen. Op deze manier kunnen belanghebbenden de voortgang van leerlingen beter volgen vanuit allerlei lokaties, mits zij toegang hebben tot het systeem. De opleidingsresultaten kunnen goed worden gebruikt voor betrouwbaarheids- en validiteitsonderzoek. De eerste versie is gemaakt in het toetssysteem Questionmark Perception. Vanwege technische problemen is echter een nieuw project gestart voor het ontwikkelen van een nieuw leerlingvolgsysteem.

Er is veel energie gestoken in de implementatie van het beoordelingssysteem, omdat dit als erg belangrijk werd beschouwd voor correct gebruik en acceptatie. De nauwe samenwerking met de gebruikers zorgde ervoor dat het systeem vrij gemakkelijk werd geaccepteerd. Bovendien hebben we het systeem zorgvuldig geïntroduceerd door te beginnen met kleine pilots en door het geven van presentaties. We hebben extra aandacht besteed aan beoordeling in de training van coaches en beoordelaars. Tussentijdse evaluaties hebben geleid tot kleine aanpassingen voordat de uiteindelijke versie werd geïmplementeerd.

Evaluatie

De eisen en randvoorwaarden in het PvE hebben geleid tot de evaluatie van het beoordelingssysteem, bestaande uit drie onderdelen. De data zijn van 34 leerlingen in ACC preOJT (188 progressierapporten; 36 simulator tests), en van 27 leerlingen in ACC OJT (407 progressierapporten), verzameld gedurende de periode januari 2001 tot en met december 2006.

Evaluatie van de psychometrische kwaliteit

We hebben de volgende zaken met betrekking tot betrouwbaarheid onderzocht: (1) *interbeoordelaarsbetrouwbaarheid*; (2) *beoordelingsfouten* (halo-effect, toegeeflijkheids- en strengheidsfout, range beperking en centrale tendentiefout); (3) *test betrouwbaarheid* (interne consistentie, split-half betrouwbaarheid). Predictieve validiteit was onderdeel van de evaluatie van leerprocessen (zie volgende paragraaf) omdat hiervoor analyses over de tijd nodig zijn.

De *interbeoordelaarsbetrouwbaarheid* is onderzocht voor simulator tests waarbij twee beoordelaars onafhankelijk van elkaar een beoordeling invullen. Hun ratings moeten uitwisselbaar zijn omdat beoordelingen niet afhankelijk mogen zijn van de specifieke persoon die beoordeelt. We hebben drie typen indicatoren bekeken: vorm, dispersie, en niveau. De resultaten hebben aangetoond dat de interbeoordelaarsbetrouwbaarheid tussen de beoordelaars voldoende hoog is voor het algemene prestatieniveau (gewogen som van ratings op de competenties), maar slechts middelmatig voor de gelijkheid van profielen (vorm, dispersie). Dit houdt in dat sommige beoordelaars lage ratings geven op sommige competenties, terwijl anderen hoge ratings geven op dezelfde competenties en omgekeerd. Beoordelaars vinden het blijkbaar moeilijk om specifieke problemen bij leerlingen aan te wijzen.

Vervolgens is de aanwezigheid van de meest voorkomende beoordelingsfouten onderzocht. Analyses van de *toegeeflijkheids-* en *strengheidsfout* hebben laten zien dat beoordelaars de neiging hebben tot toegeeflijkheid zoals verwacht: de gemiddelde ratings op de competenties zijn hoger dan het middelste schaalanker. We vonden een paar systematische verschillen in toegeeflijkheid tussen beoordelaars: er bestaan zgn. 'Santa Claus' (hoge gemiddelden) en 'Axeman' (lage gemiddelden) beoordelaars. Bovendien vertonen de meeste beoordelaars *range beperking*. Ze verdelen de ratings niet gelijkmatig over de zespuntsschaal, maar ze geven vaker positieve ratings en vermijden de extreem lage en hoge waarden op de schaal; de *centrale tendentiefout* komt dan ook niet voor. Verschillen tussen beoordelaars in hun verdeling van ratings, uitgedrukt in congruentie-indexen, werden niet gevonden. Tenslotte onderzochten we de aanwezigheid van *halo-effecten*. We vonden hoge intercorrelaties tussen ratings op sommige competenties. We herkenden de indeling van het ATC Performance Model in de resultaten. Specifieke competenties moeten hoge intercorrelaties hebben, omdat ze conceptueel met elkaar samenhangen. Daarom hebben we geconcludeerd dat de aanwezigheid van halo-effecten geen serieus probleem is. In het algemeen werden beoordelingsfouten vaker gevonden in progressierapporten. Dit bevestigt dat simulator tests betrouwbaarder zijn.

Tenslotte hebben we de *interne consistentie* en *split-half betrouwbaarheid* onderzocht. Berekeningen van Cronbach's alpha's en item-total correlaties per competentie hebben laten zien dat de interne consistentie erg hoog is. Slechts een paar criteria moeten verwijderd of aangepast worden. Split-half betrouwbaarheid werd geschat voor simulator tests met behulp van de Spearman-Brown formule. We vonden erg hoge betrouwbaarheidscoëfficiënten. De test betrouwbaarheid is dus voldoende.

De bevindingen leiden tot de conclusie dat de betrouwbaarheid van het beoordelingssysteem voldoende is, hoewel tot op zekere hoogte. Het systeem is goed ontworpen voor wat betreft de indeling in competenties, criteria en standaarden. Echter, de rol van de beoordelaar kan verbeterd worden. We zijn bezig met het organiseren van een training in beoordelen. Hierin wordt aandacht besteed aan het voorkomen van beoordelingsfouten en aan een meer uniform begrip van de competenties.

Evaluatie van leerprocessen

Een goed ontworpen beoordelingssysteem werkt voldoende als de beoordelingsresultaten een juiste afspiegeling zijn van leerprocessen. Patronen en individuele verschillen in leren (bijv. slow starters, leerplateau's) en in prestaties (sterke en zwakke punten) moeten duidelijk onderscheiden worden om te kunnen dienen als basis voor adequate feedback en interventies. Leercurves kunnen worden afgeleid van de beoordelingsresultaten, gebaseerd op een opeenvolging van prestatiemetingen over de tijd. Deze leercurves moeten voldoende representatief zijn voor leerprocessen. In dat geval kan de opleiding maximaal adaptief gemaakt worden aan de behoeftes van de leerling. Pass-fail beslissingen zullen meer valide zijn als ze gebaseerd zijn op de voorspelbaarheid van patronen in leerprocessen.

Leercurves worden normaal gesproken gepresenteerd als groeicurves op basis van herhaalde uitvoering van dezelfde taak op opeenvolgende momenten in de tijd, maar dit is niet mogelijk met ons beoordelingssysteem. De prestaties van de leerling worden beoordeeld ten opzichte van oplopende standaarden die continu worden vertaald naar dezelfde zespuntsschaal. Daarom kunnen de leercurves, geproduceerd door ons beoordelingssysteem, beter gecalibreerde leercurves worden genoemd. We hebben de leercurves, die afgeleid zijn van de progressierapporten in de preOJT en OJT, vergeleken met prototypische leercurves zoals deze worden verwacht vanuit algemene leertheorie. We hebben drie groepen gedefinieerd: (1) *goede leerlingen* (geslaagd zonder problemen); (2) *matige leerlingen* (geslaagd met problemen); (3) *slechte leerlingen* (gezakt). Drie training managers hebben de leerlingen ingedeeld in de drie groepen, gebaseerd op expert judgment dat diende als extern criterium voor succes in de opleiding. We hebben twee kwantitatieve variabelen gedefinieerd voor de leercurves, onderverdeeld in vier metingen: (1) *prestatie*: gemiddeld prestatieniveau; aantal keer dat de prestaties onvoldoende waren; (2) *progressie*: groei; mate van groei.

We hebben het onderscheid tussen de drie groepen visueel bekeken en ook kwantitatief onderzocht. De resultaten hebben laten zien dat de leercurves van de echte leerlingen voldoende representatief zijn voor leerprocessen, omdat ze voldoende gelijkenis vertonen met de prototypische leercurves. Hoewel het aantal leerlingen nog te laag is om overtuigend bewijs te verkrijgen, hebben we toch verkend wat de voorspelbaarheid was van leercurves over de tijd. De resultaten hebben gesuggereerd dat de leercurves in zekere mate voorspellend zijn voor de volgende fasen. Het is echter nog veel te vroeg om vaste afgestengrenzen op te stellen.

Een goed ontworpen beoordelingssysteem moet niet alleen leerprocessen adequaat weergeven zoals uitgedrukt in de leercurves, maar ook de competentieontwikkeling. Dit is belangrijk voor het ontdekken van mogelijke gebreken hierin. We hebben dezelfde indeling in drie groepen van leerlingen gehanteerd. De bevindingen waren in overeenstemming met de literatuur voor wat betreft de leerbaarheid van competenties. Zoals verwacht vonden we meer significante verschillen tussen individuele leerlingen voor competenties die kritischer en minder leerbaar zijn, zoals mentale beeldvorming en omgaan met werkdruk, dan voor competenties die minder kritisch en meer leerbaar zijn, zoals labelbehandeling en omgaan met apparatuur. De belangrijkste redenen voor af testen zijn respectievelijk: mentale beeldvorming, omgaan met werkdruk, aandachtsverdeling, en besluitvaardigheid.

We concluderen dat het beoordelingssysteem individuele verschillen in leerprocessen voldoende weergeeft. Meer longitudinaal onderzoek is nodig voor het vinden van bewijzen voor predictieve validiteit.

Gebruikersevaluatie

Meerdere methoden werden gebruikt om het beoordelingssysteem te evalueren met de gebruikers (leerlingen, coaches, training managers). De resultaten hebben aangetoond dat het systeem praktisch bruikbaar is en dat het zeker heeft geleid tot een verbetering van leerprocessen. De middelen, procedures en beoordelingsformulieren zijn duidelijk.

Conclusies en verder onderzoek

De doelstelling om een beoordelingssysteem voor LVNL te ontwerpen dat de opleiding meer efficiënt en effectief maakt is bereikt. Leerprocessen worden beter ondersteund, en pass-fail beslissingen zijn betrouwbaarder en meer valide. Er is voldaan aan het merendeel van eisen en randvoorwaarden van het PvE. We hebben helaas nog geen definitief bewijs dat de slagingspercentages in de opleiding omhoog zijn gegaan.

In het algemeen hebben we meer inzicht verkregen in het ontwerp van beoordelingssystemen voor het aanleren van complexe vaardigheden, gericht op de toepassing in het domein van de luchtverkeersleiding. Een beperking van deze dissertatie is echter dat de resultaten slechts op een klein sample gebaseerd zijn. Sommige zaken vereisen verder onderzoek. Meer onderzoek is nodig naar het ontwerp van beoordelingssystemen voor het aanleren van complexe vaardigheden in simulator opleidingen en werkplekopleidingen, waarbij speciale aandacht geschonken moet worden aan de betrouwbaarheid en validiteit van de ratings van menselijke beoordelaars. Bovendien is meer inzicht nodig in het aanleren van complexe vaardigheden om te bewerkstelligen dat beoordelingen optimale ondersteuning kunnen bieden aan leerprocessen. Onze nieuwe methode om leercurves af te leiden kan een bijdrage leveren aan het onderzoek naar het modelleren van leerprocessen. Daarnaast moet de mate van leerbaarheid van competenties in luchtverkeersleiding en andere domeinen verder onderzocht worden. De relatie met aangeboren cognitieve capaciteiten en persoonlijkheidskenmerken verdient extra aandacht ter verbetering van de predictieve validiteit van de selectie en opleiding. Ook is verder onderzoek vereist naar de relatie tussen adaptief opleiden en mogelijke gebreken in competenties. We hebben veel moeilijkheden ervaren in het adaptief maken van de opleiding aan de behoeftes van leerlingen, terwijl dit een sleutel tot succes kan zijn voor het succesvol afronden van de opleiding. Mogelijke interventies zijn: (dynamische) taakselectie, specifieke coaching, remedial teaching en counseling, alternatieve opleidingstrajecten, verlenging van de opleiding, heropleiden, etcetera. Tenslotte kunnen meer geavanceerde technologie en geautomatiseerde metingen bijdragen tot een verbetering van beoordelingssystemen.

Appendix A: ATC work and training at LVNL

1. Air Traffic Control The Netherlands (LVNL)

LVNL is an Air Navigation Service Provider (ANSP) responsible for providing air traffic services to civil air traffic in the Amsterdam Flight Information Region (FIR) (www.lvn.nl). The Amsterdam FIR stretches over the Dutch territory, including a large part of the North Sea. Air traffic services consist of three tasks: air traffic control (control of air traffic by issuing clearances and directions to pilots of aircraft), flight information service (providing information during flight about weather, navigation aids and airport conditions), and alerting service (warning and assisting authorities concerned on aircraft needing assistance in the form of search and rescue). Air traffic management is regulated by world-wide procedures (ICAO, 2004), European regulations (e.g., ESARR, 2000), and Dutch law. Schiphol Airport, the largest of four airports managed by LVNL, is considered to be a complex airport due to the large number of aircraft to be handled in peak hours (around 120 aircraft an hour) during multi runway operations.

There are five main departure and arrival runways. The apron and runways are controlled from the main tower at Schiphol Airport, and a second tower is available for the runway called 'Polderbaan' (18R-36L) presented at the far left of Figure A1.

Figure A1. Five runways available at Schiphol Airport

From each runway, fixed departure routes called Standard Instrument Departures (SID's) link in five directions to the international Air Traffic Service route system. Figure A2 presents these SID's in the five directions. There also exist Standard Arrival Routes (STAR's) exiting this ATS route system for airports like Schiphol (Figure A3). These routes are located in three Control Areas (CTA) of five sectors, controlled by the Area Control Center (ACC) at LVNL.

Figure A2. Airways and SID's in the Dutch airspace

Figure A3. STARS for Schiphol Airport

The organization of Air Traffic Management at LVNL divides air traffic control into two main groups of ATC functions (restricted to the controller functions that require a rating):

- *Tower and Approach control (Schiphol TWR/APP):*
 - Start-up controller (SUC): responsible for start-up control at the gate
 - Ground controller (GND): responsible for air traffic control at the ground
 - Tower controller (TWR): responsible for departures and arrivals at the runway
 - Approach (and Arrival) controller (APP, ARR): responsible for departures to the air routes and arrivals from the air routes to the vicinity of Schiphol²⁰
- *Area Control Center (Amsterdam ACC)*
 - Area controller (ACC): responsible for traffic on the air routes, departing from, arriving in, and flying over the Netherlands, divided into three sectors: South, West, East.

Apart from these licensed functions, there also exist other (related) operational functions such as assistant controllers (VLA), and Flight Information Officers (FISO).

The SUC, GND and TWR controllers work from the two towers at Schiphol Airport, the other controllers are radar controllers working from the operational control room at LVNL. Air traffic is continuously moving in different parts of the airspace, controlled by the different controllers, dependent on the flight level of the aircraft.

²⁰ ESARR (2000) prescribes different terminology for the formal licenses (e.g., ADI for GND and TWR; APS for ARR and APP, and ACS for ACC; see for the details: ESARR, 2000). However, in the practical organization of LVNL and therefore in this dissertation we use the terminology as explained in this Appendix.

Figure A4. Tower at Schiphol Airport.

Figure A5 presents the Dutch airspace in a vertical way:

Luchtruimte structuur verticaal

Module IN-OPS, dia 6

Figure A5. Vertical Dutch airspace

Figure A5 shows clearly how the Dutch airspace is divided into different parts, based on altitudes. The TWR controller controls all the traffic in the Control Zone (CTR) up to 3000 feet (and the GND controller at the ground), while the APP controller is responsible for traffic handling in the Terminal Area (TMA) till 6500/9500 feet. The ACC controllers control the aircraft in the (upper) Control Areas (CTA) till 19500 feet and above in the upper CTA up to 24500 feet (flightlevel 245). Then Eurocontrol (Maastricht UAC center) takes it over from ACC above this flightlevel. LVNL collaborates also with other adjacent centers such as foreign ATM providers (Germany, Belgium) and military ATC (Royal Netherlands Airforce).

Area control in the Amsterdam FIR is an extension of departure and approach control and therefore typically more terminal control than area control. Schiphol Airport has about 1100 take-offs and landings per day and because of the hub-and-spoke function of the airport we daily have several peaks of departing and arriving flights. The main purpose of Air Traffic control is to avoid collisions and to expedite and maintain an orderly flow of traffic. Amsterdam ACC is streaming the flow of departures and arrivals, using speed control and vectoring technique to achieve efficient and safe traffic handling to and from the FIR boundary points.

2. The LVNL training program

The internal training program at LVNL starts with Initial Training, aimed at the acquisition of generic prerequisite theory and basic skills of air traffic control, prescribed by European requirements of Eurocontrol (ESARR, 2000). Theory lessons are combined with skill training in 'basic simulators' that represent a fictive airspace and airport in a simplified air traffic control environment. Training content is generic for ATC in Europe and is designated as Common Core Content (CCC). The Initial Training period takes one year. Around 30 trainees a year start with Initial Training. A selective module, AAPRO, during Initial Training (after half a year) selects around 50% of the candidates for entering Unit Training.

Initial Training is followed by Transitional Training (2 weeks theory lessons) for the selected candidates in which Dutch-specific theory (e.g., the Dutch airspace) is taught. Finally, trainees are assigned to a specific training trajectory within Unit Training: Tower and Approach Control (TWR/APP) at Schiphol Airport, Area Control (ACC) at Schiphol Airport, or a specific regional field (Rotterdam, Beek or Eelde), but regional training does not belong to the scope of this study. The aim is that 16 trainees a year enter Unit Training, equally divided into TWR/APP and ACC, but in reality only a number of around 8 to 12 trainees a year start with Unit Training including some foreign students.

The TWR/APP training program trains for several ratings (cf. controllers' licenses): respectively Start-up Control (SUC), Ground Control (GND), Arrival Control (ARR) and Approach Control (APP) (a combined rating), and Tower Control (TWR). Trainees lose the first two ratings (SUC, GND) when they get their APP rating. The ACC training program only trains for one rating but for three sector clusters. The ACC Unit Training program starts with a training for assistant controller (VLA), followed by a internship at the Flight Information Center (FIC). Next, the trainee is submitted to ACC training for the first sector cluster (e.g., South sectors) and finally, after completing this training, an additional training is followed for the other two sector clusters (e.g., West, East sectors). The scope of this study is Unit Training for air traffic controllers: TWR/APP training (GND, ARR, APP and TWR) and ACC training. SUC, VLA and FIC are not included because these are too different from the real ATC functions. ACC (first sector cluster) is the main case in this study.

The duration of Unit Training depends on waiting times to go on with the next rating training; a period of two to four years is common to succeed the whole training program. Training for each rating starts with preOJT, which generally consists of function-specific theory lessons and simulator training. The aim of the preOJT is preparation for the OJT (on-the-job training) that follows preOJT. The main structure of the training program, focused on Unit Training, is presented in Figure A6²¹:

²¹ Since 2005, the simulator training (preOJT) for Arrival and Approach Control is combined, called the AApproach and ARrival Training (APART). However, the majority of the data used in this study are from this training structure.

Figure A6. Overview of the LVNL training program with the focus on Unit Training at Schiphol

Appendix B: Assessment reports ACC

ACC simulator test report preOJT

ACC progression report preOJT

ACC progression report OJT

Executive- & Planning Controller ACC

PreOJT Simulator

FINAL TEST

Trainee: _____

Final Test ACC (preOJT)

ASSESSMENT ITEMS SAFETY		(MAX 150)
V1 Conflicts solved adhoc but according to the correct separation minima	(- 5 a time)	(max 30)
V2 Conflicts not solved according to the correct separation minima	(- 10 a time)	(max 40)
V3 Conflicts not acknowledged/not seen	(- 20 a time)	(max 60)
V4 Traffic not transferred to adjacent ATC unit correctly and/or not in time	(- 5 a time)	(max 20)

ASSESSMENT ITEMS ACTIONS		(MAX 150)
1 Efficiency / optimum sequence		(max 65)
a) Applies vector technique correctly		
b) Applies speedcontrol correctly		
c) Takes into account differences in height between aircraft		
d) Takes into account (differences in) aircraft performances		
e) Minimises his/her own workload		
f) Realises an optimum sequence of aircraft		
g) Applies continuous climb/descent		
h) Takes into account winds aloft		
2 Verbal expression		(max 20)
a) Expresses him/herself well in '(non)-standard' phraseology		
b) Expresses him/herself concisely, to the point, unambiguously and firmly		
c) Has a clear, quiet pronunciation and good intonation		
3 Listening		(max 20)
a) Listens out well, understands messages well		
b) Interpretes messages well and reacts adequately		
c) Is easily approachable for others at the sector		
4 Coordination		(max 30)
a) Coordinates in time and with the appropriate ATC function/unit		
b) Communicates his/her plan concisely and to the point		
c) Is able to make arrangements and acknowledges these correctly		
d) Gives and receives radar handovers correctly		
5 Application of valid procedures / apparatus		(max 5)
a) Applies valid procedures and working agreements correctly		
b) Makes correct use of apparatus		
6 Strip- and label handling		(max 10)
a) Updates labels correctly		
b) Works with strips correctly		

ASSESSMENT ITEMS THINKING PROCESS		(MAX 300)
7 Mental picture / Perception		(max 80)
a) Keeps a good overview of the situation by scanning regularly		
b) Looks, observes and takes action if necessary		
c) Checks available information to be correct		
d) Guards the identification process of the label presentation		
e) Anticipates on future and variable traffic situations		
8 Attention management		
a) Can divide attention between several situations sufficiently		
b) Can perform several actions simultaneously		
c) Can park information in his/her memory without forgetting it		
9 Planning		(max 70)
a) Is creative in inventing various solutions		
b) Can plan according to valid procedures and agreements		
c) Is flexibel in adjusting plans		
10 Decision making		(max 60)
a) Takes the initiative and acts		
b) Shows confidence and takes the lead		
c) Acknowledges priorities correctly		
11 Workload management		(max 50)
a) Adapts work tempo to traffic load		
b) Stays calm, also during hectic moments

Final Test ACC (preOJT)

Trainee : _____ Date: _____

Assessors : _____

Norm final score: 70%

Overall impression assessors (max 100)

ASSESSMENT ITEM SAFETY	SCORE
Safety (max 150)	

ASSESSMENT ITEMS ACTIONS	INSUFFICIENT:			SUFFICIENT:			SCORE
	0	2	4	6	8	10	
1 Efficiency / optimum sequence							(x6,5)
2 Verbal expression							(x2)
3 Listening							(x2)
4 Coordination							(x3)
5 Application of valid procedures / apparatus							(x0,5)
6 Strip- and label handling							(x1)
Subtotal (max 150)							

ASSESSMENT ITEMS THINKING PROCESS	INSUFFICIENT:			SUFFICIENT:			SCORE
	0	2	4	6	8	10	
7 Mental picture / Perception							(x8)
8 Attention Management							(x4)
9 Planning							(x7)
10 Decision making							(x6)
11 Workload management							(x5)
Subtotal (max 300)							

Total:

Final score (divided by 7):

Remarks:

Signature assessor(s):

Signature trainee:

Progression report PreOJT Simulator Executive- & Planning Controller ACC

Trainee :
Date :
Assessor :

Competences	Insufficient				Sufficient				Remark
1 Safety									
Applies separation minima correctly									
Switches from monitoring to vectoring in time									
Builds in safety buffers sufficiently									
Transfers traffic to the adjacent ATC unit correctly and in time									
2 Efficiency									
Applies vector technique correctly									
Applies speed control correctly									
Takes into account differences in height between aircraft									
Takes into account differences in aircraft performances									
Minimises his/her own workload as much as possible									
Creates an optimal sequence of descending and climbing traffic streams									
3 Verbal expression									
Expresses him/herself well in 'non-standard' phraseology									
Expresses him/herself concisely, to the point, unambiguously and firmly									
Has a clear, quiet pronunciation and intonation									
Expresses him/herself in the Dutch language conform ICAO level 2									
4 Listening									
Listens out well, understands messages well									
Interprets messages well and reacts adequately									
Is easily approachable for others at the sector									
Has an understanding of the Dutch language conform ICAO level 2									
5 Co-ordination									
Co-ordinates in time and with the appropriate ATC function/unit									
Communicates his/her plan concisely and to the point									
Makes clear arrangements and acknowledges these correctly									
Gives and receives radar handovers correctly									

Competences	Insufficient				Sufficient				Remark
6 Equipment operation									
Applies valid procedures and working agreements correctly									
Makes correct use of equipment									
7 Strip- en label management									
Updates labels correctly									
Works with strips correctly									
8 Mental picture									
Keeps a clear overview of the situation by scanning regularly									
Looks, observes and takes action if necessary									
Controls the accuracy of available information									
Guards the identification process and the label presentation									
Anticipates future and variable traffic situations									
9 Attention management									
Divides attention between several situations sufficiently									
Performs several actions simultaneously									
Holds information in his/her memory without forgetting it									
10 Planning									
Is creative in inventing various solutions									
Plans according to valid procedures and agreements									
Is flexible in adjusting plans									
11 Decision making									
Takes initiative and acts when necessary									
Shows confidence in taking the lead									
Acknowledges priorities correctly									
12 Workload management									
Adapts work tempo to traffic load optimally									
Stays calm, also during hectic moments									
13 Attitude									
Shows responsibility during work									
Takes his/her training seriously									
Is eager to learn									
14 Team orientation									
Collaborates with others easily									
Is willing to adapt to common standards and values									
Progression									

Remarks:

General impression:		Insufficient	Sufficient
Signature assessor(s):	Signature trainee:

Progression report OJT Executive- & Planning Controller ACC

Trainee :

Total number of weeks in OJT:

Date :

Period:

Assessor :

Training phase:

Competences	Insufficient				Sufficient				Remark
1 Safety									
Applies separation minima correctly									
Switches from monitoring to vectoring in time									
Builds in safety buffers sufficiently									
Transfers traffic to the adjacent ATC unit correctly and in time									
2 Efficiency									
Applies vector technique correctly									
Applies speed control correctly									
Takes into account differences in height between aircraft									
Takes into account differences in aircraft performances									
Minimises his/her own workload as much as possible									
Creates an optimal sequence of descending and climbing traffic streams									
3 Verbal expression									
Expresses him/herself well in 'non-standard' phraseology									
Expresses him/herself concisely, to the point, unambiguously and firmly									
Has a clear, quiet pronunciation and intonation									
Expresses him/herself in the Dutch language conform ICAO level 2									
4 Listening									
Listens out well, understands messages well									
Interprets messages well and reacts adequately									
Is easily approachable for others at the sector									
Has an understanding of the Dutch language conform ICAO level 2									
5 Co-ordination									
Co-ordinates in time and with the appropriate ATC function/unit									
Communicates his/her plan concisely and to the point									
Makes clear arrangements and acknowledges these correctly									
Gives and receives radar handovers correctly									

Competences	Insufficient				Sufficient				Remark
6 Equipment operation									
Applies valid procedures and working agreements correctly									
Makes correct use of equipment									
7 Strip- en label management									
Updates labels correctly									
Works with strips correctly									
8 Mental picture									
Keeps a clear overview of the situation by scanning regularly									
Looks, observes and takes action if necessary									
Controls the accuracy of available information									
Guards the identification process and the label presentation									
Anticipates future and variable traffic situations									
9 Attention management									
Divides attention between several situations sufficiently									
Performs several actions simultaneously									
Holds information in his/her memory without forgetting it									
10 Planning									
Is creative in inventing various solutions									
Plans according to valid procedures and agreements									
Is flexible in adjusting plans									
11 Decision making									
Takes initiative and acts when necessary									
Shows confidence in taking the lead									
Acknowledges priorities correctly									
12 Workload management									
Adapts work tempo to traffic load optimally									
Stays calm, also during hectic moments									
13 Attitude									
Shows responsibility during work									
Takes his/her training seriously									
Is eager to learn									
14 Team orientation									
Collaborates with others easily									
Is willing to adapt to common standards and values									
Progression									

Remarks:

General impression:		Insufficient	Sufficient
Signature assessor(s):	Signature trainee:		

Appendix C: Data structure

We have different data for the two training trajectories, TWR/APP and ACC. We mainly used the data for ACC, collected in the period January, 2001 till July, 2006. We only used some data of TWR/APP, that is, simulator tests (chapter 9), because the number of ACC simulator tests was too small. The data reveal a hierarchical or multilevel structure (Hox, 2002). That is, there are data on courses, trainees taking parts in those courses, and assessments of those trainees. Thus, there are three levels: 1) *assessment*, 2) *trainee* and 3) *course level*. The multilevel structure appears to be 'cross-classified' (Hox, 2002), because not all trainees follow the same courses; see Figure C1:

Figure C1. Data structure

Table C1 describes the variables and their values for each level in general terms. Table C1 shows that at the highest level, i.e. the *course level* there are two training trajectories, TWR/APP and ACC and TWR/APP. Each of them has a preOJT (simulator training) and OJT training part, while OJT contains two learning phases. The TWR/APP trajectory has four courses, each with a preOJT and an OJT phase, respectively: GND, ARR, APP and TWR. The ACC trajectory only consists of the ACC course. At the *trainee level* there is a pass/fail variable: each trainee can pass or fail in a course as well as in the training part of the course, for instance, preOJT-GND or OJT-ACC. At the *assessment level* there are two types of reports – progression reports and simulator test reports – each containing certain data. Every report is filled in at a specific moment of time (date) by one or more assessors, but with irregular intervals for each trainee. For most of the assessments the assessors are different persons. Each report comprises a set of performance criteria, rated on a 6-point scale. A set of criteria represents a competence. Overall competence ratings were

obtained by averaging the ratings on the performance criteria related to each competence (equal weights). Finally, each report has a weighted final score, indicating the overall performance level, based on a weighting of the competences (see chapter 7 for ACC). The same set of competences has been rated in preOJT and OJT, and for TWR/APP and ACC, but the underlying performance criteria are different in most cases (see chapter 7). In progression reports all performance criteria are rated, while in some simulator tests (only ACC and GND, not ARR and APP) only the competences are rated. Since we mainly used data from ACC, Table C2 presents the various numbers.

Table C1. Levels, variables and values in the data set

Levels	Variables	Values
Course level	Training trajectory	TWR/APP or ACC trajectory
	Training part	PreOJT or OJT (OJT: consists of two learning phases)
	Course	TWR/APP-trajectory: respectively GND, ARR, APP, TWR ACC-trajectory: ACC
Trainee level	Trainee	Names
	Pass/fail	Pass or fail (in each course, preOJT or OJT)
Assessment level	Type of report	Progression reports (preOJT and OJT) Simulator tests (only in preOJT)
	Moments of time	Dates
	Assessor	Names
	Performance criteria	Ratings at 6-points scale (range 1 to 6)
	Competences	Averaged ratings on sets of criteria that belong to a competence
	Overall performance level	Unweighted sum of competence ratings Weighted sum of competence ratings

Table C2. Various numbers of ACC

Numbers		PreOJT	OJT
Trainees	Passed	25	16
	Failed	9	11
Assessment reports	Simulator tests	36	-
	Progression reports	188	407
Moments of time		1 < > 9	1 < > 32
Assessors		14	39
Competences		12	14

Finally, we may assume that there exist rather high intercorrelations at each level as presented in Table 1. There are two main possibilities to deal with data in a multilevel structure: 1) aggregating data to a higher or lower level (e.g., calculating averages of assessments for a single trainee), and 2) applying multilevel analysis. We choose for the first solution: we are not interested in possible explanatory variables at a higher level that could explain effects on the lower level which is the most common application of multilevel analysis. In addition, many analyses that we need are not possible in a multilevel model but could be done in another way. A multilevel analysis would make it unnecessarily complicated.

List of abbreviations

AAPRO	Approach and Area Procedural module
ACC	Area Control Center
ANSP	Air Navigation Service Provider
APP	Approach Control
ARR	Arrival Control
ASE	Assessor
ATC	Air Traffic Control
ATM	Air Traffic Management
ATS	Air Traffic Service
ATSAT	Air Traffic Selection and Training
BARS	Behavioural Anchored Rating Scale
BES	Behavioural Expectation Scale
BOS	Behavioural Observation Scale
BZO	Low visibility procedures [Beperkt Zicht Omstandigheden]
CBPM	Computer Based Performance Measure
CCC	Common Core Content
CRM	Crew Resource Management
CTA	Control Area
CTR	Control Zone
DAC	Dynamic Air traffic control Test
DATCOSS	Dutch Air Traffic Controller Selection System
DFS	German ATC [Deutsche Flugsicherung]
DLR	German Aerospace Center [Deutsches Zentrum für Luft- und Raumfahrt]
EC	Executive Controller (ACC)
EDD	Electronic Data Display
ESARR	Eurocontrol Safety Regulatory Requirements
EXM	Examiner
FAA	Federal Aviation Administration
FAB EC	Functional Airspace Blocks Europe Central
FIC	Flight Information Center
FIR	Flight Information Region
FISO	Flight Information Service Officer
FL	Flight Level
FOR	Frame-Of-Reference (training)
FT	Feet
GND	Ground Control
HERA	Human Error in European Air Traffic Management
ICAO	International Civil Aviation Organization
ICT	Information Communication Technology
IFR	Instrument Flight Rules
ILS	Instrument Landing System
IT	Initial Training
KLM	Royal Dutch Airlines [Koninklijke Luchtvaart Maatschappij]
LOE	Line Operational Evaluations
LOS	Line Operational Simulation
LOSA	Line Operations Safety Audit
LVNL	Air Traffic Control the Netherlands [Luchtverkeersleiding Nederland]
MCAT	Multiple Controller Aptitude Test
NATS	National Air Traffic Services (United Kingdom)

NAVAIDS	Navigational Aids
NM	Nautical miles
OFS	Operator Functional State
OJT	On-the-job Training
OJTI	On-the-job Training Instructor
OPS	Operational Department
OTS	Over-the-shoulder
PDA	Performance Distribution Assessment
PLC	Planning Controller (ACC)
PoR	Program of Requirements
PreOJT	Preceding On-the-job training
QMP	Questionmark Perception
R/T	Radiotelephony
SA	Situational Awareness
SATSA	Swedish Air Traffic Services Academy
SESAR	Single European Sky ATM Research
SHAPE	Self, Human interaction, Aircraft, Procedures, Environment and situation
SID	Standard Instrument Departure
SIM	Simulator
SME	Subject Matter Expert
SSR	Secondary Surveillance Radar
STAR	Standard Arrival Route
SUC	Start-up Control
TMA	Terminal Control Area
TRM	Team Resource Management
TT	Transitional training
TWR	Tower Control
TWR/APP	Tower and Approach Control (department)
UDP	Uniform Daylight Period
UT	Unit Training
VEM	Safety, efficiency, environment [Veiligheid, Efficiency, Milieu]
VFR	Visual Flight Rules
VLA	Assistant Air Traffic controller
WILSC	Work attitude, Information management, Leadership, Stress management, Cooperation

Glossary of ATM terminology

Adjacent centers	Neighbouring ANSP's
Aircraft performances	Specific properties of aircraft (e.g. maximum speed, rate of climb)
Arrival runway	Runway for landing traffic
Basic circuit	Pattern for flights arriving at the runway before setting on the ILS
Callsign	Flight identification of a specific aircraft
Clearance	ATC instruction provided to pilots
Common Core Content (CCC)	Obligatory training content for IT in Europe (ESARR, 2000)
Conflict	Possible risk of collision between aircraft
Coordination	Communication with adjacent centers or colleague controllers
Departure runway	Runway for departing traffic
Electronic data display (EDD)	Radar screen that displays the situation on the ground
En route control	Area Control
Flight level (FL)	Specified height of aircraft, expressed in hundreds of feet
Flight progress strips	Paper strips with flight information
Heading	Flight direction (in radials)
Holding	Specific area for aircraft to absorb delay
IFR traffic	Flights to be handled by instrument flight rules
Inbound traffic	Traffic approaching to land at the airport
Intersection points	Runway entry positions or crossing waypoints on air routes
Label	Label on the radar screen with flight information
Level separation	Separation by creating a minimum height between aircraft
Outbound traffic	Traffic departing from the airport
Radiotelephony (R/T)	Language used in communications with pilots
Rating	European licence needed for any ATC function (ESARR, 2000)
Readback	Repetition of the clearance or instruction by the pilot
Runway intrusion	Unauthorized entering of a runway by aircraft or vehicle
Sector	A control area in which the airspace for area controllers is divided
Separation minima	Minimum distance and height between aircraft
Sequencing	Realizing an efficient sequence or flow of aircraft
Special sight circumstances	Low visibility caused by weather circumstances e.g. fog
Standard Instrument Departure (SID)	Standard flight route after departure to the main air route system
Standard Arrival Route (STAR)	Standard flight route from the main air route system to the airport
Streaming	Avoiding (unnecessary) delay for air traffic
Taxi routes	Specific routes to taxi from the gate to the runway and vice versa
Vectoring	Providing headings to aircraft for sequence or separation purpose
VFR traffic	Traffic to be flown according visual flight rules
Wake turbulence	Vortex caused by the size of the aircraft
SSR	Radar system for position of aircraft, their identity and altitude
Safety buffer	Safe (intermediate) solution to keep aircraft separated
Pseudo pilot	Pilot played by people in simulators
Procedural control	ATC without aids such as a radar screen, only paper strips available
Overflights	Flights that do not depart or arrive at the airport served by ANSP
Executive Control (EC)	Area controller working at the radar position
Instrument Landing System (ILS)	Landing system providing information to the pilot about inbound course and descent path
Planning Controller (PLC)	Area controller working at the planner position

Index

- abilities, 33, 36, 39, 47, 48, 60, 78, 102, 177, 187, 206, 213, 225
- action regulation, 27
- action theory, 30
- action-orientation, 29, 214
- activation, 28, 29
- air traffic control, 12, 41, 42, 49, 71, 75, 76, 79, 84
- aptitude-treatment interaction, 39, 211
- arousal, 28
- assessment
 - authentic assessment, 60
 - cognitive assessment, 59
 - continuous assessment, 65, 85, 89, **107**, 117
 - diagnostic assessment, 59, 65, 99, 210
 - dynamic assessment, 58, 65, 210
 - formative assessment, 57, 59
 - peer assessment, 58, 118
 - performance assessment, 14, 55, **60**, 65, 98, 137, 150
 - self-assessment, 58, 118
 - summative assessment, 57
- assessment task, 77, 84, 100, 101, 103, 245
- assessor, 57, 63, 64, 66, 77, 84, 85, 87, 107, 112, 115, **117**, 130, 131, 137, 147, 149, 151, 164, 168, 169, 179, 183, 206, 211, 227, 229, 231
- assessor training, 64, 129, **131**, 150, 169
- attention management, 77, 80, 86, 109, 214, 223, 225, 226
- attitude, 32, 57, 75, 98, 103, 168
- authentic learning environment, 32, 35
- automated measurement, 66, 99
- behavior, 27, 29, 34
- behavioral expectation scale, 62
- behavioral markers, 65, 75, 76, 100, 109, 138
- behavioral observation scale, 62
- behaviorally anchored rating scale, **62**
- behaviorally anchored rating scales, 101, 111
- behaviorism, 34, 55
- central tendency error, 64, 115, **151**, 153, 158, 165, 168
- chunking, 38, 45, 102, 214, 227
- coaching, 15, 49, 51, **53**, 84, 99, 211, 227
- cognitive apprenticeship, 53, 106
- cognitive process, 71, 75, 99, 116, 211
- cognitive task analysis, 13, 71, 76, 77
- cognitivism, 35, 56
- compensatory control, 29, 44
- competence, 11, **31**, 32, 33, 34, 35, 49, 51, 52, 55, 58, 59, 60, 67, 68, 71, 76, 84, 88, 97, 98, 100, 103, **109**, 111, 115, 119, 125, 138, 139, 147, 150, 151, 165, 169, 172, 179, 183, 185, 206, 210, 213, 225, 227, 230, 242
- competence acquisition, **34**, 41, 46, 48, 103
- competence analysis, 71, 73, 98
- competence management, 119
- competence-based assessment, 11, 59, **98**, **100**
- complex cognitive skill, 12, 36, 42, 46, 150
- complex skill acquisition, 15, **36**, 39, 46, 177, 210
- consistent task components, **36**, 39, 48, 78, 100, 177, 188, 210, 213
- constructivism, 35, 56, 58
- control, 28, 30, 44
- crew resource management, 65, 137, 138, 140, 151, 153, 166, 169, 172

- criterion-referenced, 58, 164
- critical incidents analysis, 62
- design cycle, 16, 20, 83, 97, 129, 235
- efficiency, 12, 30, 31, 44, 75, 77, 79, 86, 106, 110, 111, 224
 - personal efficiency, 31, 44, 224
 - psychological efficiency, 30
- effort, 28, 30, 44, 52, 176, 227
- emotional-motivational regulation, 27
- examiner, 131
- experimental learning, 49
- expert, 30, 38, 51, 53
- expertise, 31, 50
- feedback, 55, 56, 65, 67, 87, 88, 98, 99, 107, 138, 150, 153, 175, 205, 227, 232, 242
 - extrinsic feedback, 53
 - intrinsic feedback, 53
- fidelity, 49
- frame-of-reference training, 64
- guided discovery learning, 53
- halo error, 64, 117, 152, 153, 160, 166, 169
- human error, 28, 42, 43
- human factors, 28, 72
- human performance analysis, 72
- information processing, 26, 37, 42, 71, 76, 80, 81, 102, 110, 115, 176, 214
 - automated information processing, 26, 37, 46, 51, 176, 213
 - controlled information processing, 26, 37, 46, 176, 213
 - information processing models, 26, 72
- instructional design theory, 19, 37, 51, 55
- internal consistency, 64, 169, 170
- interrater agreement, 64, 115, 137, 139, 147, 149, 242
- Kanfer-Ackerman task, 13, 46, 177, 205
- knowledge, 32, 33, 37, 38, 46, 52, 53, 56, 57, 59, 60, 67, 72, 73, 75, 98, 103, 168, 178, 182, 207, 211
- knowledge management, 34
- knowledge-based, 27, 30, 37, 38, 43, 213
- law of effect, 34
- learnability. *See* trainability
- learning
 - formal learning, 50
 - incidental learning, 50
 - informal learning, 13, 34, 50
 - situated learning, 50
- learning activities, 48, 51, 52, 53
- learning curves, 36, 39, 46, 58, 87, 99, 100, 101, 175, 176, 179, 181, 184, 187, 189, 205, 207, 210, 241
- learning plateau, 100, 102, 175, 176, 182, 206, 211, 242
- learning process, 27, 37, 39, 47, 53, 55, 56, 57, 87, 88, 98, 119, 150, 168, 175, 176, 205, 213, 225, 241
- learning styles, 35, 39, 52, 102, 177, 227
- leniency error, 64, 118, 151, 153, 155, 164, 168
- line checks, 66
- line operational evaluations, 66
- line operational simulators, 66
- line operations safety audit, 66
- memory, 26, 38, 42, 47, 51, 59, 63, 71, 80, 214, 224
- mental model, 44, 47
- mental picture, 44, 72, 77, 80, 113, 214, 223, 225, 226
- non-consistent task components, 39, 48, 78, 100, 177, 206, 210, 213
- non-technical skills, 65, 99
- norm-referenced, 58, 164
- novice, 30, 38, 46, 51
- on-the-job training, 13, 16, 49, 50, 52, 53, 61, 85, 86, 101, 103, 106, 177, 207, 227
- operator, 28, 41, 42, 47
- operator functional state, 27
- over-the-shoulder observation, 65, 99, 107, 153
- part-task training, 52
- perception, 77, 80

- performance, **25**, 27, 28, 34, 36, 37, 42, 52, 53, 57, 58, 59, 61, 63, 64, 65, 67, 71, 75, 76, 85, 86, 87, 88, 99, 120, 138, 149, 150, 151, 164, 175, 176, 179, 181, 184, 187, 205, 207
- performance appraisal, 55, 60, 61
- performance criteria, 68, 75, 77, 84, 85, 87, 100, 103, 107, **109**, 115, 138, 151, 165, 169, 172, 229, 230
- performance distribution assessment, 63
- performance model, 73, 77
- performance standards, 14, 61, 66, 77, 84, 86, 100, 103, 108, **111**, 169, 179, 182, 205, 211, 229, 231, 241
- performance test, 60, 108, 117, 133
- personality, 33, 48, 78, 177, 214
- portfolio, 61
- power law of practice, 36, 37, 176, 206
- problem solving, **38**, 39, 45, 71, 80
- Program of Requirements, 17, **83**, 88, 97, 107, 118, 122, 137, 235
- psychometric quality, 64
- psychophysiological state, **28**, 34, 45, 81, 214
- range restriction, **151**, 153, 158, 165, 168
- rating errors, 133, 149, 150, 164, 168, 242
- rating process, 63
- rating scale, 61, 62, 65, 86, 100, **115**, 152, 179, 205, 231
- reinforcement theory, 35
- reliability, **64**, 85, 88, 115, 119, **137**, 164, **169**, 170, 171, 172, 206, 242
 - alternative form reliability, 64
 - interrater reliability, 64, 138
 - split-half reliability, 64, 173
 - test-retest reliability, 64
 - within-group interrater reliability, 140
- rule-based, **27**, 30, 37, 38, 43, 213
- safety, 12, 31, 41, 44, 49, 75, 77, 79, 86, 106, 110, 115, 139, 224
- selection, 12, 14, 15, 47, 61, 66, 67, 74, 85, 86, 88, 98, 99, 164, 214, 226, 241
- self-efficacy, 48, 81
- self-regulation, 28
- severity error, 64, 118, **151**, 153, 155, 164, 168
- simulator test, 67, 84, 112, 115, 130, 165, 169, 172, 231
- simulator training, 13, 16, **49**, 52, 53, 61, 66, 85, 86, 101, 103, 106, 177, 179, 207, 227
- situated cognition, 35
- situation assessment. *See* situational awareness
- situational assessment. *See* situational awareness
- situational awareness, **43**, **44**, 47, 72, 73, 75, 77, 80, 214, 223, 227
- skill-based, **27**, 37, 43, 213
- skills, 32, 42, 57, 59, 60, 75, 98, 103, 168, 178, 182, 207, 211
 - non-recurrent skills, 37
 - recurrent skills, 37
 - stable skills, 36, 59, 78
 - unstable skills, 36, 59, 78
- state-orientation, 29, 214
- strategies, 29, 34, 38, **44**, 46, 56, 59, 80, 86, 99, 102, 214, 224, 227
 - adaptive strategies, 30, 44, 47
 - attention management strategies, 47
 - problem solving strategies, 30, 45, 57
 - work strategies, 30
- strategy learning, 13, 38, 46, 53
- stress, **28**, 31, 44, 81, 215, 224, 226
- subject matter expert, 73
- task, **26**, 27, 29, 30, 44, 47, 48, 57, 58, 59, 60, 61, 71, 76, 86, 98, 100, 103, 108, 112, 168, 176, 182, 211
 - objective task, 26, 30, 42, 76
 - subjective task, 26, 30, 42, 76
- task selection, 15, 49, 51, 175, 227, 228
 - dynamic task selection, 13, 52
- team resource management, 81
- trainability, 13, **33**, 36, 47, 59, 68, 75, 100, 110, 213, 216, 225, 242
- training design, **48**, 71, 74

- training objectives, 49, 51, 52, 53, 86, 102, 105, 111
- validity, 64, 67, 75, 88, 137, 164, 175, 188, 203, 207, 210, 242
 - concurrent validity, 64
 - construct validity, 65
 - content validity, 65
 - face validity, 65
 - predictive validity, 64, 176, 210, 215, 216
- varied task components, 36
- whole-task training, 52
- work sample, 61, 66, 67, 98, 100
- workload, 42, 44, 49, 80, 214, 224
- workload management, 77, 109, 214, 224, 226, 227
- Yerkes-Dodson law, 28
- zone of proximal development, 58

Curriculum Vitae

Esther Oprins werd geboren op 29 mei 1970 in Gilze-Rijen. In 1988 deed zij eindexamen Gymnasium β op het Theresialyceum in Tilburg. Vervolgens is zij Nederlandse Taal- en Letterkunde gaan studeren aan de Universiteit Utrecht. Ze deed de specialisatie Taalkunde (Syntaxis & Semantiek), en behaalde haar diploma in 1994. Een jaar later (1995) kreeg ze haar bevoegdheid tot eerstegraads lerares Nederlands na afronding van de postdoctorale lerarenopleiding (IVLOS) aan de Universiteit Utrecht. Omdat zij zich liever bezig hield met het onderwijs achter de schermen in plaats van voor de klas, is zij direct daarna gestart met de studie Onderwijskunde aan de Universiteit Utrecht. In 1999 studeerde zij af (Cum Laude) met als specialisatie Leren met Nieuwe Media. Zij was toen al twee jaar werkzaam bij het Centrum voor Innovatie van Opleidingen (CINOP) in 's Hertogenbosch, waar zij ook haar stage en afstudeeronderzoek voor Onderwijskunde had uitgevoerd. In de functie van Opleidingskundig Adviseur hield zij zich vooral bezig met onderzoek, advisering en ontwerp omtrent e-learning, simulatie en simulator training voor het bedrijfsleven en de overheid.

Per 1 januari 2002 kwam zij in dienst bij Luchtverkeersleiding Nederland (LVNL) op Schiphol als Training Expert. Vanuit de afdeling die inmiddels de naam Human Factor draagt, heeft zij het schrijven van dit proefschrift voor de Universiteit Maastricht gecombineerd met werkzaamheden voor de selectie en training van aspirant luchtverkeersleiders. Haar belangrijkste taak was om (samen met anderen) een nieuw selectie- en beoordelingssysteem voor LVNL te ontwerpen dat geïmplementeerd werd vanaf 2003, zoals beschreven staat in dit proefschrift. Sinds 1 januari 2008 werkt zij als Concepts Expert bij de afdeling Research & Development van LVNL, waar ze, voortbouwend op haar human factors expertise, onderzoek doet naar toekomstige systemen voor luchtverkeersleiding.