DEVELOPMENT OF A JOB TASK ANALYSIS TOOL FOR ASSESSING THE WORK OF PHYSICIANS IN THE INTENSIVE CARE UNIT

Kara Schultz[*~], Jason Slagle[±], Roger Brown[†], Steve Douglas[*], Brian Frederick[*], Manisha Lakhani[*], Jesse Scruggs[†], Bruce Slater[†], Matthew B. Weinger[±], Kenneth E. Wood[†], Pascale Carayon[*~]

[*]Center for Quality and Productivity Improvement
[~]Department of Industrial and Systems Engineering
University of Wisconsin-Madison
Madison, WI

[±]Center for Perioperative Research in Quality
Department of Anesthesiology
Vanderbilt University School of Medicine
Nashville, TN

[†]University of Wisconsin School of Medicine and Public Health
Madison, WI

This paper describes the development of a job task analysis tool for observing and recording physician tasks in the ICU. Real-time direct observations were conducted by outside observers using a computerized data collection tool developed to document the tasks performed by ICU physicians. The aim of the analysis was to quantify the tasks of the physicians, including measures of frequency, duration, and sequence for tasks. In this paper, we report on the development process, as well as the validity and reliability of the job task analysis tool. Initial results from our analyses provide support for the validity and reliability of the taxonomy developed for assessing work of ICU physicians.

INTRODUCTION

This paper describes the development of a job task analysis tool (Kirwan and Ainsworth, 1992) for observing and recording physician tasks in the intensive care unit (ICU) as not much is known about how ICU physicians spend their time. This job task analysis used real-time direct observations by outside observers who used a computerized data collection tool developed by Weinger, Slagle and colleagues (Weinger et al., 1994; Slagle et al., 2002) to document the tasks performed by ICU physicians. The aim of the analysis is to quantify the tasks of the physicians, including measures of frequency, duration, and sequence for both tasks and events. This paper reports data on the development, validity, and reliability of the job task analysis tool. We first begin by discussing the development of the ICU physician task taxonomy. Next, we describe the process of pilot testing the taxonomy and the data collection tool with respect to content validity and reliability. Last, we report the results from our observer training and inter-observer reliability assessment. Initial observers were three human factors engineers who were involved in the development of the training program. The participants in this study were ICU residents, as they are the physicians responsible for writing the vast majority of orders in the ICUs. This effort is part of our larger study to examine the impact of computerized provider order entry (CPOE) on physician tasks in the ICU; we are, therefore, interested in understanding order-related activities performed by ICU physicians. Outcomes to be obtained from the analysis include total time, percent time, and number of occurrences for each of the tasks to determine how ICU physicians spend their time before and after a particular technology implementation and compare with physicians in other care settings (Overhage et al., 2001; Shu et al., 2001).

DEVELOPMENT OF THE JOB TASK ANALYSIS TOOL

Content Validity

Content validity "depends on the extent to which an empirical measurement reflects a specific domain of content" (Carmines and Zeller, 1990). Thus, our first step

in developing the taxonomy was to create a list of tasks that physicians perform in the ICU. The research team began by adapting a list developed and used by Overhage et al. (2001), which was designed to observe physicians in primary care clinics. To attain content validity, the research team, including both engineers and physicians, worked together to iteratively revise the taxonomy.

Initially, one of the human factors engineering observers pilot tested a paper-based version of the task list twice in one hospital ICU and once in a non-ICU that was already using CPOE. These observations ranged from 3-6 hours each. The observer used the coded task list (numbers were given to each task) to record the tasks performed by physicians. Additional input was sought from other physicians and members of the research team. It was critical that the final taxonomy was exhaustive (i.e., that it captured all clinical duties performed) and that the tasks were mutually exclusive (i.e., that there was no overlap in the task definitions). Moreover, the taxonomy had to be easy to train and to use. As the taxonomy reached a more mature state, it was implemented in the data collection software (see below) and installed on two tablet PCs. Observers then piloted the system in several ICUs.

Approximately 16 iterations of the tasks list were completed, ranging from additions/deletions of tasks, reorganization of categories, task name changes, and definition clarifications. The list developed by Overhage and colleagues was manipulated to be relevant to the inpatient rather than the outpatient setting, while covering all tasks performed by ICU physicians.

Data Collection Software

The data collection software allowed trained observers to record tasks in real-time via direct observation using custom software on a tablet PC. Each recorded task was automatically time-stamped and logged into a data file (Weinger et al., 1994; Slagle et al., 2002). Using a stylus on the touch screen of a tablet PC facilitated data collection by permitting observers to enter data while standing or walking. The initial data entry screen prompted the observer to enter participant demographic information (e.g., medical vs. surgical service and adult vs. pediatric ICU). Date, day of the week, and time of day for the start of the observation were automatically recorded by the program. Once data were entered into this screen, the file was saved and the observer proceeded to the task screen (see Figure 1).

There were three main actions the observer could take using the task screen: select a task, select an event marker, or make an annotation. Each time the physician began a new task, the observer selected that task from the task list organized into categories on the screen. Tasks logged with the software could be coded as sequential (i.e., one task follows another) or concurrent (i.e., occurring simultaneously). Simultaneous task occurrences were captured using a toggling function. When observers selected "toggle on", all tasks entered subsequently were considered concurrent until "toggle off" was selected. This function allowed task data to more accurately capture the frequent multifunction activities of ICU physicians. Event markers identified specific events and allowed tasks to be associated with those events. For example, the event 'daily bedside rounds' typically involves numerous tasks such as physician communication with other physicians, nurses, and pharmacists, as well as documentation and order-related tasks. The annotation feature in the software allowed observers to write brief notes about their observations and the data. More detailed observations were made using Microsoft Windows Journal. Annotations allowed observers to correct data collection errors (e.g., replacing 'Nurse conversation' with 'Respiratory therapist conversation') and to capture qualitative data that placed the task and event data in proper clinical context. Narrative information was recorded using handwriting recognition software embedded in the tablet PC operating system (Microsoft Windows XP Tablet PC Edition 2005™).

Task data from each case, automatically saved as a tab-delimited text file, was processed and collated using custom task analysis software (Slagle et al., 2002). The number of individual occurrences of each task, mean duration of each occurrence (i.e., dwell time) of each task, the total time (in minutes) spent on each task category over the entire observation period, the percentage of total time spent on each task category, and aggregated task groups (e.g., all observation tasks) were calculated.

Observer Training and Reliability

Reliability of this behavioral task analysis methodology was dependent on a number of factors. First, each observer had to understand and be familiar with the task list, the software, and the tablet PC. To optimize reliability and minimize the potential for observation error, the research team developed a training manual and procedures for formal observer training. Each observer read literature relevant to CPOE technology and implementation, and about job task

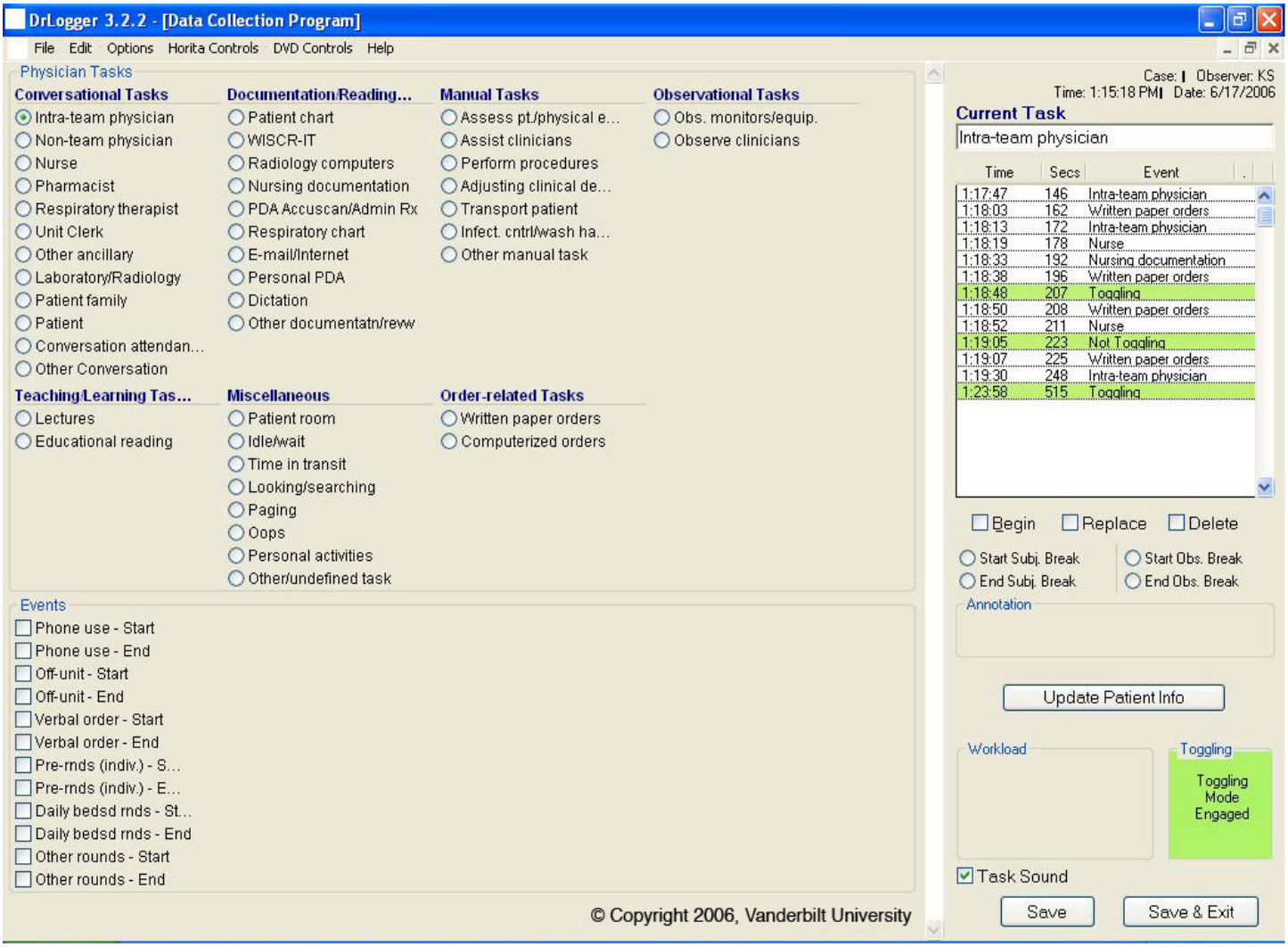

Figure 1. Screenshot of data collection program: task recording screen.

analysis. Observers also familiarized themselves with the ICU environment, physician tasks, and the data collection software by touring the ICUs and conducting practice observations in the units. The early pilot observations aided the research team in making further revisions to the software user interface, taxonomy, observation procedures and observer training. For example, we determined whom to follow (residents vs. fellows) in each unit and in which units we would attempt to observe at night depending on each unit's organizational structure and call schedule. Additionally, we wanted the tasks to be easy for observers to identify, which required us to have physically observable categories. For example, the categories of 'Documentation' and 'Data Review' were combined into one category of 'Documentation/Reading Tasks'. In terms of the software, we were able to make the screen more user friendly by changing the fonts and column widths. The logistics involved with conducting observations and collecting data also generated concern for reliability. The data collection equipment had to be as lightweight as possible and this consideration affected the purchase of a second tablet PC. Additionally, battery life was a constraint on the duration of observations. Issues of where to stand and how to conduct oneself during an observation, as well as tips on collecting data and using the tool were discussed in detail in the training manual.

Assessment of Inter-observer Reliability

Inter-observer agreement refers to the "extent to which two or more observers obtain the same results when measuring the same behavior" (Robson 2002, p.340). According to Robson (2002), studies that involve structured observation, such as the job task analysis should include more than one observer collecting data. Once the taxonomy and data collection instrument were ready, the engineers began conducting observations to assess inter-observer reliability. These observations

Time	Seconds	Task	Duration	Obs	Time	Seconds	Task	Duration	Obs		Lag	Hit
9:47:06	26	Conversation attendance	4	observer1	9:47:08	91	Conversation attendance	13	observer2		2	1
9:47:10	30	Patient chart	18	observer1								0
9:47:28	48	Intra-team physician	107	observer1	9:47:21	104	Intra-team physician	13	observer2		7	1
9:49:15	155	Conversation attendance	14	observer1	9:49:17	220	Conversation attendance	12	observer2		2	1
9:49:29	169	Intra-team physician	82	observer1	9:49:29	232	Intra-team physician	18	observer2		0	1
					9:49:47	250	Patient chart	10	observer2			0

Figure 2. Excerpt of CLIP results

involved two observers simultaneously recording the tasks of the same clinician during patient care. Each observer positioned himself or herself so as to accurately observe the clinician's activities, but not be able to view the screen of the second observer's tablet PC. The observers were instructed not to interact with one another during the observation, so that discussion of data or the data collection process would not bias the reliability measures. Inter-observer reliability was assessed between a human factors engineer and a physician, as well as between the three trained human factors engineers (two at a time) on the observation team. Data collected from inter-observer observations were compared to assess the degree to which the same tasks were recorded during identical clinical observation periods.

In all, twelve inter-observer reliability observations were conducted (10 prior to data collection and 2 about 6 weeks following the start of data collection). Each observation was scheduled for 30 minutes and at various times of the day and the week in order to test reliability of the observers in a variety of situations. One of these pilot observations was conducted by a human factors engineer and a physician; all other inter-observer observations were between the engineers.

In order to evaluate inter-observer reliability, we developed a program called Continuous Logger Improvement Program (CLIP) using Microsoft Visual Basic for Applications. Initially, the data file (in Microsoft Excel format) from each observer's observation is imported into the CLIP file and merged into a temporary spreadsheet. The program is designed to go through each line of the data in the merged sheet, looking for corresponding tasks between observers. For example, if Observer 1 recorded 'Nurse conversation' starting at time X, CLIP will determine if Observer 2 recorded the same task starting within a pre-determined window of time around time X (e.g., within 10 seconds of Observer 1's recording). Each time the program finds a correspondence between Observer 1 and Observer 2's recorded tasks, it marks the line as a 'hit' with a logical value of 1; each non-hit, or 'miss', line is marked with a value of zero (see Figure 2). Thus, a miss occurs when the lag between when one observer records the start of a particular task and when the second observer records the start of same task is greater than the pre-determined time window. Likewise, if both observers record the start of different tasks at the same time (e.g., 'Intra-team physician conversation' and 'Non-team physician conversation'), the line is marked as a miss. The result obtained from CLIP is the percentage of hits out of the total number of hits and misses in the observation (see Figure 2). Table 1 presents the results of our inter-observer reliability observations using CLIP for 7 pre-data collection observations and 2 observations following the start of data collection. CLIP results were not obtained from the other three observations due to poor quality of the data.

	Correspondence Percentage	
Pre-data collection	10-second window	20-second window
1†	21.7	26.1
2	37.7	44.3
3	40.0	51.1
4	54.3	62.9
5	53.6	57.3
6	84.6	84.6
7	42.9	42.9
6 weeks into data collection		
1	57.5	60.9
2	57.1	62.9

†observation between engineer and physician

Table 1. CLIP results for inter-observer reliability.

For all inter-observer reliability observations, the observers examined the data and discussed discrepancies found by CLIP. Table 2 presents results from the last inter-observer reliability observation (#2, 6 weeks into data collection), showing how close the observers were to one another with respect to total time (in seconds) of tasks recorded – an important outcome measure for the job task analysis.

We found CLIP to be useful for observer training and clarifying definitions in the tasks list. The merged spreadsheet pointed out discrepancies between observers that they could discuss and clarify for future

	Observer 1	Observer 2
Conversational Tasks	1530	1532
Intra-team physician	326	248
Patient family	12	13
Conversation attendance	1181	1244
Other conversation	11	27
Documentation/Reading Tasks	236	234
Patient chart	228	221
Other documentation /review	8	13
Order-related Tasks	0	0
Manual Tasks	0	0
Observational Tasks	0	0
Miscellaneous Tasks	41	41
Time in transit/walking	12	18
Paging	29	23
Teaching/Learning Tasks	0	0
Total Time	1807	1807

Table 2. Duration (in seconds) of tasks and categories for two observers conducting an inter-rater reliability observation (see Figure 1 for complete list of tasks and event markers).

observations. For example, initially, there was some confusion regarding how to record certain documents used by physicians. In other cases, we found that observers disagreed in their identification of individuals in conversational tasks. The inter-observer reliability observations provided a means to identify these issues early on and address them prior to beginning data collection.

In addition to the analysis and review of CLIP data, the observers recorded notes from each observation and held weekly meetings to discuss these notes and other issues encountered while conducting observations. During these meetings, the observers would go through each of the past week's observation notes, making decisions regarding necessary task definition revisions, potential task classification errors and corrections to be made to the observation logs that week, or any global issues concerning how the observations were being conducted (i.e., scheduling of observations).

DISCUSSION

This type of behavioral task analysis methodology has been shown to be both valid and reliable in observing physicians in the inpatient setting (Weinger et al., 1994; Slagle et al., 2002). Initial results from our analyses provide similar support for the task taxonomy we developed for assessing work of ICU physicians. Inter-observer reliability needs to be reassessed periodically throughout the study to ensure the continued reliability of observation data. The extensive iterative development of the taxonomy at the start was critical to the project's overall success. This tool provides researchers with an objective description of the task characteristics for physicians in the ICU environment and thus, a better understanding of ICU physician workflow. The process of developing the physician task list and the software benefited greatly from the involvement of an interdisciplinary research team. A similar process would likely be beneficial for other researchers wanting to develop a job task analysis tool in other realms. Likewise, this job task analysis tool could be extended to other professions (e.g., nursing, pharmacy) as well as other contexts (e.g., general hospital wards). We expect the tool will be valuable in our examination of the impact of CPOE technology on the job tasks of physicians in the ICU with data collected both pre- and post-implementation of the technology. Future reliability analyses can be conducted to compare outcomes between observers such as total task time, percent time, number of occurrences and dwell time (task duration).

ACKNOWLEDGMENTS

Funding provided by Agency for Healthcare Research and Quality - AHRQ (1 R01 HS015274-01) and AHRQ Institutional Training Grant (T32 HS00083).

REFERENCES

Carmines EG, Zeller RA (1990). *Reliability and Validity Assessment*. Beverly Hills, California, Sage Publications.

Overhage JM, Perkins S, Tierney WM, McDonald CJ. (2001). Controlled trial of direct physician order entry: effects on physicians' time utilization in ambulatory primary care internal medicine practices. *Journal of the American Medical Informatics Association* 8(4): 361-371.

Kirwan B, Ainsworth LK. (1992). *A Guide to Task Analysis*. Washington, DC,Taylor & Francis, Inc.

Robson C. (2002). *Real World Research – A Resource for Social Scientists and Practitioner-Researchers*. Malden, MA, Blackwell Publishers.

Shu K, Boyle D, Spurr C, Horsky J, Heiman H, O'Connor P, Lepore J, Bates DW. (2001). Comparison of time spent writing orders on paper with computerized physician order entry. *Medinfo* 10(Pt 2): 1207-1211.

Slagle J, Weinger MB, Dinh M-TT, Wertheim VV, Williams K. (2002). Assessment of the intrarater and interrater reliability of an established clinical task analysis methodology. *Anesthesiology* 96(5): 1129-1139.

Weinger MB, Herndon OW, Zornow MH, Paulus MP, Gaba DM, Dallen, LT. (1994). An objective methodology for task analysis and workload assessment in anesthesia providers. *Anesthesiololgy* 80(1): 77-92