Development and reliability of the explicit professional oral communication observation tool to quantify the use of non-technical skills in healthcare

Peter F Kemper,[1] Inge van Noord,[1] Martine de Bruijne,[1] Dirk L Knol,[2] Cordula Wagner,[1,3] Cathy van Dyck[4]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/bmjqs-2012-001451).

[1]Department of Public and Occupational Health, EMGO Institute for Health and Care Research, VU University Medical Center, Amsterdam, The Netherlands
[2]Department of Epidemiology and Biostatistics, EMGO Institute for Health and Care Research, VU University Medical Center, Amsterdam, The Netherlands
[3]The Netherlands Institute of Health Services Research (NIVEL), Utrecht, The Netherlands
[4]Faculty of Social Sciences, Department of Organizational Science, VU University, Amsterdam, The Netherlands

Correspondence to
Peter F Kemper, Department of Public and Occupational Health, EMGO+ Institute for Health Care and Research, VU University Medical Center, Van der Boechorststraat 7, Amsterdam 1081 BT, The Netherlands; p.kemper@vumc.nl

Received 15 August 2012
Revised 28 December 2012
Accepted 16 January 2013
Published Online First
13 February 2013

To cite: Kemper PF, van Noord I, de Bruijne M, *et al*. *BMJ Qual Saf* 2013;**22**:586–595.

ABSTRACT

Background A lack of non-technical skills is increasingly recognised as an important underlying cause of adverse events in healthcare. The nature and number of things professionals communicate to each other can be perceived as a product of their use of non-technical skills. This paper describes the development and reliability of an instrument to measure and quantify the use of non-technical skills by direct observations of explicit professional oral communication (EPOC) in the clinical situation.

Methods In an iterative process we translated, tested and refined an existing checklist from the aviation industry, called self, human interaction, aircraft, procedures and environment, in the context of healthcare, notably emergency departments (ED) and intensive care units (ICU). The EPOC comprises six dimensions: assertiveness, working with others; task-oriented leadership; people-oriented leadership; situational awareness; planning and anticipation. Each dimension is specified into several concrete items reflecting verbal behaviours. The EPOC was evaluated in four ED and six ICU.

Results In the ED and ICU, respectively, 378 and 1144 individual and 51 and 68 contemporaneous observations of individual staff members were conducted. All EPOC dimensions occur frequently, apart from assertiveness, which was hardly observed. Intraclass correlations for the overall EPOC score ranged between 0.85 and 0.91 and for underlying EPOC dimensions between 0.53 and 0.95.

Conclusions The EPOC is a new instrument for evaluating the use of non-technical skills in healthcare, which is reliable in two highly different settings. By quantifying professional behaviour the instrument facilitates measurement of behavioural change over time. The results suggest that EPOC can also be translated to other settings.

BACKGROUND

A lack of non-technical skills is increasingly recognised as an important underlying cause of adverse events in healthcare.[1 2] Non-technical skills are ‘the cognitive, social and personal resource skills that complement technical skills and contribute to safe and efficient task performance’.[3] Examples of non-technical skills are task management, teamwork, situation awareness and leadership.[4 5]

It can be reasoned that the nature and number of things that professionals communicate to each other can be perceived as a product of their use of non-technical skills. Application of non-technical skills implies that tasks, situations, decisions and team roles are made more explicit. Task management, for instance, becomes more explicit when a physician discusses with a colleague who is responsible for a patient, rather than assuming this is implicitly clear. Or when transport of a patient is standardised, it can be expected that abnormalities will be proactively managed instead of troubleshooting along the way.

In order to improve patient safety through the better use of non-technical skills, dedicated training is required,[6 7] such as crew resource management (CRM),[8] which is increasingly being applied in healthcare.[9] The number of evaluations of this training and the corresponding use of non-technical skills is also increasing rapidly, with results being promising but still limited.[10] Classroom-based training has shown mixed results with regard to behavioural change.[11]

A possible explanation for these mixed results might be that non-technical skills are difficult to measure. Non-technical skills are broad concepts that capture a wide range of aspects that can be relevant, depending on the situation. Furthermore, most non-technical skills are automatic and consist of routine behaviour, of which people have no realistic perception regarding the extent to which they use them. Most studies rely on self-reported questionnaires to measure non-technical skills,[11] or use proxy measures such as incident reporting[12] and adherence to guidelines.[13] Although these outcomes are relevant, they are not a measure of the actual demonstration of a non-technical skill.

Probably the best way to measure non-technical skills is by systematic direct behavioural observation because observations have the advantage of measuring behaviour as it actually occurs. There are several existing structured observation methods that assess the use of non-technical skills in healthcare.[3] [14–16] Most of these methods are setting-specific (eg, the operating theatre or anaesthetics) and demand clinical knowledge to assess the use of non-technical skills. Moreover, all these instruments appraise the use of non-technical skills, which, although highly structured, is a subjective assessment. Up to now, most studies have used the assessment of non-technical skills by direct observations in descriptive studies, for example as an educational feedback tool during a training session. Only a few studies have applied observations in evaluation studies.[17] [18] There is a need for an observation method that can be used independent of context by observers without or with limited clinical expertise, and that systematically quantifies non-technical skills rather than appraises them. Therefore, building on the shoulders of our predecessors, we set out to develop this method.

Our starting point has its origin in aviation, in which training of non-technical skills by means of CRM training was widely established by the mid-1990s in airlines across Europe and North America.[8] To give aircraft personnel structured feedback regarding their non-technical skills during CRM training, Antersijn and Verhoef[19] developed a checklist of important non-technical skills for the staff of the Royal Dutch Airlines. This checklist, called SHAPE, structured non-technical skills into five domains, notably self, human interaction, aircraft, procedures and environment. The domains aircraft (A) and procedures (P) are bound to the aviation context. In healthcare, the aircraft-specific non-technical skills should be replaced by department-specific clinical skills, although we did not use the specific clinical skills in this study.

Within the other domains (S, H and E), we can distinguish specific and general non-technical skills. The general non-technical skills in SHAPE are context independent and are applicable to different settings and situations without specific knowledge of the situation. For instance, a person can be assertive in the cockpit as well as on the hospital ward, which means that these general non-technical skills of the SHE domains can be translated to the healthcare context.

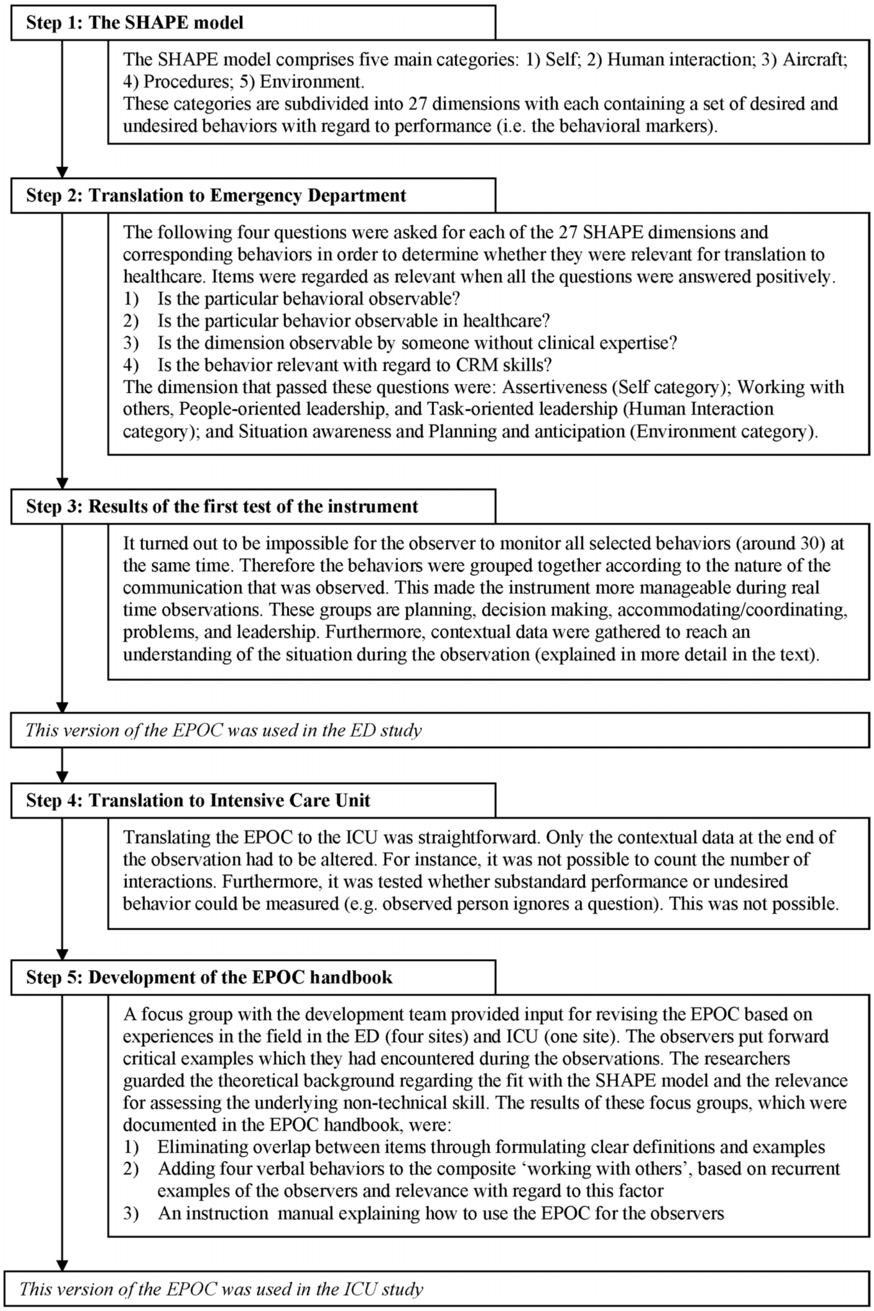
Like most of the existing observational instruments in healthcare, SHAPE uses behavioural markers, defined as ‘observable non-technical behaviours that contribute to superior or substandard performance within a work environment’.[20] Normally, these behavioural markers are used to appraise the use of a non-technical skill. Within a well-structured and complete framework, non-technical skills can also be quantified by counting the number of times a behavioural marker is explicitly expressed. We used the general non-technical skills defined in SHAPE to quantify non-technical skills by systematically observing professional communication on the work floor. The present paper describes the development of this new observation instrument, called the explicit professional oral communication (EPOC) measurement. We also present the interobserver variability in two different settings—the emergency department (ED) and the intensive care unit (ICU).

METHODS

Development of the EPOC measurement

The development of the EPOC comprised five steps (see figure 1). We started with the SHE domains of the original SHAPE. In an iterative process we translated, tested and refined the instrument in the healthcare context, first to ED and later to ICU. Decisions within the developmental steps and the progression through these steps were made by the development team. This team consisted of four researchers (PFK, IvN, MdB, CvK) with a background in medicine, epidemiology and psychology. In addition, various international experts in the field of non-technical skills were consulted on invitation and during international conferences. Furthermore, experienced EPOC observers were included in the team after the first studies ended.

Within these three categories, the final version of EPOC consists of six dimensions to classify explicitly the professional oral communication of an observed person. The self category of EPOC measures assertiveness. The human interaction category is divided into working with others, task-oriented leadership, and people-oriented leadership. The environment category consists of situation awareness, and planning and anticipation. Each dimension is subdivided into several concrete verbal behaviours that together represent the dimension. Table 1 displays the categorisation of EPOC and provides definitions for the categories



Step 1: The SHAPE model

The SHAPE model comprises five main categories: 1) Self; 2) Human interaction; 3) Aircraft; 4) Procedures; 5) Environment.
These categories are subdivided into 27 dimensions with each containing a set of desired and undesired behaviors with regard to performance (i.e. the behavioral markers).

Step 2: Translation to Emergency Department

The following four questions were asked for each of the 27 SHAPE dimensions and corresponding behaviors in order to determine whether they were relevant for translation to healthcare. Items were regarded as relevant when all the questions were answered positively.

1) Is the particular behavioral observable?
2) Is the particular behavior observable in healthcare?
3) Is the dimension observable by someone without clinical expertise?
4) Is the behavior relevant with regard to CRM skills?

The dimension that passed these questions were: Assertiveness (Self category); Working with others, People-oriented leadership, and Task-oriented leadership (Human Interaction category); and Situation awareness and Planning and anticipation (Environment category).

Step 3: Results of the first test of the instrument

It turned out to be impossible for the observer to monitor all selected behaviors (around 30) at the same time. Therefore the behaviors were grouped together according to the nature of the communication that was observed. This made the instrument more manageable during real time observations. These groups are planning, decision making, accommodating/coordinating, problems, and leadership. Furthermore, contextual data were gathered to reach an understanding of the situation during the observation (explained in more detail in the text).

This version of the EPOC was used in the ED study

Step 4: Translation to Intensive Care Unit

Translating the EPOC to the ICU was straightforward. Only the contextual data at the end of the observation had to be altered. For instance, it was not possible to count the number of interactions. Furthermore, it was tested whether substandard performance or undesired behavior could be measured (e.g. observed person ignores a question). This was not possible.

Step 5: Development of the EPOC handbook

A focus group with the development team provided input for revising the EPOC based on experiences in the field in the ED (four sites) and ICU (one site). The observers put forward critical examples which they had encountered during the observations. The researchers guarded the theoretical background regarding the fit with the SHAPE model and the relevance for assessing the underlying non-technical skill. The results of these focus groups, which were documented in the EPOC handbook, were:

1) Eliminating overlap between items through formulating clear definitions and examples
2) Adding four verbal behaviors to the composite ‘working with others’, based on recurrent examples of the observers and relevance with regard to this factor
3) An instruction manual explaining how to use the EPOC for the observers

This version of the EPOC was used in the ICU study

Figure 1 Chronological display of the development of the explicit professional oral communication. CRM, crew resource management; ED, emergency department; EPOC, explicit professional oral communication; ICU, intensive care unit.

and dimensions, and examples of verbal behaviours. The instrument is described in a handbook with instructions, definitions and examples.

Measurement: observation

Assessing EPOC meant that only work-related interactions between professionals were counted. Every time the observed person expressed one of the verbal behaviours of the EPOC the observer had to tally this on the observation form. One exception was made for ‘listens’. When someone nods his or her head, this was also tallied. Social talk or conversations with the patients or family were not included.

An observation lasted 30 min. During an observation one person at work was observed and his or her work-related verbal expressions were tallied. An observation was carried out by one observer (an individual observation) or, in order to calculate the inter-observer reliability, by two independent observers simultaneously (a contemporaneous observation). The observations were carried out directly and were not recorded on video. All observations were conducted during daily practice between 07:00 and 19:00 hours.

In addition to the verbal behaviours, contextual information was gathered before, during and after the observation. Contextual information consisted of the

Table 1 Overview of categories, dimensions and items, with definitions and examples and the descriptive results

		ED				ICU			
		Baseline (n=179)		Follow-up (n=199)		Baseline (n=179)		Follow-up (n=199)	
Categories, dimensions and items	Definitions of the categories and dimensions and examples of the items	N	%	N	%	N	%	N	%
Overall EPOC score	All verbal expressions taken together	1439	100	1887	100	22979	100	20854	100
Self category	Expressing non-technical skills in relation to oneself	11	0.74	4	0.21	136	0.59	210	1.01
1. Assertiveness	Takes action on his/her own accord; stands up for him/herself								
Expresses concerns	'I'm not sure that this is going to end well'	4	0.27	2	0.11	97	0.42	197	0.94
Speaks up even when faced with resistance	'I already said that I was too fatigued for these tasks'	7	0.47	2	0.11	29	0.13	13	0.06
Speaks up aggressively (–)*	'I am fed up with it!'	1	0.07	0	0.00	10	0.04	0	0
Neglects others (–)*		1	0.07	0	0.00				
Human interaction category	Expressing non-technical skills in relation to others	1300	87.07	1760	93.27	21637	94.16	18746	89.89
2. Working with others	Takes initiative and remains an active part of the team; interacts with others	726	48.63	1062	56.28	18639	81.11	15699	75.28
Reacts to suggestions from others†	'Yes, you are right'					2317	10.08	2393	11.48
Asks others for contributions	'Does the medicine cabinet need refilling?'	24	1.61	12	0.64	2213	9.63	2055	9.85
Gives suggestions	'Shall we turn her on her left side?'	52	3.48	38	2.01	2441	10.62	2112	10.13
Confirms task	'Yes, I will do that'	196	13.13	357	18.92	2390	10.40	1198	5.74
Repeats task†	'Yes, I will lower the bed'					188	0.82	204	0.98
Asks for confirmation about information†	'Did you understand what I just said?'					457	1.99	453	2.17
Asks for confirmation about tasks	'Do you know what you have to do?'	238	15.94	465	24.64	340	1.48	266	1.28
Tells others what he/she is going to do	'I am going to prepare patient X for transfer'	144	9.65	142	7.53	1106	4.84	1150	5.51
Asks 'follow-up' questions for clarification†	'What do you mean by that?'					1031	4.49	801	3.84
Gives contributions†	Saying something without an explicit cause, like a question					2819	12.27	2367	11.35
Answers a question†	In answer to a question about the result of a test: 'It was negative'					3058	13.31	2253	10.8
Checks the correctness of the information	'So the patient received 200 ml you said?'	60	4.02	41	2.17	279	1.21	447	2.14
Demonstrates self-control in the case of errors		0	0.00	2	0.11				
Takes part in decision making		12	0.80	5	0.26				
3. Task-oriented leadership	Plan and organise the crew's tasks	151	10.11	83	4.40	1576	6.86	1193	5.72
Coordinates tasks	'You are going to wash this patient'	13	0.87	3	0.16	717	3.12	416	1.99
Checks and corrects tasks	'I said 5 ml, but you gave 50 ml'	66	4.42	27	1.43	295	1.28	337	1.62
Uses authority	'I'm the physician, so I will make that call'	0	0.00	0	0.00	5	0.02	3	0.01

Continued

Table 1 Continued

		ED				ICU			
		Baseline (n=179)		Follow-up (n=199)		Baseline (n=179)		Follow-up (n=199)	
Categories, dimensions and items	Definitions of the categories and dimensions and examples of the items	N	%	N	%	N	%	N	%
Gives instructions, coaches	'It is more comfortable for the patient if you do it this way'	72	4.82	53	2.81	559	2.43	437	2.10
4. People-oriented leadership	Motivate and encourage crew cooperation for performing tasks	423	28.33	615	32.59	1422	6.19	1854	8.89
Provides support and shows appreciation	'You have done this perfectly!'	21	1.41	13	0.69	336	1.46	425	2.04
Gives others space	A senior to a junior: 'Do you want to add anything else?'	1	0.07	1	0.05	38	0.17	64	0.31
Shows that one is listening	When someone else is talking: 'Yes'	394	26.39	599	31.74	1048	4.56	1365	6.55
Appreciate suggestions		7	0.47	2	0.11				
Environment category	Expressing non-technical skills in relation to the environment and the situation	182	12.19	123	6.52	1206	5.25	1898	9.10
5. Situation awareness	Be aware of the operational situation based on relevant factors	84	5.63	58	3.07	99	0.43	296	1.42
Names environmental factors that influence the situation	'The constant ringing of the phone distracts me'	6	0.40	0	0.00	57	0.25	148	0.71
Takes action based on these environmental factors	'I'm leaving the receiver off the hook'	0	0.00	0	0.00	4	0.02	4	0.02
Keeps an overview of all the patients	While treating patient X: 'How is patient Y doing?'	78	5.22	58	3.07	38	0.17	144	0.69
6. Planning and anticipation	Plan and structure actions	98	6.56	65	3.44	1107	4.82	1602	7.68
Discusses expectations that can influence the situation	'It is likely that another patient will be admitted to the unit today'	12	0.80	6	0.32	141	0.61	232	1.11
Specifies goal	'The aim is to transfer this patient to the ward today'	1	0.07	0	0.00	56	0.24	176	0.84
Indicates priorities	'There has to be a bed available before we can admit a new patient'	6	0.40	7	0.37	71	0.31	94	0.45
Indicates what has to be done	'This bed has to be lowered'	71	4.76	49	2.60	41	0.18	59	0.28
Adjusts the plan if necessary	'The aim was to use treatment X, but seeing the condition of this patient, treatment Y is better'	8	0.54	3	0.16	798	3.47	1041	4.99

Note: Contact author for the extensive EPOC handbook.
*Results of negative items are not part of the sum score of the dimensions and categories.
†Added verbal behaviours that were not part of the preliminary version of the EPOC that was used in the ED study.
ED, emergency department; EPOC, explicit professional oral communication; ICU, intensive care unit.

starting time and the occupation of the observed person, the type and number of patients the observed person saw, how many times and with whom the observed person interacted. Directly after observation, both the observer and observed person filled out the National Aeronautics and Space Administration (NASA) task load index (NASA TLX)[21] to measure the perceived workload during the observation period. Next, the observer indicated which tasks the observed person had performed, by means of a short description of the observation period and ticking a number of preselected tasks (eg, handover or multidisciplinary meeting).

All observers received a 1-day theoretical training course to enable them to learn the definitions of the verbal behaviours, practise with written examples and becoming familiar with the common sources of rating biases (eg, halo effect). Specific attention was paid to maintain their sensitivity to verbal behaviours that occur less frequently. This was followed by a 1-day practical training course in the clinic with an experienced observer who supervised the observations and discussed the outcomes afterwards. Contemporaneous observations were carried out regularly and discussed afterwards. To make sure that all observers rated behaviour in the same way and to check whether they were consistent during the whole data collection period, regular meetings were organised to discuss contemporaneous or doubtful examples. Furthermore, these meetings were used to receive feedback about the EPOC with regard to the further development of the instrument.

Evaluation of the EPOC: ED and ICU

The evaluation of the EPOC consisted of two parts. First, we examined the occurrence of EPOC items by assessing how many times each item of the EPOC was observed. Second, we determined the reliability of EPOC by assessing the interobserver reliability.

Data from two distinct studies that applied the EPOC were used for this evaluation, one conducted in four ED and the other in six ICU. Both studies assessed the effect of a medical team training in a controlled trial, comprising a baseline measurement and a follow-up measurement.[22] In the ED the first version of the EPOC was applied, whereas in the ICU the second, revised, version was used. The data sets of the ED and ICU were therefore examined separately. The measurements within each site were also studied separately, as it was not possible to recruit the same observers during the post-measurement as in the premeasurement in the ICU departments.

Statistical analysis

The descriptive results of all individual observations were analysed in order to determine the occurrence of the EPOC items. Three parameters were used to assess the interobserver reliability: the intraclass correlation coefficient (ICC), the SEM and the limits of agreement (LOA). Due to the comprehensive number of graphs that the analysis of the LOA creates, the results and discussion are described in online supplementary appendix A (available online only).

The ICC examines the proportion of the total variance that can be attributed to ‘true’ differences between observed persons. The ICC for a single measurement based on absolute agreement[23] was determined for each category and dimension. The ICC was derived from both the contemporaneous and individual observations, a method following from the work of Euser *et al*.[24] The restricted maximum likelihood method was used to estimate the variance components and the delta method was used to calculate the corresponding CI.[24] To minimise the influence of the observed person, it was made sure that only one observation per unique person was used in the analysis. As the observations were carried out in the context of a controlled trial, the observed persons (subjects) were nested in either the intervention or control unit, which is incorporated in the model as a fixed factor. This resulted in four variance components: (1) the subjects nested within the intervention or control unit; (2) the observer; (3) an interaction between the observer and the intervention or control unit; (4) the residual variance (error). The ICC was estimated dividing the variance of the subject by the total variance, as described by Molenberghs *et al*.[25]

The SEM was estimated by taking the square root of the sum of three components of variance, these being the observers, the interaction between intervention or control and observer, and the residue. The SEM can be considered as the estimation of the ‘noise’ of the EPOC.[26 27]

RESULTS

In the ED, 378 individual and 51 contemporaneous observations of individual staff members were conducted in two measurement periods during 240 h of observation. In the baseline measurement, on average 8.3 (range 1–30) verbal behaviours per 30 min observation were counted, compared to 9.5 (range 2–28) in the follow-up measurement. In both measurements the most frequent item was ‘shows that one is listening’ (respectively n=394 and n=599), representing approximately 29% of the observed behaviours. Items belonging to the category ‘self’ were infrequently observed (less than 1% of all observed behaviour in both measurements). Some of the EPOC items were never observed at all (eg, ‘uses authority’).

In the ICU, 1144 individual and 68 contemporaneous observations of individual staff members were conducted in two measurement periods during 640 h of observation. In the baseline measurement, on average 41 verbal behaviours per 30 min observation were counted (range 2–129), compared to 35.5 in the follow-up measurement (range 1–95). This is

Table 2 Overview of the ICC with its CI, the mean score per item and the SEM

	ED*				ICU							
	Follow-up				Baseline				Follow-up			
	ICC	(95% CI)	Mean	SEM†	ICC	(95% CI)	Mean	SEM†	ICC	(95% CI)	Mean	SEM†
Overall EPOC score	0.91	(0.84 to 0.95)	0.40	0.05	0.85	(0.74 to 0.92)	0.64	0.14	0.88	(0.76 to 0.94)	0.66	0.13
Self category	0.00	(−0.30 to 31)	0.01	0.07	0.41	(0.12 to 0.63)	0.08	0.16	0.09	(−0.12 to 0.28)	0.10	0.25
Assertiveness	Same as category ‘self’											
Human interaction	0.90	(0.80 to 0.95)	0.61	0.08	0.90	(0.78 to 0.96)	1.46	0.26	0.95	(0.91 to 0.97)	1.33	0.16
Working with others	0.85	(0.73 to 0.92)	0.69	0.13	0.87	(0.73 to 0.94)	2.86	0.51	0.84	(0.71 to 0.92)	2.29	0.44
Task-oriented leadership	0.84	(0.72 to 0.91)	0.14	0.10	0.84	(0.73 to 0.91)	0.66	0.40	0.74	(0.58 to 0.85)	0.53	0.36
People-oriented leadership	0.70	(0.52 to 0.82)	0.80	0.24	0.76	(0.62 to 0.86)	0.87	0.47	0.62	(0.35 to 0.80)	1.14	0.71
Environment	0.85	(0.75 to 0.91)	0.08	0.05	0.63	(0.42 to 0.77)	0.36	0.26	0.56	(0.35 to 0.72)	0.54	0.33
Situation awareness	0.77	(0.63 to 0.86)	0.10	0.09	0.85	(0.75 to 0.91)	0.05	0.07	0.53	(0.24 to 0.73)	0.16	0.24
Planning and anticipation	0.84	(0.73 to 0.90)	0.07	0.05	0.65	(0.45 to 0.79)	0.68	0.50	0.58	(0.37 to 0.73)	0.92	0.60

All results were calculated using the mean score per item of a category or dimension.
Note. The follow-up measurement of the ED comprised 57 unique individual observations and 43 contemporaneous observations, for the baseline ICU these were 274 and 33 and for the follow-up they were 309 and 35.
*Too few contemporaneous observations were conducted during the baseline measurement to enable the interobserver reliability characteristics to be calculated.
†The SEM is calculated using the following formula: $\mathrm{SEM}=\sqrt{(\mathrm{Variance}_{\mathrm{observer}}+\mathrm{Variance}_{\mathrm{interaction}}+\mathrm{Variance}_{\mathrm{error}})}$.
ED, emergency department; EPOC, explicit professional oral communication; ICC, intraclass correlation coefficient; ICU, intensive care unit.

approximately five times higher than the average of the ED. The most frequent item in the baseline measurement was 'answers a question' (n=3094), representing 13% of all observed verbal behaviours. The most frequent item in the follow-up measurement was 'reacts to suggestions from others' (n=2393), representing 11.5% of all observed verbal behaviours. There were no items that were never observed.

The ICC in the ED ranged from 0.70 to 0.91, and in the ICU from 0.53 to 0.95, with the self category as an exception in both settings (table 2). The graphs of the LOA (see supplementary appendix A, available online only) show that all measurements stay well within the LOA, although due to insufficient numbers for the ED at baseline the LOA could not be computed. The LOA are small, reflecting low variation in differences between the observers.

DISCUSSION

The EPOC is a new observational method for assessing non-technical skills through quantifying EPOC of healthcare professionals. We assessed the amount of explicit professional communication in two settings as well as the interobserver reliability. Our results show that some of the verbal behaviours and dimensions occur less often than others. It is plausible that these behaviours do indeed not arise very often, such as 'uses authority'. Some behaviours may take place more frequently after dedicated training, for instance, 'explicitly coordinating tasks with each other'. In addition, some concepts may occur but may be difficult to classify correctly due to close overlap with other concepts, such as 'expresses concerns' and 'gives suggestion'.

The results show good interobserver reliability for the EPOC. Although there is no consensus concerning what constitutes a good ICC,[28] the general convention is that ICC below 0.40 are poor, between 0.41 and 0.60 are moderate, and above 0.60 are good or even very good (>0.80).[29] Most categories and dimensions exceed 0.60. Interobserver reliability of the overall EPOC score, the human interaction category and its underlying dimensions, 'working with others' and 'task-oriented leadership' are very good in both studies. These findings indicate that the observers have been well trained and that the framework is comprehensive and clear. Furthermore, it means that the EPOC is solid for use in scientific research.

The self category, and its dimension 'assertiveness', has the lowest agreement. It can be argued that this category was observed too infrequently in the ED to calculate a valid ICC. During the ICU study, the self category was observed more often. However, the ICC for this dimension was also low in ICU, especially in the follow-up measurement. This suggests it is hard to assess this category reliably.

The follow-up measurement of the ICU study has somewhat lower ICC than the baseline measurement of this study. The environment category even has moderate ICC in the follow-up measurement. This difference is probably due to more formal and informal discussions about the definitions between the baseline observers, resulting in a higher mutual calibration. This signifies the need for an intensive and involving training of the observers, and to keep stimulating discussions about the application of EPOC with each other.

EPOC was applicable both in ED and ICU. Although the transfer of the EPOC from the ED to the ICU went very smoothly (see step 4 of figure 1), both settings differed significantly from each other in outcomes. The ED has overall a smaller CI range in ICC scores than both ICU measurements. A possible explanation for this finding is that the ED observations were conducted by two observers and in the ICU measurements a total of eight observers carried out observations. Another reason could be that the average amount of verbal behaviours per 30-min observation is almost five times higher in the ICU compared to the ED. This difference is probably due to the nature and organisation of work in both departments. In the ED, work processes are mainly organised along a chain of care. This chain starts with the triage and ends with the patient being referred to other providers or being sent home. Healthcare professionals in ED work sequentially rather than simultaneously. Providing care in the ICU is more of a team effort, with regular meetings to discuss the status of a patient. For adequate transfer of the EPOC across medical settings it is important to recognise such differences and details.

A major benefit of EPOC is that the explicit communication as a whole can be quantified. Experience with the EPOC showed that all professional communication during an observation can be classified along the verbal behaviours of the instrument. Moreover, it enables tracking differences in the sorts of professional communication. This is highly relevant when studying the effects of, for instance, a medical team training directed at improving communication, leadership and decision making.

Compared to existing instruments for observing non-technical skills,[3 15 30] the EPOC is distinctive as it quantifies general verbal behaviours rather than appraising context-specific behavioural markers that require clinical expertise. As general verbal behaviours are context independent and occur in every professional interaction, observing these skills does not require context-specific knowledge of the situation, such as clinical expertise. In addition, due to a minimum interpretation of what is being said, even complex situations can still be reliably observed. Quantifying the verbal behaviours makes it especially useful for evaluating changes in occurrence and patterns of non-technical skills.

When using EPOC as an instrument for evaluation, it should be noted that the expected effects of improving non-technical skills may fluctuate depending on the setting, previous training, motivation and so on. In the current context, for example, improving non-technical skills in the ED will probably result in more explicit communication, as there is very little verbal communication to begin with. Yet, for the ICU, in which a lot of communication between team members occurs round the clock, improving non-technical skills may rather change the content than the amount of verbal communication; for instance, to proactive planning instead of troubleshooting. This could even end in a decrease of verbal behaviours in the ICU, as communication becomes more efficient.

The difference in expected effects also emphasises that improving non-technical skills will not always result in more explicit communication. It has been proposed that there is an optimum after which the number of things being said damages the efficiency. For instance, Stachowski *et al*[31] showed that during a simulated crisis, fewer verbal statements were associated with high-performing teams in the control room of a nuclear power plant. In other words, the situation determines what effect can be expected and should be taken into account. Therefore the EPOC should first be adequately tested before the evaluation starts.

Limitations

A limitation of the EPOC is that it only assesses verbal communication, whereas non-verbal behaviour or things that should have been said are equally relevant. For instance, ignoring a question or purposefully turning your back on someone expresses more than can be said in words. During the development phase of the EPOC, several non-verbal behaviours were tested as part of the observation. However, as observing non-verbal communication is hard to standardise, these non-verbal items did not pass the testing phase.

The EPOC also has limitations related to observation schemes in general. Flin *et al*[32] summarise the boundaries of observational methods in three points. First, a classification of behaviour can never capture every aspect of performance. Second, important but infrequent behaviours are hard to measure once they occur. Third, to err is human also applies to observers. Observers can be distracted, fatigued or faced with too complex situations. An additional fourth caveat in line with the previous one is observer bias, which means that observers are more likely to find those things that they are looking for.

It may occur that the observed person is influenced by the observer, the so-called Hawthorne effect. In our experience this influence was marginal. Observed persons stated that they very quickly became used to the presence of the observers or even forgot that they were being observed at all.

Further research should explore other psychometric properties of this measurement, as described by Mokkink *et al.*[26] The level of reliability could be further increased by studying the test–retest reliability and internal consistency. The validity of the EPOC should also be explored. Studying the criterion validity of the EPOC should answer the question of what the optimum explicit communication is in a particular situation. Furthermore, attention could be paid to cross-validate the EPOC with a measure of non-technical performance. This should reveal to what degree the scores of the EPOC are consistent with changes in the use of non-technical skills (construct validity). In addition, the ability of the EPOC to measure changes in non-technical skills over time (responsiveness) should be assessed.

CONCLUSION

We developed a new instrument for evaluating the use of non-technical skills in healthcare, which is reliable in two highly different settings. By quantifying professional behaviour, our instrument facilitates the measurement of behavioural change over time. Our results suggest the EPOC can also be applied to other settings.

Acknowledgements The authors would like to thank the Royal Dutch Airlines for letting them use the SHAPE checklist and Patricia Antersijn for explaining the development of the SHAPE. Furthermore, they wish to express gratitude to the observers, who helped develop the EPOC by critically using and discussing it. The preliminary results of this study were presented in September 2010 during the Behavioral Science Applied to Acute Care Teams (BSAACT) meeting in Amsterdam. The authors would like to thank the organisation for this opportunity.

Contributors PFK drafted the final manuscript, all other authors (IvN, MdB, DLK, CW, CvD) read, revised and approved the manuscript. PK, IvN, MdB and CvD participated in the development team of the EPOC, as described in the manuscript, which conceived, designed and executed this study. CW participated in the design of the study and was regularly consulted by the development team. DLK helped with the methodological/statistical part of the present study. All authors read and approved the final manuscript.

Funding This study was partly funded by Zon-Mw, the Dutch Organisation for Health Research and Development.

Competing interests None.

Ethics approval This study received ethics approval from the ethical committee of the VU University Medical Centre.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Additional results have been published in a web-only appendix. Other results or data are available on request from the corresponding author.

REFERENCES

1 Helmreich RL. On error management: lessons from aviation. *BMJ* 2000;320:781–5.

2 Sevdalis N, Brett SJ. Improving care by understanding the way we work: human factors and behavioural science in the context of intensive care. *Crit Care* 2009;13:139.

3 Flin R, Patey R, Glavin R, *et al*. Anaesthetists' non-technical skills. *Br J Anaesth* 2010;105:38–44.

4 Reader T, Flin R, Lauche K, *et al*. Non-technical skills in the intensive care unit. *Br J Anaesth* 2006;96:551–9.

5 Yule S, Flin R, Paterson-Brown S, *et al*. Non-technical skills for surgeons in the operating room: a review of the literature. *Surgery* 2006;139:140–9.

6 Helmreich RL, Foushee HC. Why crew resource management? Empirical and theoretical bases of human factors training in aviation. In: Wiener EL, Kanki BG, Helmreich RL.eds. *Cockpit resource management*. San Francisco: Academic Press Inc, 1993: 3–45.

7 Catchpole K, Mishra A, Handa A, *et al*. Teamwork and error in the operating room: analysis of skills and roles. *Ann Surg* 2008;247:699–706.

8 Flin R, Maran N. Identifying and training non-technical skills for teams in acute medicine. *Qual Saf Health Care* 2004;13:i80–4.

9 McConaughey E. Crew resource management in healthcare: the evolution of teamwork training and MedTeams. *J Perinat Neonatal Nurs* 2008;22:96–104.

10 Ostergaard D, Dieckmann P, Lippert A. Simulation and CRM. *Best Pract Res Clin Anaesthesiol* 2011;25:239–49.

11 Rabol LI, Ostergaard D, Mogensen T. Outcomes of classroom-based team training interventions for multiprofessional hospital staff. A systematic review. *Qual Saf Health Care* 2010;19:e27.

12 Sax HC, Browne P, Mayewski RJ, *et al*. Can aviation-based team training elicit sustainable behavioral change? *Arch Surg* 2009;144:1133–7.

13 Halverson AL, Andersson JL, Anderson K, *et al*. Surgical team training: the Northwestern Memorial Hospital experience. *Arch Surg* 2009;144:107–12.

14 Undre S, Healey AN, Darzi A, *et al*. Observational assessment of surgical teamwork: a feasibility study. *World J Surg* 2006;30:1774–83.

15 Mishra A, Catchpole K, McCulloch P. The Oxford NOTECHS System: reliability and validity of a tool for measuring teamwork behaviour in the operating theatre. *Qual Saf Health Care* 2009;18:104–8.

16 Yule S, Flin R, Maran N, *et al*. Surgeons' non-technical skills in the operating room: reliability testing of the NOTSS behavior rating system. *World J Surg* 2008;32:548–56.

17 Weaver SJ, Rosen MA, DiazGranados D, *et al*. Does teamwork improve performance in the operating room? A multilevel evaluation. *Jt Comm J Qual Patient Saf* 2010;36:133–42.

18 McCulloch P, Mishra A, Handa A, *et al*. The effects of aviation-style non-technical skills training on technical performance and outcome in the operating theatre. *Qual Saf Health Care* 2009;18:109–15.

19 Antersijn PAM, Verhoef MC. Assessment of non-technical skills: is it possible? In: McDonald N, Johnston N, Fuller R. eds. *Applications of psychology to the aviation system: Proceedings of the 21st Conference of the European Association for Aviation Psychology (EAAP)*, Vol. 1. Aldershot, UK: Avebury Aviation, 1995: 243–50.

20 Klampfer B, Flin R, Helmreich R, *et al*. Group interactions in high risk environments: behavioural markers workshop. 2001. homepage.psy.utexas.edu/homepage/group/helmreichlab/publications/pubfiles/pub262.pdf (accessed 10 Mar 2012).

21 Hart SG, Staveland LE. Development of the NASA-TLX (task load index): results of empirical and theoretical research. In: Hancock A, Meshkati N.eds. *Human mental workload*. Amsterdam: North Holland Press, 1988.

22 Kemper PF, De Bruyne M, Van Dyck C, *et al*. Effectiveness of classroom based crew resource management training in the intensive care unit: study design of a controlled trial. *BMC Health Serv Res* 2011;11:304.

23 McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods* 1996;1:30–46.

24 Euser AM, Le Cessie S, Finken MJ, *et al*. Reliability studies can be designed more efficiently by using variance components estimates from different sources. *J Clin Epidemiol* 2007;60:1010–14.

25 Molenberghs G, Laenen A, Vangeneugden T. Estimating reliability and generalizability from hierarchical biomedical data. *J Biopharm Stat* 2007;17:595–627.

26 Mokkink LB, Terwee CB, Patrick DL, *et al*. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 2010;63:737–45.

27 De Vet HC, Terwee CB, Mokkink LB, *et al*. *Measurement in medicine*. Cambridge: Cambridge University Press, 2011.

28 Shrout PE. Measurement reliability and agreement in psychiatry. *Stat Methods Med Res* 1998;7:301–17.

29 Altman DG. *Practical statistics for medical research*. London: Chapman & Hall, 1991.

30 Yule S, Flin R, Paterson-Brown S, *et al*. Development of a rating system for surgeons' non-technical skills. *Med Educ* 2006;40:1098–104.

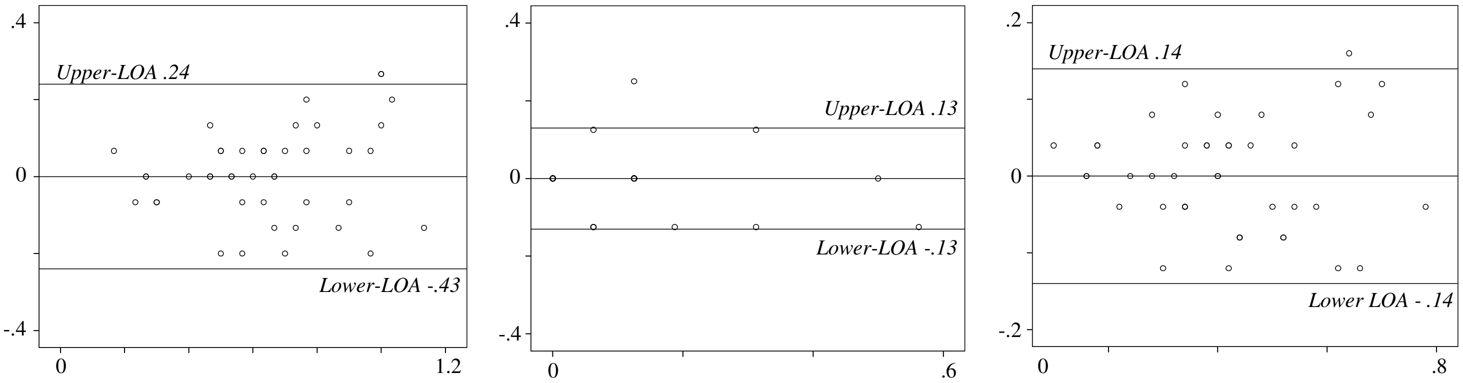
31 Stachowski AA, Kaplan SA, Waller MJ. The benefits of flexible team interaction during crises. *J Appl Psychol* 2009;94:1536–43.

32 Flin R, O'Conner P, Crichton M. *Safety at the sharp end: a guide to non-technical skills*. Farnham: Ashgate, 2008.

Web only appendix A: Text and figures that show the results of the analyses of the Limit of agreements.

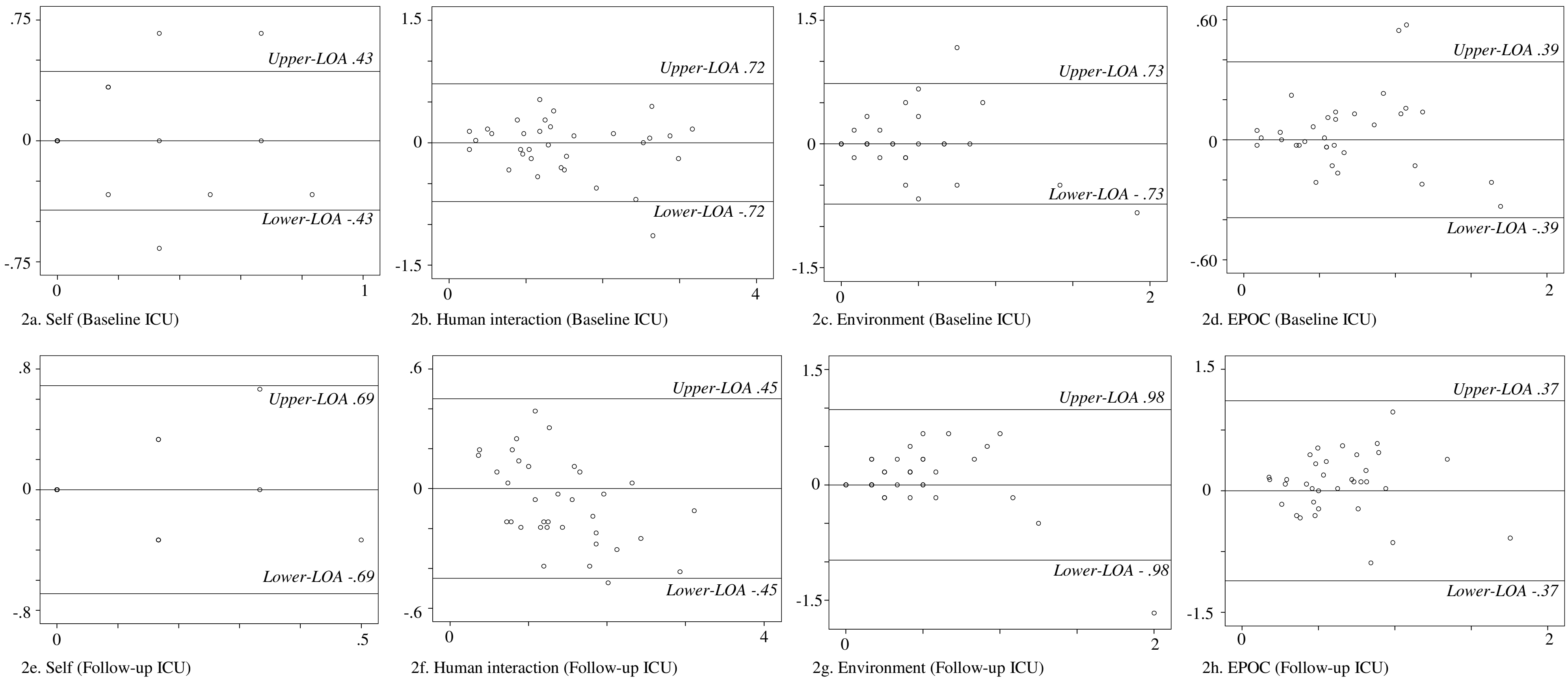
Web only appendix A presents the graphs of the Limits of Agreement. In this plot the average outcome of two contemporaneous observers (x-axis) is compared with the difference between these two observers (y-axis). The Limits of Agreement (LOA) were calculated using the method of Euser et al. [1]. This method was used because the observers are considered random in the present study. The results are graphically displayed by plotting the average outcome of the two observers (x-axis) against the difference between their scores (y-axis) in a 'Bland and Altman plot'. By randomly subtracting the score of observer A from observer B or vice versa, the mean difference was centered at zero.

The graphs show that the results stay well within the upper and lower limits of agreement. The number of contemporaneous observations during the baseline measurement in ED was insufficient to calculate a valid ICC, SEM and LOA. The number of verbal behaviors observed in the category Self was too low in the follow-up measurement of the ED study to make a meaningful plot of the LOA.



1a. Human interaction 1b. Environment 1c. EPOC

Figure 1^{abc}. Limit of Agreement-plots for the overall EPOC score and the category scores in the ED in the follow-up measurement. The average outcome of the two observers (x-axis) is plotted against the difference between their scores centered at zero (y-axis). The Limits of Agreement (0 ± 1.96*√2*SEM) are depicted in every figure. Too few contemporaneous observations were conducted during the baseline measurement to enable the LOA to be calculated. The category Self was under-observed in the follow-up measurement to make a meaningful plot.



2a. Self (Baseline ICU)

2b. Human interaction (Baseline ICU)

2c. Environment (Baseline ICU)

2d. EPOC (Baseline ICU)

2e. Self (Follow-up ICU)

2f. Human interaction (Follow-up ICU)

2g. Environment (Follow-up ICU)

2h. EPOC (Follow-up ICU)

Figure 2[abcdefgh]. Limit of Agreement-plots for the overall EPOC score and the category scores on the ICU in both measurements (baseline and follow-up). The average outcome of the two observers (x-axis) is plotted against the difference between their scores centered at zero (y-axis). The Limits Of Agreement (0 ± 1.96*√2*SEM) are depicted in every figure.

Reference list

1. Euser AM, Dekker FW, Le Cessie S. A practical approach to Bland-Altman plots and variation coefficients for log transformed variables. J Clin Epidemiol. 2008;**61**(**10**):978-982.