2007

Analysis of Learning Curves: Evaluation of Assessment in Atc Training

Esther Oprins

Ernst Burggraaff

Hans van Weerdenburg

ANALYSIS OF LEARNING CURVES: EVALUATION OF ASSESSMENT IN ATC TRAINING

Esther Oprins, Ernst Burggraaff & Hans van Weerdenburg
Air Traffic Control the Netherlands (LVNL)
Amsterdam, The Netherlands

This paper describes the analysis of learning curves as part of the evaluation of a competence-based assessment system designed for air traffic control (ATC) simulator and on-the-job training (OJT) at LVNL. This system should contribute to an increased output from training by making learning processes more effective and efficient. Learning curves were derived from assessment results obtained in training. Patterns and individual differences in learning should be recognized in these (recalibrated) learning curves if the assessment system is well-designed, needed for adequate feedback and interventions. We divided the trainees into three groups and we compared their learning curves with patterns of prototypical learning curves. Next, we analysed differences in development of competences. The results show that the assessment system represents patterns and individual differences in learning adequately.

Introduction

At Air Traffic Control the Netherlands (LVNL) we designed a competence-based assessment system for simulator and on-the-job training and we evaluated the reliability and validity of the system (Oprins, Burggraaff & Van Weerdenburg, 2006). Another part of the evaluation involves analysis of learning curves derived from assessment results: a well-designed assessment system works adequately if it represents learning processes optimally. The assessment system should clearly distinguish patterns and individual differences in learning processes (e.g., slow starters, learning plateaus) and in performance (strengths and weaknesses) to provide a basis for adequate feedback and interventions with the purpose of adapting the training to trainees' needs (task selection, coaching, remedial teaching). In this way, assessment would contribute to make training more effective and efficient. Pass/fail decisions would be more valid if they are based on the recognition and predictability of patterns in learning processes.

Design of the assessment system

The assessment system designed for ATC training is competence-based. Competence is defined as the successful integration of knowledge, skills and attitudes and their application in realistic environments. The assessment of competences as they develop during training is a base for adequate feedback, needed to improve individual performance. All facets of competence are assessed (technical, cognitive, emotional, social) to get a complete picture of trainees' performance. We did a competence analysis with air traffic controllers and based on literature research. This resulted in the 'ATC Performance Model' (Oprins et al., 2006) which has served as a framework for assessment design. There are 14 competences to be assessed in training. Each competence is represented by a set of *performance criteria*. The criteria were directly derived from the model and formulated in the jargon of controllers for maximizing comprehension and recognition of behaviors. They are formulated as 'behavioral markers' (O'Connor et al., 2002) and rated at a six points rating scale. In order to follow trainees' 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. However, augmenting *performance standards* ('norms') are defined for the subsequent training phases in simulator training and OJT. These standards are formulated as behavioral examples, a variant on behaviorally anchored scales (BARS; Berk, 1986): they do not specify the scale positions, but standards to be achieved at the end of each phase. Consequently, assessors agree more on what is expected from trainees in intermediate phases, while for trainees it is clearer to which extent specific competences should be further developed.

Multiple assessors measure trainees' progression over one to two week periods in a *continuous assessment* system. A web-based assessment tool is used to make assessments, to store the results in a database and to generate overviews of trainee performance. In this way, authorized officials with access from several positions within LVNL can follow trainees' progression, and training results can easily be used for data analyses such as of learning curves.

Learning curves

Learning curves are usually presented as growth curves by measuring performance of the same task execution at successive moments of time, aimed at modelling learning processes based on general learning theory. A simplified ATC task, the Kanfer-Ackerman task (Ackerman, 1988), has often been used for examining complex skill acquisition (Lee & Anderson, 2001; Taatgen & Lee, 2003). General learning theory says that each learning curve ends in an asymptote (learning plateau) conform the power law of practice (Newell & Rosenbloom, 1981):

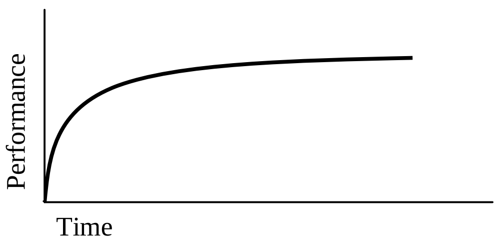

Figure 1. General learning curve

However, in practical training these learning curves are not as equal and smooth. As ATC is considered to be a complex cognitive skill, continued practice does not always lead to improved performance (Schneider, 1990); complex skill acquisition is more complicated, especially in practical settings. The ATC task consists of many subtasks to be learned. The total learning process has many parts with different learning curves because a trainee needs time to assimilate new knowledge and skills with previous experiences. Several learning plateaus may occur for an individual and individual differences in learning result in many (allowed) variations of learning processes.

The LVNL assessment system does not produce learning curves as usual. Continuous assessment is applied but recorded measurements are not really continuous because they only reflect performance at specific moments. Theoretically, we may not draw a straight line through the points of measurement because we don't know the precise form of the lines in-between. Besides, trainees are assessed against augmenting performance standards, corresponding to the training phases, while these standards are translated into the same rating scale. This means that the rating scale, which has 6-points with a value of 4 or more being sufficient, is constantly being recalibrated. When ratings stay the same over time this implies that the trainee constantly achieves the required standards and thus shows progression. Therefore, the 'learning curve' produced by our assessment system can only be a derivative of the real learning curve. Figure 2 shows three variations of recalibrated learning curves. Progression is measured four times (t1-t4) but not equally divided over the training period. The points are connected although the lines in-between do not necessarily reflect learning adequately. The red line represents a trainee with constant sufficient performance (horizontal black line), the green line refers to a slow starter and the blue line shows a learning plateau in-between:

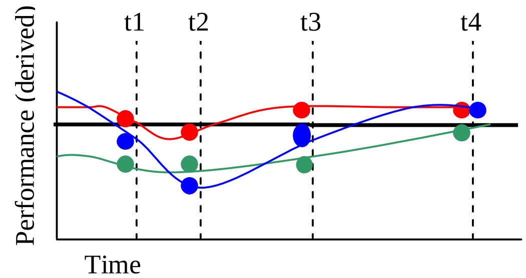

Figure 2. Learning curves (recalibrated)

Assuming that the assessors' ratings are sufficiently reliable, we examined these kinds of recalibrated learning curves resulting from assessments. Patterns and individual differences in learning should be recognized if the assessment system is well-designed.

Method

The scope of this paper is restricted to simulator training in area control (ACC) although we did the same analyses for OJT and other ATC functions. We included 146 progression measurements made for 25 'real' trainees at successive moments of time (once in two weeks) with different intervals for the trainees.

Prototypical learning curves

The first step was to define prototypical learning curves with which learning curves of actual trainees can be compared. In theory, trainees would start at a comparable performance level under the assumption that they do not differ in prerequisite knowledge and skills. Due to differences in learning, differences in performance level across trainees would become larger over time when asymptotes of learning are being reached. In the recalibrated learning curves some trainees would show decreasing performance because it becomes more difficult to achieve the increasing standards. As a result, trainees may fail during or at the end of training. We defined three prototypical patterns of recalibrated learning curves for *low*, *moderate* and *high performers*. Figure 3 provides an abstract visualization of this distinction:

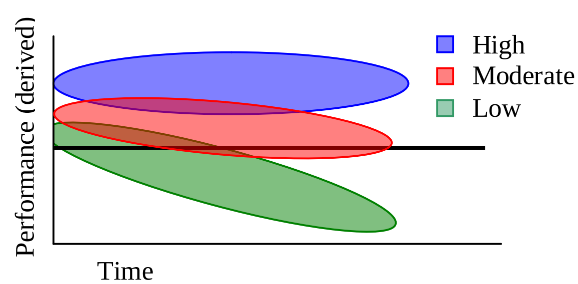

Figure 3. Patterns of prototypical learning curves

Performance of low performers would decrease; they do not achieve the required standards. The high performers would show a constant (sufficient or higher) performance level. The moderate performers would take an intermediate position, showing more variations (slow starters, learning plateaus) with generally sufficient performance at the end.

Analysis of learning processes

We distinguish two main variables in learning processes to be used in analyzing learning curves:
1) *Performance level:* weighted (averaged) sum of ratings on each competence over time.

2) *Progression:* the rate of growth over time. We consider the beta coefficient in the linear regression model to be indicative for progression although the linear model may not always fit for individual learning curves. Using these variables we made learning curves with graphs and curve fitting in SPSS. We examined general patterns of learning as well as differences between individual trainees. We did additional analyses to get quantitative evidence for findings visualized in the graphs. Besides, we analyzed differences between trainees in more detail by comparing (progression on) singular competences. Finding differences in competences between trainees, including possible deficiencies, is important because they are the base for adequate feedback in learning.

Results

Patterns in learning curves

We divided the trainees in ACC simulator training into the groups low, moderate and high performers based on training success also in OJT that follows simulator training. The failures in simulator training (N=9) were considered to be the low performers. The passed trainees were divided into two equal groups (N=8) based on a ranking of the overall performance level in OJT: the moderate performers did have an averaged sum of ratings lower than 4.3 in OJT, while this sum for high performers was higher than 4.3. We did not use the criterion pass-fail in OJT, because simulator training was not completely predictive for OJT; six of the eight moderate performers failed in OJT and only one of the high performers.

We did a discriminant analysis to check whether this classification was correctly predicted by the variables used to define learning curves, i.e. performance level and progression in simulator training. An analysis of variance (ANOVA) shows that the groups differ significantly on performance level (F=30.83, df=22, p=.000) and progression (F=7.12, df=22, p=.004). The first discriminant function is significantly different (chi-square=42.80, df=4, p=.000) across groups while performance level is a better predictor than progression. In total, the discriminant function successfully predicted group membership for 84%.

Next, we made graphs (time sequence) of recalibrated learning curves resulting from assessments for each group. These graphs are presented in figure 4, 5 and 6. At value 4 a horizontal black line is drawn: the norm for sufficient performance. The moments of time are only a rank order because the intervals differ across trainees. Not all trainees achieved the latter moments of time since they failed (selection effect) or they received fewer progression measurements:

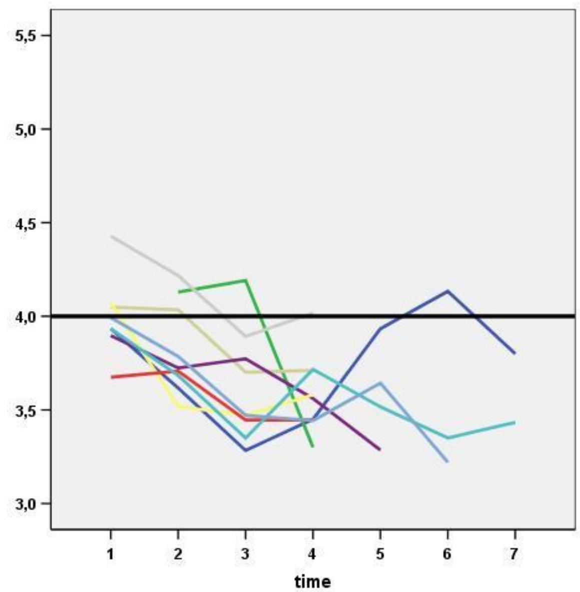

Figure 4. Learning curves of low performers

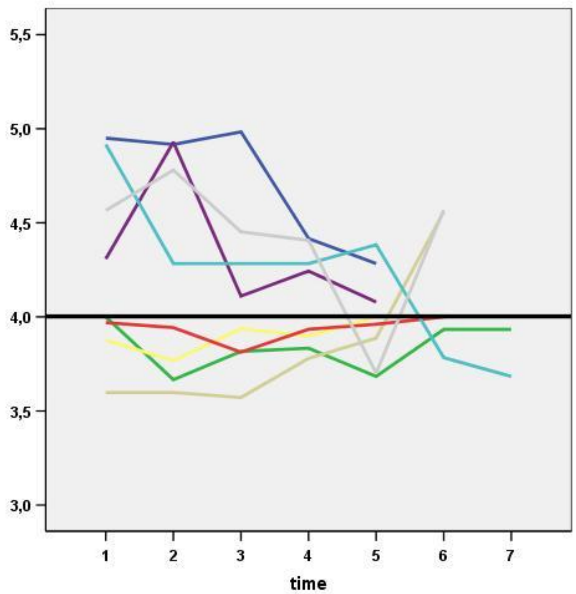

Figure 5. Learning curves of moderate performers

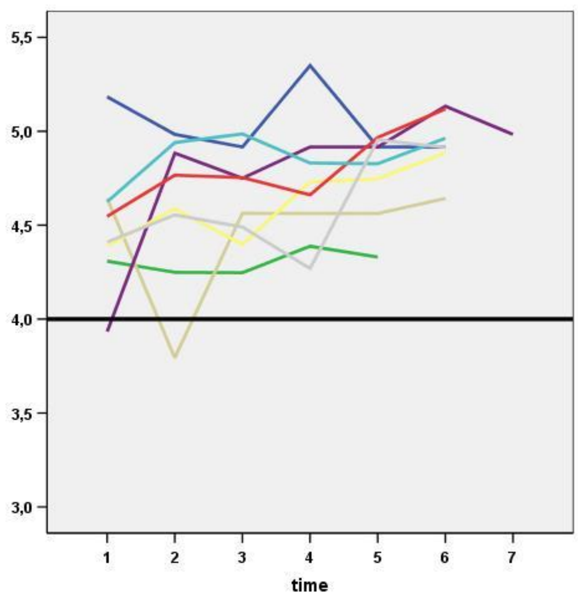

Figure 6. Learning curves of high performers

The correspondence with the prototypical patterns of learning curves (figure 3) can easily be recognized Some exceptions exist, for instance, the blue line in figure 4 at the end. But we should realize that the assessors' ratings probably are not always reliable in practical training. As expected, many differences within the group of moderate performers exist; we recognize a slow starter in the brown line. The high performers generally show increasing performance above sufficient standards; this could point at a higher (implicit) norm for really high performers.

Finally, we applied curve fitting for each group (all trainees combined in one model) and we found the following linear regression models with the beta coefficient (b1) as indication for progression:

Table 1. Lineair regression models of curve fitting

Group	Rsq	F	df1	df2	Sig.	b1
Low	.191	9.906	1	42	.003	-.072
Moderate	.064	3.371	1	49	.072	-.053
High	.022	1.069	1	48	.306	.024

The model of the low performers fits best and even better if we would delete the only real outlier (blue line in figure 4). Individual differences that exist within each group explain why a fitting (lineair) model is not likely to occur as we found for the moderate and high performers. Other (non-linear) models were tried but these did not fit better. The fact that the moments of time are only successive with different intervals for various trainees could have affected the results. As expected, the low (and moderate) performers show negative progression; positive progression is found for the high performers.

Differences in learning curves

The assessment system should help to differentiate between trainees, needed for individualized feedback. We argued that differences between trainees in performance level would increase over time: their entrance level is similar but asymptotes of learning arise. We consider the standard deviation to be an indication of differences in performance level between trainees. We made graphs of the means (figure 7) and standard deviations (figure 8) of the three groups and in total. Figure 7 summarizes the graphs of individuals in figure 4-6 by calculating their means at successive moments of time:

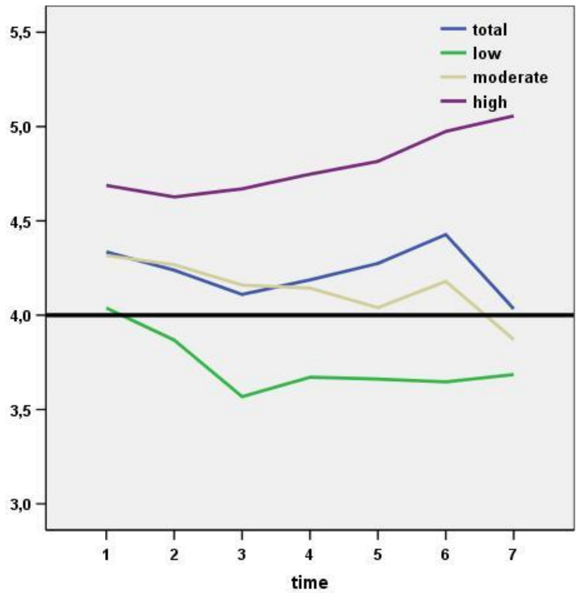

Figure 7. Mean performance level over time

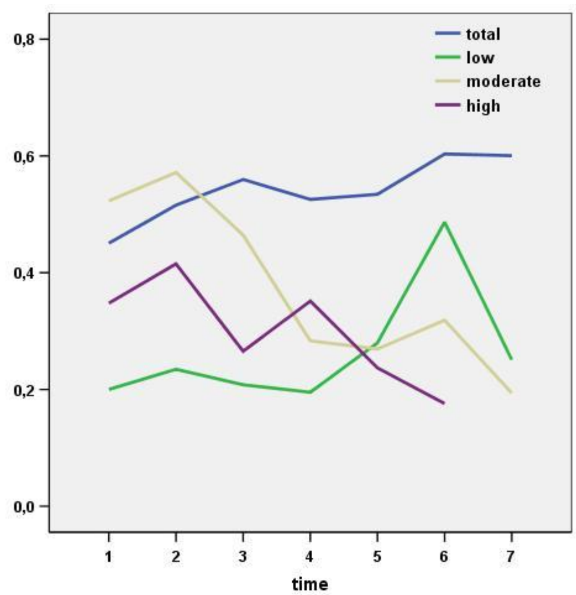

Figure 8. Standard deviation (differences) over time

Figure 7 shows that the means of the total number of trainees stay rather similar over time. The graphs clearly differ in means across the three groups. As expected, all trainees start at a sufficient level but the means of low (and moderate) performers decrease over time, due to more difficulties with achieving the required standards. In figure 8 the standard deviations of the total number of trainees increase over time conform our expectations. Differences within a group of comparable performers are not as likely to increase which is visible for moderate and high performers.

Differences in competences

Singular competences are even more important for supporting individual learning: they help to identify deficiencies of trainees required for appropriate feedback and interventions. We calculated the means of singular competences for each trainee. We did an analysis of variance (ANOVA) to examine whether the means of competences were significantly different across the three groups. This was verified for all competences, significant at $p<.001$.

Next, we calculated correlations (Pearson) of the competences with the variable 'group', taken as an overall measure of ATC performance. We found significant correlations ($p<.001$) respectively for *decisiveness* (.86), *mental picture* (.84), *workload management* (.83), *attention management* (.82), *efficiency* (.81), *safety* (.80), *planning* (.73), *listening* (.69), *co-ordination* (.67), *verbal expression* (.55) and *equipment operation* (.53). The lowest correlations but significant ($p<.05$) were respectively found for *team orientation* (.48), *attitude* (.46) and *label management* (.46). The competences with the highest correlations are assumed to be most critical for ATC performance. In previous studies (Oprins et al., 2006) we found that the same competences were the main reasons for failing in the group of low performers. We also found high intercorrelations between the (averaged) competences which may have influenced the high correlations found in this study. These high intercorrelations suggest that only one construct is measured, air traffic management. However, they are undesirable because the competences cannot easily be distinguished while this is needed for the detection of individual strengths and weaknesses. Alternatively, we calculated another measure that should point at possible deficiencies: we counted the number of insufficient ratings (value < 4) on each competence over time for each trainee and next we averaged them. The results presented in figure 9 show that this leads to more differentiation across the three groups than averaged competences. 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 performers never have 50% or more insufficient ratings. Less critical competences *equipment operation*, *attitude* and *team orientation* are hardly rated as insufficiently.

Differences in competence development

Analyses of averaged competences ignore time influences while trainees may differ in development of specific competences. Some competences may be more trainable than others. This relates to the distinction between ‘consistent components’ that improve by more practice (cf. *label management*) and ‘not-consistent components’ that are often reasons for failing (cf. *mental picture*) (Schneider, 1990). We examined progression of competences by calculating rank order correlations (Spearman) of the variable time (successive moments) with the ratings on competences, serving as a measure for progression. Table 2 presents the results for the three groups.

Table 2 shows that differences in progression on specific competences are minimal for the total group of trainees while they are present when we split up the trainees into the three groups. The low performers have negative correlations for all competences except *equipment operation* and *label management* which are more trainable indeed.

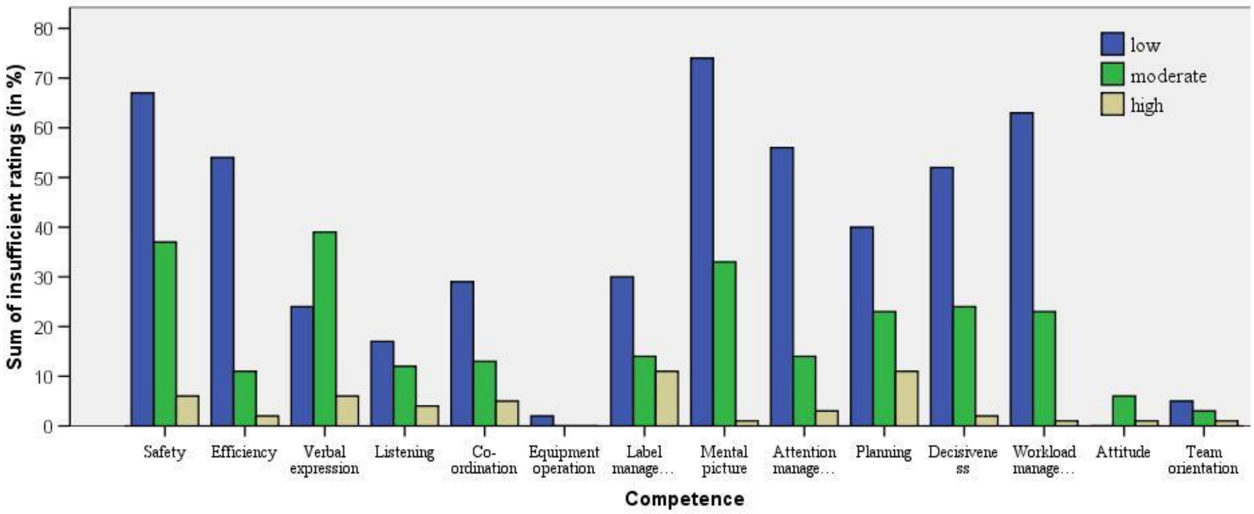

Figure 9. Averaged sum of insufficient ratings (in %) for competences for low, moderate and high performers

Table 2. Spearman correlations time vs. competences

Competence	Total (N=146)	Low (N=44)	Mod. (N=52)	High (N=50)
Safety	-.04	-.29	-.17	.20
Efficiency	.10	-.22	-.03	.36**
Verbal expression	-.07	-.19	-.20	.09
Listening	.02	-.24	-.11	.20
Co-ordination	.05	-.04	.02	.04
Equipment operation	.08	.11	.01	.10
Label management	.14	.04	.03	.31*
Mental picture	-.02	-.54**	-.10	.25
Attention managem.	.11	-.19	-.08	.28*
Planning	.01	-.44**	-.25	.41**
Decisiveness	.03	-.30*	-.18	.29*
Workload managem.	.03	-.19	-.09	.06
Attitude	-.08	-.17	-.16	-.08
Team orientation	-.05	-.11	-.20	.07

The highest significant correlation ($p<.001$) is found for *mental picture*. Apparently, this competence differentiates the most in both its mean (see figure 9) and in progression; trainability of *mental picture* appears to be very low. In this context, *planning* and *decisiveness* are also relevant because they indicate a significantly high negative progression for low performers but a high positive progression for high performers. Standards of *efficiency* grow during training; the results suggest that high performers work more efficiently over time. The results confirm the findings presented in figure 7: high performers tend to show positive progression in contrast with low and moderate performers.

Discussion and conclusions

The assessment system designed for ATC training at LVNL can produce (recalibrated) learning curves that are sufficiently representative for general learning processes, although complex skill acquisition in practical training (simulator, OJT) is complicated and assessors' ratings are not always completely reliable. We distinguished three groups of trainees based on training success. Their recalibrated learning curves derived from assessment results agreed with the three defined patterns of prototypical learning curves. Classification based on the variables performance level and progression used in learning curves was correctly predicted. Individual differences in learning (slow starters, learning plateaus) were recognized as well. As expected, differences in performance level between trainees increased over time due to arising asymptotes of learning. Besides general performance, trainees differ in progression on competences. *Mental picture* differentiated most between low and high performers for ratings over time and for progression. It appears to be minimally learnable in contrast with less critical competences (e.g. *label management).*

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, needed for providing adequate feedback and adapting training to trainees' needs, making training more effective and efficient. The next step is to analyze predictability of learning curves and development of competences in order to make pass/fail decisions more valid. Therefore, more longitudinal research in learning curves is needed.

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.

Berk, R.A. (1986). *Performance assessment: methods and applications.* Baltimore: The Johns Hopkins University.

Lee, F. & Anderson, J.R. (2001). Does learning a complex task have to be complex? A study in learning decomposition. *Cognitive psychology, 42,* 267-316.

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.* Hillsdale, NJ: Erlbaum.

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.

Oprins, E., Burggraaff, E. & Van Weerdenburg, H. (2006). Design of a competence-based assessment system for ATC training. *The international journal of aviation psychology, 16*(3), 297-320.

Schneider, W. (1990). Training high-performance skills: fallacies and guidelines. In M. Venturino (Ed.). *Selected readings in human factors.* Santa Monica, CA: The human factors society.

Taatgen, N.A., & Lee, F.L. (2003). Production compilation: a simple mechanism to model complex skill acquisition. *Human Factors, 45*(1), 61-76.