2007

Procedural Control in ATC Selection Tests to Predict Situational Awareness

Esther Oprins

Esther Geven

Emiel Veldhuijzen

Robert A. Roe

Repository Citation

Oprins, E., Geven, E., Veldhuijzen, E., & Roe, R. A. (2007). Procedural Control in ATC Selection Tests to Predict Situational Awareness. *2007 International Symposium on Aviation Psychology*, 505-511.
https://corescholar.libraries.wright.edu/isap_2007/49

PROCEDURAL CONTROL IN ATC SELECTION TESTS TO PREDICT SITUATIONAL AWARENESS

Esther Oprins[1], Esther Geven[1], Emiel Veldhuijzen[1], Robert A. Roe[2]
[1]Air Traffic Control the Netherlands (LVNL)
[2]Maastricht University, Department of Organization & Strategy

At LVNL we have developed a new selection system, called DATCOSS, which should contribute to a higher output of qualified controllers from training. Two job sample tests are part of this selection system, specifically designed to measure the candidate's potential for Situational Awareness by using simplified procedural control tasks. Grading takes half a day while AAPRO is a selective training module of five weeks. We examined the psychometric quality of both tests and the predictive validity of Grading for AAPRO. We may conclude that SA is sufficiently measured in the two job samples and that Grading results are rather predictive for performance in AAPRO. We made a start with analyzing predictive validity in relation to training success; this will be further examined in the near future.

Introduction

Air Traffic Control the Netherlands (LVNL) faces a persistent shortage of qualified controllers. The previous system of recruitment, selection and training has not been able to produce sufficient numbers of qualified controllers. Therefore, we have developed the Dutch Air Traffic Controller Selection System (DATCOSS), which was implemented in 2003 (see also Oprins, Geven, Veldhuijzen, & Roe, 2006). Simultaneously, an overhaul and redesign of the training system was started. Special attention was paid to the design of two job sample instruments, one administrated during entrance selection, and one during pass-on selection in the form of a training module of five weeks. These job samples both involve simplified procedural control tasks, particularly aimed at measuring the candidate's potential for Situational Awareness (SA) since we consider SA to be one of the most critical competences in ATC work. This paper describes the design of these job samples as part of DATCOSS, the methods and the results of analyses, demonstrating the added value of using procedural control in selection tests for predicting SA in ATC.

Situational Awareness in ATC

A commonly made assumption is that operators in dynamic and complex tasks such as ATC create a mental representation of the changing environment, which makes it possible to keep 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' these patterns that help them to create order in a seemingly chaos by streaming traffic flows. Much research has been done on the three-dimensional 'mental picture' that controllers develop of the traffic situation. This is usually referred to as situation assessment, defined by Endsley (1995) as follows: '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 to be the product of the process situation assessment that takes place at three levels, respectively perception (SA1), interpretation (SA2), and anticipation (SA3). Attention management strategies are crucial to keep this continuously changing 'picture' up-to-date (Shebilske, Goetle & Garland, 2000).

We consider SA to be a very critical competence in ATC work. A lack of SA is one of the main reasons for failing in our ATC training. Procedural control tasks are assumed to require a high level of SA since the candidates cannot rely on visual aids such as a radar screen. In the job samples used in DATCOSS candidates only have a time clock, paper strips and a map of the airspace available to form their own mental picture of the traffic situation. Based on their SA, candidates should handle the traffic by issuing instructions to (pseudo) pilots by means of radiotelephony. In this way, these procedural control tests are particularly designed to predict SA in ATC work among other essential competences. This makes them different from many other existing job samples in ATC selection.

The selection system DATCOSS

The selection system DATCOSS combines sign and sampling methods and consists of two main parts, covering entry selection and pass-on selection.

Entry selection

Entry selection uses a battery of tests measuring cognitive abilities, information processing abilities, low fidelity job samples, and self-descriptive personality scales. After having made these tests candidates are successively subjected to an interview and an assessment centre. The entry selection ends with administration of the first high-fidelity job sample based on procedural control, called 'Grading'. The prediction model was chosen to be accumulative and compensatory using a priori weights (Roe, 2005, 2006). This implies that each stage contributes to the prediction of success as a controller in two ways. At each stage a certain number of candidates is first rejected on the basis of a (weighted) composite score obtained in this stage (p% remain). Next, a certain number of the remaining candidates is rejected based on this score compensated with (weighted) composite scores obtained in previous stages (q% remain).

The Grading consists of an exercise followed by a test that takes 45 minutes. The required theory should be prepared at home and is examined prior to the exercise to control for level of preparation and understanding. The candidate's performance is scored by means of rather objective methods; this is possible because only 14 aircraft should be handled. Observation checklists are used and all possible safety violations are predefined. *Safety* is measured objectively by counting these violations. In addition, assessors rate a set of ATC competences at a 6-points rating scale: *attention management, mental picture, planning*, *decisiveness* and *workload management.* The first three are most strongly related to SA according to Endsley's definition, but the others are relevant for SA as well. Competences to be rated that are less critical for SA are: *communication, strip management*, *efficiency*. A weighted sum of ratings is calculated into a percentage which serves as a cut-off for selection with thresholds: rejection below 55%, admission above 70%, and qualitative compensation with previous selection results between 55 and 70%.

Pass-on selection

The remaining candidates are enrolled into an Initial Training program of one year. After half a year the candidates are subjected to pass-on selection which consists of the training module 'AAPRO' (Area Approach PROcedural control), the second high-fidelity job sample. The candidates are trained in a basic simulator during a period of five weeks. Their performance is assessed by multiple controllers in two ways: progression reports are made weekly in which also progression is assessed, and two simulator tests are taken, based on a one hour exercise. Because we assume that performance in the job samples optimally reflects and predicts performance in training and in operational work, the same set of ATC competences is assessed in Grading, AAPRO and in subsequent training, based on the ATC Performance Model (Oprins, Burggraaff & Van Weerdenburg, 2006). Each competence is represented by a set of performance criteria (behavioural markers), rated by the assessors at a 6-points scale. On the basis of the ratings a weighted sum is calculated into a percentage for each assessment. In progression reports also progression is withdrawn in this weighted sum to serve as a measure for learning potential. Based on the assessment scores a weighted final AAPRO score is calculated that serves as a cut-off with thresholds for the selection decision. Candidates can be assigned to two different function categories: area, tower and approach controllers for mainport Schiphol (CAT1); and tower and approach controllers at regional airports or ground controllers at Schiphol (CAT2). As such, different cut-offs are used for the two categories. Candidates are always rejected below 55% and admitted above 75%. Qualitative compensation is applied in-between score range 55-75% (although with different thresholds for CAT1 and CAT2), based on assessment results obtained in the entry selection and a group discussion with the assessors. Unless rejected at this stage, candidates complete the Initial Training and enter in Unit Training. Figure 1 illustrates the selection system and its relationship with training:

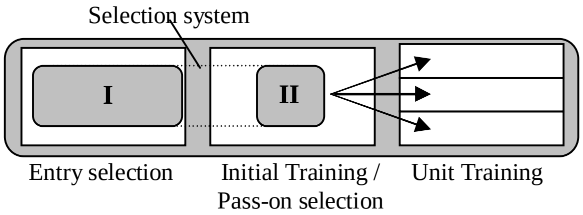

Figure 1. Selection and training system

Although the primary aim of DATCOSS is to predict job success, it was expected that Grading would also predict performance in AAPRO due to the high similarity between the job samples, bearing in mind the high costs of training and scarce availability of controllers who are needed as assessors.

Method

In total, we included 206 Grading and 92 AAPRO candidates who were assessed during the period November 2003 – December 2006; 77 candidates did both Grading and AAPRO. We investigated the psychometric quality of the two selection instruments separately, using the same methods. We examined construct validity, the internal structure of the competence ratings, and the relationship between the competences and the selection decisions. Only for Grading we also analysed possible differences across assessors. Next, we investigated the predictive validity of Grading with regard to AAPRO by comparing final scores, selection decisions and competence ratings in the two job samples. Finally, we made a first start with predictive validity of AAPRO by presenting some training results. Unfortunately, given the training time demanded in Unit Training, many candidates have not completed the full training yet. As such, we are awaiting more results to be able to start a full predictive validity study against overall job success.

Results

Grading

The selection ratio of Grading in our data set is 55.8% (N=206). Table 1 presents the distribution of final scores; some exceptions below or above the thresholds were made on qualitative grounds:

Table 1. Distribution of final scores in Grading

	Fail	Pass	Total
< 55%	61	8	69
55 < > 70%	26	24	50
> 70%	4	83	87
Total	91	115	206

There are 7 assessors. We cannot examine interrater agreement because only one controller assesses the same candidate, but we compared the means of the final scores across assessors with an analysis of variance (ANOVA). The results show that the means do not differ significantly from each other (*F*=1.283, df=6, p=.267). Comparable results are found for the competence ratings with only one exception for *strip management* (*F*=2.245, df=6, p=.041). The results imply that possible severity or leniency errors do not exist, which refer to the assessors' tendency to give respectively lower or higher ratings systematically.

Besides the calculated final score, assessors provide a subjective impression on a 5-points scale, based on their experience as an assessor, before they start to rate the competences. This subjective impression was assumed to be an appropriate indication of the candidates' performance. The relationship between the two measures may serve as a measure for construct validity: the extent to which the final score represents ATC performance. The correlation coefficient (Pearson) between the two measures is .86, significant at p<.001, very high as expected.

Next, we examined the internal structure of competence ratings with a factor analysis (principal components). The results are presented in Table 2:

Table 2. Component matrix of competences (N=206)

Grading competences	Component 1
Mental picture	.88
Attention management	.88
Decisiveness	.86
Planning	.86
Efficiency	.83
Workload management	.83
Strip management	.87
Safety	.79
Communication	.78

Table 2 shows that only one component emerged, which indicates that only one construct is measured: ATC performance. High intercorrelations (Pearson) found between the competence ratings confirm this result: between .46 and .77, all significant at p<.001.

Finally, we explored how the competence ratings were related to pass/fail (criterion variable) and to which extent this classification was made correctly by means of a discriminant analysis. Analysis of variance (ANOVA) showed that the failed and passed candidates differ significantly on each competence (p<.001). A single discriminant function was calculated which was significantly different between pass and fail (chi-square=148.16, df=9, p<.001). The best predictors based on the discriminant function coefficients are respectively *planning*, *attention management*, *safety*, *decisiveness*, *mental picture*, *workload management.* These are considered to be the most critical competences and strongly related with SA in comparison with less critical competences and the least predictors as expected: respectively *strip management*, *efficiency*, *communication.* The discriminant function successfully predicts group membership for 85.1% in total.

AAPRO

By the end of 2006 seven groups of candidates (N=92) had participated in AAPRO. We only included their assessments from week 3 onwards because the first two weeks are aimed at trainees' familiarization with procedural control and learning to apply the rules; they are not part of the final AAPRO score that determines the selection decision. There are five assessments for each trainee: three progression reports made from week 3 and two simulator tests. They are all made by two assessors together for maximizing objectivity, but therefore we cannot examine differences between assessors.

Table 3 presents for the three groups the output in the two function categories and the means of final scores at which the selection decision is mainly based:

Table 3. Output and means of AAPRO groups

Group	N	Output			Mean
		CAT1	CAT2	Exit	(in %)
1	14	6	2	6	63.6
2	10	5	1	4	64.9
3	12	3	3	6	59.7
4	14	6	1	7	64.2
5	14	7	2	5	65.2
6	14	4	3	7	60.5
7	14	6	1	7	63.8
Selection ratio		40.2%	14.1%	45.7%	

Table 3 shows that the groups are rather comparable with each other, although the output of group 3 and 6 as well as their means are relatively low. But the means are not significantly different across the groups (*F*=.366, df=6, p=.90). The overall selection ratio (CAT1 and CAT2) is 54.3%. The distribution of final scores in categories is presented in table 4:

Table 4. Distribution of final scores

	CAT1	CAT2	Exit
< 55%	0	0	26
55 < > 60%	0	2	10
60 < > 65%	2	9	4
65 < > 70%	4	2	0
> 70%	31	0	0

Table 4 indicates that the thresholds for the two function categories are different, and that the rules for compensation between the thresholds (55-70%) have been consistently used in the selection decisions.

In the same way as in Grading, the assessors gave a subjective impression before they start to rate the competences, but only for the weekly progression reports (not for simulator tests) and not for group 1 yet. The scale, however, differs with Grading: a percentage was used instead of a rating scale. As in Grading, the relationship between the two measures can be taken as an index of construct validity. The correlation coefficient (Pearson) between the two measures for all reports (N=217) is .96, significant at p<.001. Because the scales are similar (percentages), we also calculated the absolute difference between the two measures, averaged for all reports. This D-index, serving as another measure for construct validity, is extremely low: 2.2%. However, we should realize that the assessors see the objective final score afterwards, although they officially are not allowed to change their ratings anymore.

Next, we examined the internal structure of the competence ratings with factor analysis (principal components). We used the averaged ratings on the five assessments for each single trainee. The results of the factor analysis are presented in table 5:

Table 5. Component matrix of competences (N=92)

AAPRO competences	Component 1
Decisiveness	0.98
Planning	0.96
Mental picture	0.96
Attention management	0.95
Workload management	0.95
Efficiency	0.94
Safety	0.85
Co-ordination	0.80
Strip management	0.77
Communication	0.70
Attitude	0.68

As in Grading only one component emerged, thus we may conclude again that one construct is measured, ATC performance. High intercorrelations (Pearson) were also found between the competences in AAPRO: between .45 and .93, significant at p<.001.

Finally, we did a discriminant analyses to explore how the competence ratings predict membership in the three groups (CAT1, CAT2 and exit). Again we used the averaged ratings for each trainee. An analysis of variance (ANOVA) shows that the means of all competences are significantly different across the three groups at p < .001. The value of the first function is significantly different (chi-square=124.30, df=22, p<.001). The best predictors following from correlations between discriminant variables and the first function are respectively *safety, mental picture, planning, attention management, decisiveness, efficiency,* and *workload management.* The best predictors are those measuring the competences mostly related to SA. As expected, the least effective predictors are respectively *strip management, co-ordination,* and *communication* and *attitude.* The discriminant function successfully predicted group membership for 82.6% in total.

Grading and AAPRO

We examined predictive validity of Grading for AAPRO at the overall performance level and at the competence level, using the subset of 77 candidates who did both Grading and AAPRO. The correlation coefficient (Pearson) between the final scores of Grading and AAPRO is .50. The rank order correlation coefficient (Spearman) between the final score of Grading and the selection decision in

AAPRO (CAT1, CAT2, exit) is .45, both significant at p<.001. Thus, there is a strong relationship, certainly if restriction of range is taken into account. This result is confirmed by the fact that 25 CAT1 candidates (81% of this group; see table 4) have a Grading final score above 75%; thus this group of the best candidates in AAPRO did show a very high performance in Grading as well, although there also exist a certain number of false positives.

Next, we calculated the rank order correlation coefficients (Spearman) for each competence rated in Grading with the final score in AAPRO in order to examine the predictiveness of each competence rated in Grading. Table 6 presents the results, ordered from high to low, not corrected for restriction of range:

Table 6. Correlations (Spearman) of Grading competences with AAPRO final score (N=77)

Grading competences	Correlation with AAPRO
Decisiveness	.50**
Attention management	.45**
Safety	.41**
Workload management	.38**
Planning	.35**
Communication	.25*
Mental picture	.24*
Efficiency	.23
Strip management	.18

The highest correlations were found with the most critical competences and which are strongly related with SA as is shown in Table 6.

In addition, we calculated the correlation coefficients (Pearson) of the competences rated in Grading with the same competences rated in AAPRO in order to examine their relationship. They are presented in table 7 ordered from high to low, not corrected for restriction of range:

Table 7. Correlations (Pearson) between the same competences rated in Grading and in AAPRO

Competences	Correlation Grad./AAPRO
Workload management	.47**
Decisiveness	.38**
Mental picture	.34**
Attention management	.34**
Strip management	.32**
Communication	.28*
Efficiency	.25*
Safety	.24*
Planning	.22

Table 7 presents high correlations which are expected because of the high intercorrelations found between the competences in both job samples and because of the high correlations with the selection decision in AAPRO (see table 6). However, some competences rated in Grading are more correlated with other competences rated in AAPRO (not presented here). We should realize that their meaning may be different if they are assessed in different tasks (job samples), and that the candidates' competences in Grading do not necessarily predict how the candidates acquire the competences in AAPRO from the point-of-view that competences are not innate abilities but that they are the result of learning processes.

AAPRO and training results

Finally, we did the first step towards analysis of predictive validity by examining the candidates' success in training. Because most candidates are still in training, we consider the achievement of the first rating as a criterion besides overall training success. Generally, trainees do not fail in subsequent training phases for next ratings. Table 8 presents the status of candidates in training divided into the two categories:

Table 8. Status of candidates in training

Status of candidates	CAT1	CAT2
Passed 1st rating (and 2nd rating)	12	5
Failed in Unit Training (1st rating)	7	4
In Unit Training (1st rating)	10	3
In Initial Training	6	1
Stopped (own choice)	2	-

Table 8 shows that a rather high number of trainees has failed already. However, five of them are from group 1 with a different order in training that preceded the AAPRO, and these candidates were not selected with the new selection system DATCOSS yet. The failure rate in Unit Training was extremely high in the last 10 years (around 50%), but we really see an increasing pass rate since two years and a better performance of candidates in Unit Training, especially of CAT1 candidates. Other factors probably have played a role as well such as a redesign of the training and assessment system (Oprins, Burggraaff & Van Weerdenburg, 2006).

Discussion and conclusions

The two job sample tests Grading and AAPRO, as part of the new selection system (DATCOSS) in use by LVNL, should contribute to increase the output of qualified controllers from training. These job samples are both simplified procedural control tasks aimed at measuring the candidates' potential for Situational Awareness (SA). We examined the added value of using procedural control in selection tests for predicting SA in ATC work.

Psychometric quality

First, we examined the psychometric quality of each job sample instrument separately. The results point at many similarities between Grading and AAPRO, in which a comparable set of competences is rated by multiple assessors. The cut-offs and thresholds for the selection decisions, comparable in both tests, have consistently been maintained. Construct validity was estimated by comparing the final score, objectively calculated, with a subjective impression provided by the assessors. In both tests the two measures were strongly related. This confirms that construct validity is sufficiently high: ATC performance is measured and not something else. Only for Grading we examined possible differences across assessors, but rating errors such as leniency and severity did not have influenced the candidates' assessments. Next, the results of factor analysis show that only one component is measured in both Grading and AAPRO: ATC performance. The competences to be rated are highly intercorrelated. Finally, we explored how the set of competences was related to the selection decisions in the tests by means of a discriminant analysis. The classifications were quite correctly predicted while the competences that are most strongly related to SA were the best predictors in both Grading and AAPRO, such as *mental picture, attention management* and *planning.* This agrees with Endsley's definition of SA divided in the three levels. These results suggest that SA is sufficiently measured in the procedural control tasks Grading and AAPRO.

Predictive validity

Next, we analysed predictive validity of Grading for AAPRO. Although DATCOSS is aimed at predicting overall job success, Grading was assumed to predict AAPRO as well due to their similarities. Based on the results we may conclude that predictive validity is sufficiently high. Candidates' performance in AAPRO, expressed in final scores and in selection decisions, is strongly correlated with candidates' performance in Grading. Competences that are rated in Grading and that are most critical and highly related to SA appear to be most predictive for performance in AAPRO, although this result can be influenced by the high intercorrelations between the competence ratings.

Finally, we started to examine the predictive validity of AAPRO for training success, although the period since the implementation of DATCOSS is too short yet for drawing conclusions. Training success of the candidates in the first AAPRO group, not selected with DATCOSS yet, is not very high. However, the pass rate has been increased since two years.

Despite of the lack on quantitative evidence at this moment due to small numbers, Grading and AAPRO have led to other positive results. First, the assessors, who are also coaches in Unit Training, have more confidence in their trainees who have been selected by themselves. Coaches put more effort in training interventions that may increase the trainee's chance on success. For the same reason, important changes in Unit Training have been made such as a more gradual sequence of simulator exercises. The fact that the same competences are assessed in the job samples and in training makes candidates' behaviour better recognizable for assessors.

These competences will be the basis for long-term validation research. The training results, including the competence ratings, are stored in a database so that they can be related to selection results. In this way, validation research can be done at a more detailed level than only using the pass/fail criterion. We are analysing the main reasons for failing related to specific competences such as SA both quantitatively and qualitatively. Besides, we are trying to get more insight into learning processes (learning curves) and the role of progression for making better predictions in selection and in training ultimately.

References

Endsley, M. (1995). Towards a theory of situational awareness in dynamic systems. *Human factors, 37 (1),* 32-64.

Garland, D.J., Stein, E.S. & Muller, J.K. (1999). Air traffic controller memory: capabilities, limitations and volatility. In: Garland, D.J., Wise, J.A. & Hopkin, V.D (Eds.). *Handbook of aviation human factors* (pp. 455-496). Mahwah, NJ: Erlbaum.

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.

Oprins, E., Geven, E., Veldhuijzen, E., & Roe, R.A (2006). Development of a new selection system for air traffic controllers: design, implementation and initial validity evidence. *Proceedings of the 27th EAAP conference, Potsdam.*

Roe, R. A. (2005). The design of selection systems: Context, 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.

Shebilske, W.L., Goetll, B.P. & Garland, D.J. (2000). Situation awareness, automation and training. In: Endsley, M.R. & Garland, D.J. (Eds.). *Situation awareness: analysis and measurement* (pp. 303-323). Mahwah, NJ: Lawrence Erlbaum Associates.