9 Validation of TACOM Measure

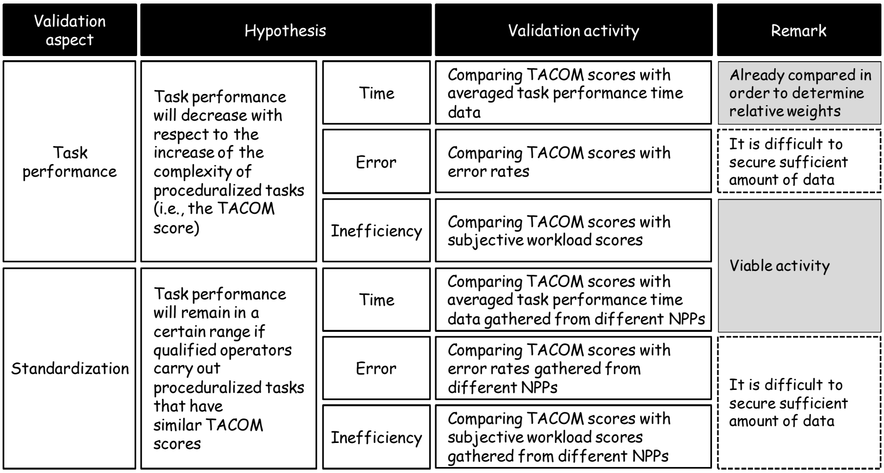
From the previous chapter, the TACOM measure is now available to quantify the complexity of proceduralized tasks. Therefore, the last question about the development of the TACOM measure would be: *is the TACOM measure meaningful for quantifying the complexity of proceduralized tasks*?

In order to answer this question, we can consider two kinds of validation. The first one is to directly compare the performance of qualified operators with the associated TACOM scores. That is, one should be able to validate the appropriateness of the TACOM measure from the point of view of three performance dimensions – *time, error,* and *efficiency*. The second kind of validation can be deduced from one of the canonical advantages of a good procedure. As stated in Sect. 2.1, good procedures guarantee at least three major advantages, and one of them is the standardization of the performance of qualified operators. This means that if the TACOM measure can quantify the complexity of proceduralized tasks, then the performance of qualified operators should be similar when they are performing proceduralized tasks with similar TACOM scores.

9.1 Validation Activity – Outline

Let us look at Fig. 9.1, which illustrates the overall validation scheme regarding the appropriateness of the TACOM measure. In Fig. 9.1, detailed activities belonging to the first validation aspect correspond to TACOM scores vs. three kinds of performance data that represent the basic performance dimensions. Unfortunately, since the error rate of qualified operators is generally low, it is very difficult to collect a sufficient amount of error-related data. In addition, since the relative weights that are indispensable for quantifying TACOM scores have been determined by averaged task performance time data, it is reasonable to expect that there would be a significant correlation between averaged task performance time data and TACOM scores. For this reason, the only viable activity would be comparing TACOM scores with subjective workload scores to reflect the *inefficient* dimension.

Meanwhile, the validation activities belonging to the second category are very straightforward because the standardization aspect of the TACOM measure will be clarified by comparing TACOM scores with the associated performance data that were gathered not only from the reference NPPs but also from other NPPs. Unfortunately, although the standardization aspect should be clarified from the other two dimensions (i.e., the *error* and the *inefficiency*), the only viable activity seems to be comparing averaged task performance time data (i.e., the *time* dimension) due to the difficulty in securing the associated performance data.



Validation aspect	Hypothesis		Validation activity	Remark
Task performance	Task performance will decrease with respect to the increase of the complexity of proceduralized tasks (i.e., the TACOM score)	Time	Comparing TACOM scores with averaged task performance time data	Already compared in order to determine relative weights
		Error	Comparing TACOM scores with error rates	It is difficult to secure sufficient amount of data
		Inefficiency	Comparing TACOM scores with subjective workload scores	Viable activity
Standardization	Task performance will remain in a certain range if qualified operators carry out proceduralized tasks that have similar TACOM scores	Time	Comparing TACOM scores with averaged task performance time data gathered from different NPPs	
		Error	Comparing TACOM scores with error rates gathered from different NPPs	It is difficult to secure sufficient amount of data
		Inefficiency	Comparing TACOM scores with subjective workload scores gathered from different NPPs	

Fig. 9.1 Validation scheme of TACOM measure

9.2 Comparing with Subjective Workload Scores

9.2.1 NATA–TLX Technique

As stated by Henneman and Rouse (1984), the diagnostic performance of qualified operators will be ineffective if they reach a final decision through many subdecisions. This means that qualified operators who follow an ineffective way of thinking are likely to feel a high level of cognitive demand compared to those who follow an effective way of thinking, because the former expended more efforts than the latter. Thus, it is necessary to emphasize that a subjective workload is susceptible to a certain level of cognitive demands (Campbell 1988). This strongly suggests that a subjective workload would be a good indicator to represent the *inefficiency* dimension of human performance. In addition, since the amount of effort to be spent by qualified operators will increase as task complexity increases, the subjective workload should increase in proportion to the complexity of tasks to be performed (Stassen et al. 1990; Maynard and Hakel 1997; Li and Wieringa 2000; Hancock 1996; Wei et al. 1998).

Therefore, although many researchers have criticized the meaning of subjective workload scores, the TACOM measure can be regarded as a proper indicator of the complexity of proceduralized tasks, if there is a tendency whereby subjective workload scores increase as TACOM scores increase. For this reason, TACOM scores and subjective workload scores are compared in order to investigate the appropriateness of the TACOM measure from the point of view of the *inefficient* dimension.

Many kinds of subjective workload measurement techniques have been developed in recent decades (Vidulich and Tsang 1986; Nygren 1991; Dickinson et al. 1993; Hendy et al. 1993; Hancock 1996; Svensson et al. 1997; Hill et al. 1992). Of these, the NASA–TLX (National Aeronautics and Space Administration – task load index) technique has been selected as the reference method to measure subjective workload scores because it (1) provides detailed as well as diagnostic results (Hill et al. 1992), (2) is able to support the general prediction model for a subjective workload (Nygren 1991), and (3) is known as one of the most suitable techniques for evaluating the level of subjective workloads (Liu and Wickens 1994).

The NASA–TLX technique was first developed in the 1980s (Hart and Staveland 1988), and it quantifies a subjective workload by a weighted average of ratings on six dimensions, such as mental demand (MD), physical demand (PD), temporal demand (TD), performance (PE), effort (EF), and frustration (FR) (NASA 2009). To this end, the evaluators are asked to identify the relative weights of six dimensions about the workload of a given task based on their knowledge and experience. Then, the evaluators are asked to assess subjective scores about six dimensions using an arbitrary scale ranging from 0 to 100, which represent the level of subjective workload they felt in the course of performing the required task. Finally, based on the relative weights and subjective ratings, the overall workload can be quantified by their weighted average:

$$\mathrm{NASA-TLX} = \mathrm{a}_1 \times \mathrm{MD} + \mathrm{a}_2 \times \mathrm{PD} + \mathrm{a}_3 \times \mathrm{TD} + \mathrm{a}_4 \times \mathrm{PE} + \mathrm{a}_5 \times \mathrm{EF} + \mathrm{a}_6 \times \mathrm{FR}$$

where a_i ($i = 1, \ldots, 6$) denotes the relative weight

However, since evaluators have to follow a quite tricky process to determine relative weights (Hart and Staveland 1988), an equally weighted average has been suggested as an alternative method, such as $\mathrm{a}_i = 1/6$ (Nygren 1991).

9.2.2 Gathering Subjective Workload Scores

In order to gather subjective workload scores pertaining to the performance of emergency tasks, SROs working in the MCR of the reference NPPs were chosen, for two reasons. First, it is reasonable to assume that most of the burden that may arise in the course of performing emergency tasks will be loaded on the SRO of each operating team, because the SRO is responsible for the performance of emergency tasks (Moray 1999; Reinartz and Reinartz 1992). As outlined in Sect. 5.5, most of the actions included in emergency tasks should be carried out by the command as well as the confirmation of SROs. Under this operation scheme, it seems to be less meaningful to consider the subjective workload of board operators (i.e., ROs, TOs, and EOs).

Second, it should be emphasized that SROs have sufficient experience with emergency tasks prescribed in EOPs owing to regular retraining (for a period of about 6 months) for various kinds of initiating conditions. In other words, since the NASA–TLX technique quantifies a subjective workload based on personal experience with a given task to be evaluated, it is essential to select qualified operators who are familiar with the performance of emergency tasks. From these concerns, in total 18 SROs were asked to rate 6 dimensions about 23 emergency tasks that had been selected from the EOPs of reference NPPs. Table 9.1 summarizes the list of selected emergency tasks.

Table 9.1 Emergency tasks selected from the reference NPPs (Park and Jung 2006, © IEEE)

ID	Corresponding EOP	Procedural step		Remark
		Start	End	
1	ESDE (excess steam demand event)	4.0	5.0	–
2	LOCA (loss of coolant accident)	6.0	7.0	Group A
3	ESDE	7.0	8.0	Group A
4	ESDE	13.0	16.0	Group B
5	ESDE	17.0	18.0	–
6	SGTR	6.0	7.0	Group A
7	ESDE	24.0	28.0	–
8	ESDE	29.0	30.0	–
9	SGTR	8.0	10.0	–
10	SGTR	11.0	14.0	–
11	LOCA	11.0	13.0	–
12	LOCA	21.0	24.0	Group B
13	LOCA	15.0	19.0	–
14	ESDE	37.0	38.0	Group C
15	LOOP (loss of off-site power)	3.0	4.0	–
16	SGTR	15.0	18.0	Group B
17	LOOP	8.0	13.0	–
18	LOCA	27.0	28.0	Group C
19	LOAF (loss of all feed water)	5.0	10.0	–
20	LOAF	11.0	16.0	–
21	SBO (station blackout)	4.0	6.0	–
22	SBO	7.0	13.0	–
23	SBO	14.0	18.0	–

In Table 9.1, *Start* and *End* in the *Procedural step* column refer to procedural steps that denote, respectively, the commencement and the accomplishment of a given emergency task. For example, the first task is started from the fourth procedural step of the ESDE procedure, and then completed when the performance of the fifth procedural step has been finished. It is to be noted that the meaning of the three groups in the *Remark* column of Table 9.1 will be explained later.

On the basis of the selected emergency tasks, eight tasks were assigned to each SRO by the following sequence: (1) three emergency tasks belonging to *Groups A, B, and C* were evenly assigned and (2) the remaining emergency tasks not belonging to the three groups were randomly assigned. Table 9.2 summarizes the emergency tasks assigned to each SRO.

Table 9.2 Emergency tasks assigned to each SRO (Park and Jung 2006, © IEEE)

SRO ID	Task ID about 8 tasks assigned to each SRO							
1	3	4	9	11	13	14	17	23
2	1	3	4	5	8	18	20	23
3	1	3	9	12	14	19	22	23
4	3	7	9	12	15	18	19	22
5	3	5	8	14	15	16	17	20
6	1	3	9	15	16	18	20	23
7	2	4	8	11	13	14	15	21
8	2	4	7	10	11	13	18	23
9	2	5	7	10	12	14	15	19
10	2	8	9	10	12	13	18	22
11	1	2	5	11	14	16	17	21
12	2	5	10	16	18	19	21	23
13	4	6	7	10	13	14	17	20
14	1	4	5	6	8	18	20	21
15	6	10	12	14	17	19	21	22
16	1	6	7	9	12	17	18	22
17	6	8	11	14	15	16	19	22
18	6	7	11	13	16	18	20	21

Then, SROs gave subjective scores on six dimensions, which represent the amplitude of the workload they felt in the course of performing the assigned emergency tasks. Consequently, Table 9.3 shows subjective workload scores with the associated emergency tasks. It is to be noted subjective workload scores appearing in the each row of Table 9.3 indicate all the NASA–TLX scores given by SROs who were asked to assess emergency tasks. Accordingly, since a total of nine SROs participated in the evaluation of the 14th and 18th emergency tasks (refer to *Group C* in Table 9.1), those tasks have two more NASA–TLX scores than the others. In addition, *Average* represents the mean value of NASA–TLX scores for a given emergency task.

Table 9.3 Summary of subjective workload scores (Park and Jung 2006, © IEEE)

Task ID	Average	Subjective workload score								
1	38.1	34.2	69.2	29.2	40.0	35.0	20.8	–	–	–
2	41.3	51.7	46.7	38.3	43.3	29.2	38.3	–	–	–
3	44.7	55.0	31.7	58.3	43.3	51.7	28.3	–	–	–
4	45.6	48.3	35.0	55.0	50.0	40.0	45.0	–	–	–
5	46.3	41.7	56.7	47.5	43.3	35.0	53.3	–	–	–
6	38.8	40.0	41.7	44.2	30.0	43.3	33.3	–	–	–
7	53.9	49.2	62.5	63.3	48.3	55.0	45.0	–	–	–
8	52.2	60.0	35.0	55.0	65.8	38.3	59.2	–	–	–
9	55.0	65.0	71.7	53.3	30.0	48.3	61.7	–	–	–
10	54.6	63.3	54.2	50.0	41.7	61.7	56.7	–	–	–
11	52.9	45.0	37.5	63.3	55.0	55.0	61.7	–	–	–
12	43.1	60.0	38.3	38.3	41.7	42.5	37.5	–	–	–
13	48.6	44.2	51.7	60.8	43.3	51.7	40.0	–	–	–
14	53.9	58.3	69.2	26.7	65.0	43.3	61.7	56.7	45.8	58.3
15	47.9	61.7	24.2	60.0	30.8	56.7	54.2	–	–	–
16	39.5	48.3	24.2	36.7	35.0	58.3	34.2	–	–	–
17	47.1	45.0	51.7	55.0	43.3	27.5	60.0	–	–	–
18	48.8	36.7	28.3	55.0	65.0	45.0	46.7	62.5	58.3	41.7
19	55.7	61.7	67.5	40.0	57.5	40.0	67.5	–	–	–
20	49.4	45.8	46.7	58.3	30.8	55.0	60.0	–	–	–
21	63.7	35.8	65.0	73.3	55.0	82.5	70.8	–	–	–
22	61.3	65.0	79.2	58.3	51.7	70.0	43.3	–	–	–
23	51.0	56.7	42.5	66.7	38.3	51.7	50.0	–	–	–

9.2.3 Reliability of Subjective Workload Scores

As summarized in Table 9.3, NASA–TLX scores on 23 emergency tasks have been successfully obtained. However, before comparing NASA–TLX scores with the associated TACOM scores, it is essential to check their reliability. In this regard, it is necessary to consider two aspects related to the reliability of subjective ratings – *consistency* and *reproducibility*.

First, the consistency (or the agreement) of NASA–TLX scores should be clarified because SROs' ratings on six dimensions could be changed for various reasons, such as aptitude or personality, for example. In other words, if SROs' ratings fluctuate due to factors besides the performance of emergency tasks, the reliability of NASA–TLX scores would be questionable. From this concern, an intraclass correlation (ICC) coefficient was used to confirm the consistency of SROs' ratings (Bartko 1966; Bartko 1976).

The ICC coefficient ranges from $-\infty$ to 1, and the level of consistency increases with increases in the ICC coefficient. Accordingly, one indicates perfect consistency, while a negative value of the ICC coefficient denotes that subjective ratings are unreliable because of the lack of consistency. Table 9.4 summarizes the classes of ICC coefficients that have been frequently adopted as a basis for determining the consistency level of subjective ratings (Landis and Koch 1977).

Table 9.4 Levels of consistency of subjective ratings

Level of consistency	Corresponding ICC coefficient
Poor	Negative value
Slight	0 to 0.2
Fair	0.21 to 0.4
Moderate	0.41 to 0.6
Substantial	0.61 to 0.8
Almost perfect	0.81 to 1.0

In addition, the result of existing studies found that subjective ratings would be consistent when their ICC coefficient locates at least in the moderate level (Landis and Koch 1977; Marinus et al. 2004). Consequently, 0.41 is used as the threshold value from which the consistency of NASA–TLX scores can be determined. As a result, Table 9.5 summarizes TACOM scores as well as the associated NASA–TLX scores with the ICC coefficients of all the emergency tasks. It is to be noted that a strikethrough in Table 9.5 indicates an emergency task having an unreliable NASA–TLX score**.**

Second, the reproducibility (or repeatability) of NASA–TLX scores should be considered in order to confirm the reliability of subjective ratings (Bruton et al. 2000; Levy et al. 1999). In other words, even if there is consistency, if SROs assigned different scores to the same emergency tasks, then it may be difficult to use the collected NASA–TLX scores as the reference data to validate the appropriateness of the TACOM measure. Therefore, in order to clarify the reproducibility, it is necessary to internally compare NASA–TLX scores of the same emergency tasks. To this end, three groups of emergency tasks are selected and then randomly assigned to SROs, as noted in Table 9.2 (i.e., *Groups A, B, and C*).

For example, let us consider the second, third, and sixth emergency tasks in Table 9.1, which belong to *Group A*. Here, the goal of the sixth emergency task is *checking the necessity of stopping RCPs*, which consists of two procedural steps prescribed in the SGTR procedure, as illustrated in Fig. 5.5. The interesting point is that, in order to accomplish the same goal, identical procedural steps are also stipulated in both a LOCA (i.e., the second emergency task) and an ESDE procedure (i.e., the third emergency task).

Table 9.5 TACOM scores, NASA–TLX scores, and ICC coefficients

Task ID	TS	TR	TU	TACOM	Average	ICC
~~1~~	~~4.688~~	~~2.506~~	~~5.012~~	~~4.321~~	~~38.10~~	~~0.33~~
2	4.868	2.160	3.784	4.223	41.25	0.77
3	4.868	2.160	3.784	4.223	44.73	0.41
4	4.841	2.526	5.223	4.461	45.57	0.50
5	4.586	1.765	6.393	4.419	46.30	0.51
6	4.868	2.160	3.784	4.223	38.73	0.49
7	5.973	2.757	6.624	5.488	53.90	0.48
8	5.481	2.471	5.306	4.905	52.20	0.41
~~9~~	~~5.711~~	~~2.792~~	~~6.515~~	~~5.297~~	~~55.00~~	~~0.37~~
10	6.089	2.407	6.355	5.483	54.58	0.53
~~11~~	~~5.293~~	~~2.708~~	~~4.884~~	~~4.742~~	~~52.92~~	~~0.39~~
12	4.841	2.526	5.223	4.461	39.43	0.53
13	5.502	2.494	6.442	5.108	48.61	0.47
14	5.881	2.235	6.731	5.386	53.85	0.44
~~15~~	~~5.387~~	~~2.645~~	~~3.889~~	~~4.670~~	~~47.92~~	~~0.33~~
16	4.841	2.526	5.223	4.461	43.08	0.42
17	5.717	2.403	7.083	5.357	47.08	0.46
18	5.881	2.235	6.731	5.386	48.78	0.43
~~19~~	~~5.871~~	~~2.854~~	~~6.204~~	~~5.361~~	~~55.69~~	~~0.38~~
~~20~~	~~6.064~~	~~2.392~~	~~7.026~~	~~5.578~~	~~49.44~~	~~0.38~~
~~21~~	~~4.768~~	~~2.021~~	~~3.866~~	~~4.145~~	~~63.75~~	~~0.38~~
22	5.727	2.675	6.091	5.222	61.25	0.46
23	5.120	2.473	5.266	4.650	50.97	0.42

This means that the reproducibility can be investigated by comparing whether or not SROs give similar NASA–TLX scores to the same emergency tasks. Based on this concern, Table 9.6 shows the results of one-way ANOVA conducted for three groups of emergency tasks**.**

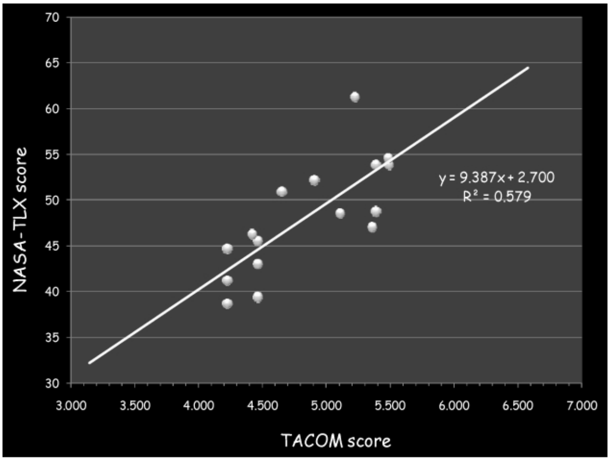
From Table 9.6 it seems to be evident that there is no significant difference among NASA–TLX scores for the three groups of emergency tasks. For example, the mean values of NASA–TLX scores for the three kinds of emergency tasks belonging to *Group A* are similar because their ANOVA result strongly indicates that the difference among NASA–TLX scores is due to random variability (i.e., p = 0.54). Similarly, the ANOVA results of other groups indicate that SROs have given similar NASA–TLX scores when they are asked to rate the same emergency tasks. Consequently, one could reasonably expect reproducibility of NASA–TLX scores.

Table 9.6 ANOVA results of three groups of emergency tasks (Park and Jung 2006, © IEEE)

Group	Task ID	Corresponding NASA –TLX score rated by SROs									p^*
A	2	51.7	46.7	38.3	43.3	29.2	38.3	–	–	–	0.54
	3	55.0	31.7	58.3	43.3	51.7	28.3	–	–	–	
	6	40.0	41.7	44.2	30.0	43.3	33.3	–	–	–	
B	4	48.3	35.0	55.0	50.0	40.0	45.0	–	–	–	0.55
	12	60.0	38.3	38.3	41.7	42.5	37.5	–	–	–	
	16	48.3	24.2	36.7	35.0	58.3	34.2	–	–	–	
C	14	58.3	69.2	26.7	65.0	43.3	61.7	56.7	45.8	58.3	0.41
	18	36.7	28.3	55.0	65.0	45.0	46.7	62.5	58.3	41.7	

*Significance level

The above rationales uphold the notion that NASA–TLX scores are meaningful as the reference data by which the appropriateness of the TACOM measure can be established. For this reason, a linear regression analysis is conducted using the data summarized in Table 9.5. Figure 9.2 shows the results of a statistical analysis with ANOVA table.



ANOVA table

Item	Degree of freedom	Sum of squares	Mean square	F statistics
Model	1	326.498	326.498	19.207
Error	14	237.982	16.999	
Total	15	564.480		

$F_{0.05}(1, 14) = 4.600$
$p < 10^{-4}$

Residual analysis

- Residual mean: -9.770×10^{-15}
- Normality test: passed ($p = 0.842$)
- Constant variance test: passed ($p = 0.512$)

Fig. 9.2 Result of linear regression analysis – TACOM scores with associated NASA–TLX scores

Figure 9.2 shows a remarkable correlation between TACOM scores and the associated NASA–TLX scores. In addition, the ANOVA table elucidates that the variation in NASA–TLX scores is largely attributable to the variation in TACOM scores ($p < 10^{-4}$). Therefore, it is reasonable to say that the TACOM measure is meaningful for explaining subjective workload scores perceived by SROs.

9.3 Comparing Task Performance Time Data Obtained from Other NPPs

In studying human-performance-related issues, one of the important findings is that the performance of qualified operators (or unqualified operators) is predictable when they are carrying out tasks having similar complexities (Chater 2000; Feldman 2000; Hamilton and Clarke 2005; Johannsen et al. 1994; Johnson and Payne 1985; Ogawa 1993; Stassen et al. 1990; Stanton and Young 1999; Zandin 2003). From the point of view of proceduralized tasks, one plausible explanation of this finding is that procedures strongly affect the actual behavior of qualified operators by institutionalizing detailed instructions. In other words, since proceduralized tasks institutionalize what is to be done and how to do it, it is assumed that the performance of qualified operators is, to some extent, predictable. Actually, the results of existing studies have provided a theoretical as well as an empirical clue supporting the reasonability of this assumption (Hollnagel et al. 1999; Kim et al. 2003; Stanton and Baber 2005).

If we adopt this assumption, it is natural to expect that the appropriateness of the TACOM measure can be consolidated by comparing TACOM scores with task performance time data gathered from other NPPs. For the sake of convenience, it should be noted that NPPs from which task performance time data were additionally collected will henceforth be referred as the subsidiary reference NPPs.

Similar to the case of the reference NPPs, a full-scope simulator has been installed in the training center of the subsidiary reference NPPs. This simulator is designed based on the MCR of a PWR that has 950 MWe capacity with conventional control devices. In addition, qualified operators working in the MCR of the subsidiary reference NPPs must be regularly retrained in order to increase their skills or knowledge related to various operating conditions including emergencies. Therefore, it is possible to collect audiovisual records on emergency operations under SGTR conditions that were carried out by 6 MCR operating teams. This collection was conducted from April to August 2005, and as a result, averaged task performance time data on 9 distinctive emergency tasks were obtained. Table 9.7 summarizes averaged performance time data on emergency tasks with their associated TACOM scores.

Based on the task performance time data shown in Table 9.7, a direct comparison was conducted to clarify whether averaged task performance time data obtained from the subsidiary reference NPPs remained within a certain range predicted by those from the reference NPPs. Figure 9.3 depicts the results of this comparison.

Table 9.7 Averaged task performance time data with the associated TACOM scores that are collected from the subsidiary reference NPPs (Park and Jung 2008, © Elsevier)

ID	TS	TR	TU	TACOM	Avg.(s)[1]	SD(s)[2]
1	4.626	1.774	4.112	4.051	41.9	25.5
2	4.630	1.496	3.495	3.944	12.0	2.9
3	4.042	1.821	3.979	3.627	17.9	5.6
4	4.691	1.799	4.262	4.121	33.9	22.3
5	5.486	2.203	4.134	4.716	55.4	27.8
6	4.847	1.680	3.879	4.168	38.9	16.0
7	4.433	1.537	3.778	3.843	34.7	10.3
8	5.976	2.740	6.344	5.441	97.0	28.6
9	5.742	2.547	5.227	5.084	77.1	24.1

[1] Avg.(s) denotes the mean value of task performance time data for each emergency task

[2] SD: standard deviation

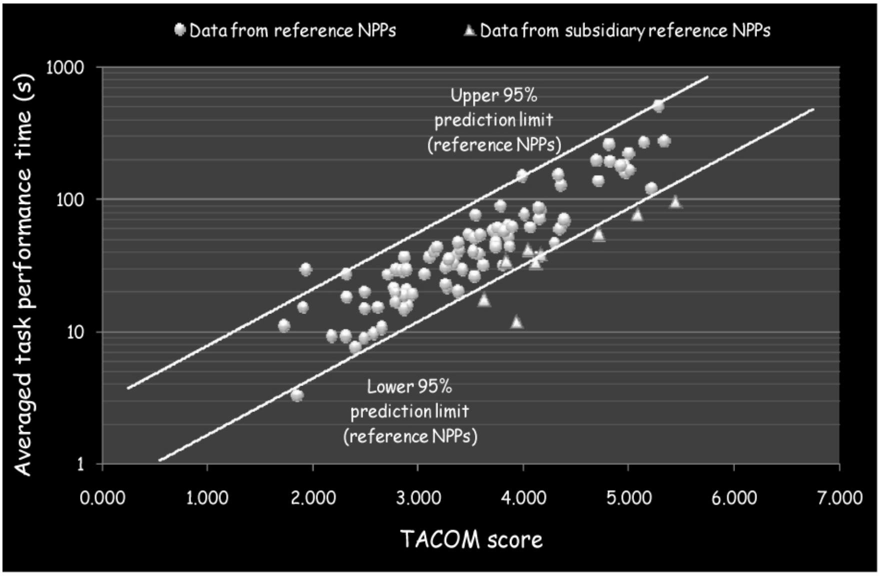


Fig. 9.3 Comparing two sets of task performance time data

In Fig. 9.3, there are two lines, *Upper 95% prediction limit* and *Lower 95% prediction limit*. Here, the meaning of the former is that, with a 95% confidence level, most of the averaged task performance time data obtained from the reference NPPs are expected to not exceed this limitation. Similarly, *Lower 95% prediction limit* indicates that, with a 95% confidence level, most of the averaged task performance time data will be greater than this limitation. Under these prediction limits, it is anticipated that two sets of task performance time data will be comparable with respect to TACOM scores because most of the task performance time data obtained from the subsidiary reference NPPs seem to be located near the lower prediction limit. In other words, although the contents of emergency tasks to be done by qualified operators working in the reference NPPs are quite different from those of the subsidiary reference NPPs, averaged task performance time data are predictable to some extent when the complexity score of a task (i.e., TACOM score) is given.

This expectation becomes more evident when averaged task performance time data obtained from the subsidiary reference NPPs are compared with those of the reference NPPs, which are obtained under similar conditions. Table 9.8 summarizes averaged task performance time data extracted from the OPERA database and collected under SGTR conditions of the reference NPPs. In addition, Fig. 9.4 depicts the results of these comparisons.

Table 9.8 Averaged task performance time data with the associated TACOM scores pertaining to the SGTR condition of the reference NPPs (Park and Jung 2008, © Elsevier)

ID	TS	TR	TU	TACOM	Avg.(s)	SD(s)
1	2.807	1.612	2.846	2.579	10.5	6.14
2	3.384	1.434	2.404	2.900	13.5	7.55
3	4.005	2.186	4.901	3.804	32.0	11.14
4	4.698	2.450	4.884	4.299	49.5	17.87
5	3.226	1.612	2.846	2.867	18.6	9.23
6	4.429	2.450	4.549	4.064	48.4	11.72
7	3.724	1.478	3.374	3.276	36.8	30.56
8	4.317	1.806	2.856	3.674	49.1	24.71
9	4.264	2.099	4.863	3.956	44.1	19.70
10	4.846	2.154	3.814	4.210	89.0	62.20
11	5.447	2.550	6.214	5.038	169	66.70
12	6.007	2.285	6.178	5.385	507	239.40

Figure 9.4 is very important for clarifying the appropriateness of the TACOM measure. According to Stassen et al. (1990), it was pointed out that human performance could be predictable if tasks are well defined. In addition, laboratory experiments have shown that the performance of human operators would be the

same if systems to be supervised had the same complexity, although the systems might differ in the number of functions and the degree of interactions (Wieringa and Stassen 1993). Therefore, the concept of an iso-complexity curve was suggested based on the number of functions and the degree of interactions (Johannsen et al. 1994; Visser and Wieringa 2001). This strongly suggests that, even though qualified operators have to accomplish different tasks, if there is a proper measure that can evaluate the complexity of a well-defined task, then their performance should not only be predictable but also be standardized as a function of a task complexity score. Subsequently, it is possible to say that the TACOM measure is meaningful for quantifying the complexity of a task to be done by qualified operators.

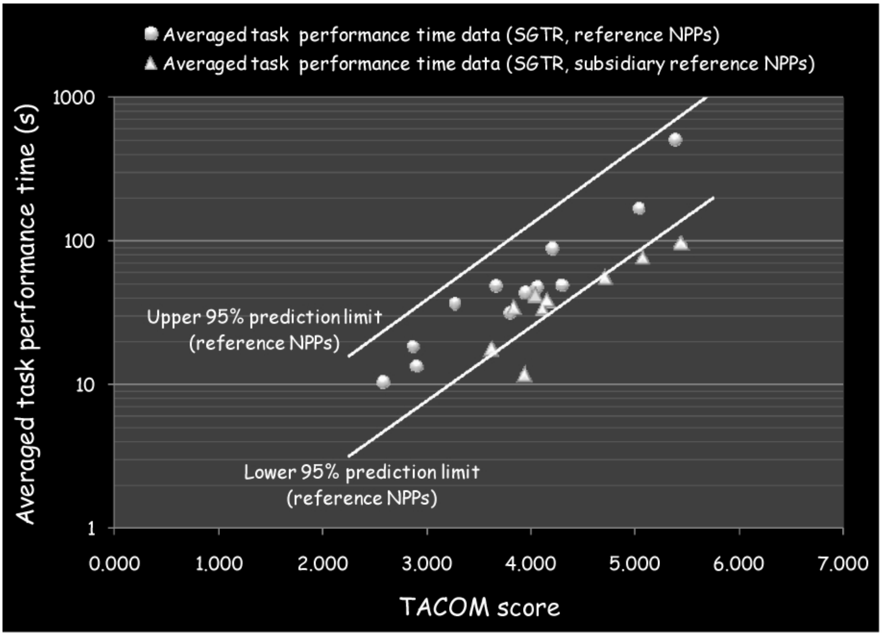


Fig. 9.4 Comparing two sets of averaged task performance time data collected under SGTR conditions

References

Bartko JJ (1966) The intraclass correlation coefficient as a measure of reliability. Psychol Rep 19:3–11

Bartko JJ (1976) On various intraclass correlation reliability coefficients. Psychol Bull 83:762–765

Bruton A, Conway JH, Holgate ST (2000) Reliability: what is it and how is it measured? Physiotherapy 86(2):94–99

Campbell DJ (1988) Task complexity: a review and analysis. Acad Manage Rev 13(1):40–52

Chater N (2000) The logic of human learning. Nature 407:572–573

Dickinson J, Byblow WD, Ryan LA (1993) Order effects and weighting process in workload assessment. Appl Ergonom 33(1):17–33

Feldman J (2000) Minimization of Boolean complexity in human concept learning. Nature 407:630–633

Hamilton WL, Clarke T (2005) Driver performance modelling and its practical application to railway safety. Appl Ergonom 36:661–670

Hancock PA (1996) Effects of control order, augmented feedback, input device and practice on tracking performance and perceived workload. Ergonomics 39(9):1146–1162

Hart SG, Staveland LE (1988) Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: Hancock PA, Meshkati N (eds) Human Mental Workload, Elsevier, Amsterdam, pp.139–183

Hendy KC, Hamilton KM, Landry LN (1993) Measuring subjective workload: when is one scale better than many? Hum Factors 35(4):579–601

Henneman RL, Rouse WB (1984) Measures of human problem solving performance in fault diagnosis tasks. IEEE Trans Syst Man Cybern 14:99–112

Hill SG, Iavecchia HP, Byers JC, Bittner, Jr., AC, Zaklad AL, Christ RE (1992) Comparison of four subjective workload rating scales. Hum Factors 34(4):429–439

Hollnagel E, Kaarstad M, Lee HC (1999) Error mode prediction. Ergonomics 42:1457–1471

Johannsen G, Levis AH, Stassen HG (1994) Theoretical problems in man-machine systems and their experimental validation. Automatica 30:217–231

Johnson EJ, Payne JW (1985) Effort and accuracy in choice. Manage Sci 31:395–414

Kim JH, Lee SJ, Seong PH (2003) Investigation on applicability of information theory to prediction of operator performance in diagnosis tasks at nuclear power plants. IEEE Trans Nuclear Sci 50:1238–1252

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174

Levy AS, Lintner S, Kenter K, Speer KP (1999) Intra- and interobserver reproducibility of the shoulder laxity examination. Am J Sport Med 27(4):460–463

Li K, Wieringa PA (2000) Understanding perceived complexity in human supervisory control. Cognit Technol Work 2:75–88

Liu Y, Wickens CD (1994) Mental workload and cognitive task automaticity: an evaluation of subjective and time estimation metrics. Ergonom 37(11):1843–1854

Marinus J, Visser M, Stiggelbout AM, Rabey JM, Martinez-Martin P, Bonuccelli U, Kraus PH, Hilten JJ (2004) A short scale for the assessment of motor impairments and disabilities in Parkinson's disease: the SPES/SCOPA. J Neurol Neurosurg Psychiatr 75:388–395

Maynard DC, Hakel MD (1997) Effects of objective and subjective task complexity on performance. Hum Perform 10(4):303–330

Moray N (1999) Advanced displays, cultural stereotypes and organizational characteristics of a control room. In: Misumi J, Wilpert M, Miller R (eds) Nuclear Safety: A Human Factors Perspective. Taylor & Francis, New York

NASA (2009) http://humansystems.arc.nasa.gov/groups/TLX/

Nygren TE (1991) Psychometric properties of subjective workload measurement techniques: implications for their use in the assessment of perceived mental workload. Hum Factors 33(1):17–33

Ogawa K(1993) A complexity measure of task content in information-input tasks. Int J Hum-Comput Interact 5(2):167–188

Park J, Jung W (2006) A study on the validity of a task complexity measure for emergency operating procedures of nuclear power plants – comparing with a subjective workload. IEEE Trans Nuclear Sci 53(5):2962– 2970

Park J, Jung W (2008) A study on the validity of a task complexity measure for emergency operating procedures of nuclear power plants – comparing task complexity scores with two sets of operator response time data obtained under a simulated SGTR. Reliabil Eng Syst Saf

93:557–566

Reinartz SJ, Reinartz G (1992) Verbal communication in collective control of simulated nuclear power plant incidents. Reliabil Eng Syst Saf 36:245–251

Stanton N, Young M (1999) What price ergonomics? Nature 399:197–198

Stanton NA, Baber C (2005) Validating task analysis for error identification: reliability and validity of a human error prediction technique. Ergonomics 48:1097–1113

Stassen HG, Johannsen G, Moray N (1990) Internal representation, internal model, human performance model and mental workload. Automatica 26(4):811–820

Svensson E, Angelbrog-Thandrez M, Sjoberg L, Olsson S (1997) Information complexity: mental workload and performance in combat aircraft. Ergonomics 40:362–380

Vidulich MA, Tsang PS (1986) Technique of subjective workload assessment: a comparison of SWAT and the NASA-Bipolar methods. Ergonomics 29(11):1385–1398

Visser M, Wieringa PA (2001) PREHEP: Human error probability based process unit selection. IEEE Trans Syst Man Cybern C Appl Rev 31(1):1–15

Wei ZG, Macwan AP, Wieringa PA (1998) A quantitative measure for degree of automation and its relation to system performance and mental load. Hum Factors 40(2):277–295

Wieringa PA, Stassen HG (1993) Assessment of complexity. In: Wise JA, Hopkin VD, Stager P (eds) Verification and validation of complex systems: Human Factors Issues, Springer, Berlin, Heiddelberg, New York, pp.173–180

Zandin KB (2003) MOST Work Measurement Systems, 3rd edn. Marcel Dekker, New York