Anesthesiology 2002; 96:1129–39

Assessment of the Intrarater and Interrater Reliability of an Established Clinical Task Analysis Methodology

Jason Slagle, M.S., Matthew B. Weinger, M.D.,† My-Than T. Dinh, B.S.,‡ Vanessa V. Brumer, M.B.A.,‡ Kevin Williams, Ph.D.§*

Background: **Task analysis may be useful for assessing how anesthesiologists alter their behavior in response to different clinical situations. In this study, the authors examined the intraobserver and interobserver reliability of an established task analysis methodology.**

Methods: **During 20 routine anesthetic procedures, a trained observer sat in the operating room and categorized in real-time the anesthetist's activities into 38 task categories. Two weeks later, the same observer performed task analysis from videotapes obtained intraoperatively. A different observer performed task analysis from the videotapes on two separate occasions. Data were analyzed for percent of time spent on each task category, average task duration, and number of task occurrences. Rater reliability and agreement were assessed using intraclass correlation coefficients.**

Results: **Intrarater reliability was generally good for categorization of percent time on task and task occurrence (mean intraclass correlation coefficients of 0.84–0.97). There was a comparably high concordance between real-time and video analyses. Interrater reliability was generally good for percent time and task occurrence measurements. However, the interrater reliability of the task duration metric was unsatisfactory, primarily because of the technique used to capture multitasking.**

Conclusions: **A task analysis technique used in anesthesia research for several decades showed good intrarater reliability. Off-line analysis of videotapes is a viable alternative to real-time data collection. Acceptable interrater reliability requires the use of strict task definitions, sophisticated software, and rigorous observer training. New techniques must be developed to more accurately capture multitasking. Substantial effort is required to conduct task analyses that will have sufficient reliability for purposes of research or clinical evaluation.**

TASK analysis is a commonly used human factors technique to identify, quantify, and evaluate job-related tasks.[1–7] Typically, task analysis methods involve the structured decomposition of work activities or decisions and the classification of these activities as a series of tasks, processes, or classes. Task analysis has been used in a wide range of nonmedical domains, for example, to facilitate design (*e.g.*, information systems and human-computer interfaces)[1,3,5,6,8] or to evaluate human, team, or system performance (*e.g.*, nuclear power plant operation, baggage handling, military tactical fire control teams).[4,7,9,10]

A number of task analysis studies have been performed in anesthesiology over the last 25 yr, resulting in the acceptance of an established standardized methodology. Early studies were exploratory in nature and focused on obtaining data to aid in the redesign of anesthesia equipment or processes.[11–16] More recently, researchers have used task analysis to study the impact of new technologies and equipment, such as electronic anesthesia record-keeping systems[17,18] or transesophageal echocardiography,[18] on anesthesia providers' clinical task patterns. Task analysis has also been used to formally study the effects of clinical experience on anesthesia job performance.[19] For example, the task patterns, workload, and vigilance of first-year anesthesia residents have been compared with those of more experienced anesthesia providers.[19] These studies demonstrated that, compared with their more experienced colleagues, novice residents' intraoperative task patterns are significantly different (*e.g.*, novice providers are less efficient, despite manifesting increased workload and decreased vigilance). Task analysis has also been applied in the assessment of the impact of sleep deprivation and fatigue on anesthesia residents during extended call shifts.[20]

Thus, task analysis methodologies can be used to assess the effects of new clinical technologies or processes of care. Task analysis may also be used to measure the outcomes and effectiveness of anesthesia training processes and procedures. Since task analysis requires observers to identify a subject's every action in a highly complex work environment, there is a fundamentally subjective aspect to this methodology. However, the reliability of current anesthesia task analysis methods has not been formally assessed.

This study attempted to address three questions. First, what is the test–retest reliability of the existing task analysis methodology? That is, if the same observer studies the same anesthesiologist doing the same case on two separate occasions, will the same results be obtained (*i.e.*, intrarater reliability)? Second, will two different trained observers viewing the same anesthesiologist doing the same case obtain the same results (*i.e.*, interrater reliability)? Third, are data obtained during viewing of a videotape of a case after-the-fact equivalent to the data

* Staff Research Associate, Department of Anesthesiology, and Doctoral Student, Industrial and Organizational Psychology, California School of Professional Psychology, San Diego, California. † Professor, Department of Anesthesiology, University of California San Diego, Director, Anesthesia Ergonomics Research Laboratory, and Staff Physician, San Diego VA Medical Center, San Diego, California. ‡ Research Assistant, Department of Anesthesiology, University of California–San Diego. § Associate Professor, Department of Psychology, University at Albany, State University of New York, Albany, New York.

Received from the Anesthesia Ergonomics Research Laboratory, Veterans Affairs San Diego Healthcare System, San Diego, California; and the Department of Anesthesiology, University of California–San Diego, La Jolla, California. Submitted for publication September 22, 2000. Accepted for publication January 3, 2002. Supported by a grant from the Anesthesia Patient Safety Foundation, Pittsburgh, Pennsylvania, and a grant from the National Patient Safety Foundation, Chicago, Illinois. Presented in part at the 73rd Clinical and Scientific Congress of the International Anesthesia Research Society, Los Angeles, California, March 14, 1999.

Address correspondence to Dr. Weinger: San Diego VA Healthcare System (125), 3350 La Jolla Village Drive, San Diego, California 92161-5085. Address electronic mail to: mweinger@ucsd.edu. Reprints will not be available from the authors. Individual article reprints may be purchased through the Journal Web site, www.anesthesiology.org.

Table 1. Case, Patient, and Provider (Subject) Demographics

Case No.	Patient Age (yr)	Sex	ASA Physical Status	Surgical Procedure	Duration (min)	Provider*
1	57	F	2	Laparoscopic cholecystectomy	155	CRNA-1
2	42	M	2	Inguinal hernia repair	120	CRNA-1
3	54	M	3	Diagnostic laparoscopy	130	CRNA-2
4	42	M	1	Knee arthroscopy	130	CRNA-3
5	35	M	2	Knee arthroscopy	130	CRNA-3
6	24	F	1	Excision of anal warts	33	CRNA-4
7	54	F	2	Breast biopsy	70	CRNA-4
8	79	M	2	Inguinal herniorrhaphy	144	CA2-1
9	65	M	2	Inguinal herniorrhaphy	130	CRNA-2
10	52	M	2	Sinus cystectomy	115	CA3-1
11	51	M	2	Panendoscopy and biopsy	83	CA3-1
12	69	M	2	Laparoscopic cholecystectomy	166	CA2-1
13	50	M	3	Toe resection	105	CA3-2
14	55	M	2	Hydrocelectomy	95	CRNA-5
15	75	M	2	Neck mass biopsy	90	CA3-3
16	54	M	2	Panendoscopy	75	CRNA-1
17	30	F	2	Laparoscopic cholecystectomy	85	CRNA-6
18†	35	F	2	Incisional herniorrhaphy	85	CRNA-6
19	38	F	2	Laparoscopic cholecystectomy	110	CRNA-7
20	45	F	2	Labial excisional biopsy	95	CRNA-5
Totals‡	50 ± 3 yr	13/7 (M/F)			107 ± 7	n = 12 providers

* CRNA indicates Certified Registered Nurse Anesthetist. CA indicates residents in their second (2) or third (3) year of training. The number after the hyphen indicates a specific anesthesia provider to identify which anesthesia providers performed more than one case (*e.g.*, CA3-1 and CRNAs 1–6 were the anesthesia providers in more than one case). † The videotapes of case 18 inadvertently became unavailable to the second observer. Data for this case was only available for OB1-OR and OB1-VID. ‡ Mean ± standard error of the mean, OB1: n = 20, OB2: n = 19.

ASA = American Society of Anesthesiologists.

obtained from the same case studied in real-time in the operating room? The latter issue is of practical importance because real-time data collection may be logistically difficult because of constraints on observer availability or physical space in the operating room. In addition, controlled studies are enhanced by blinding the observers, and this can be more easily accomplished when data are collected from videotaped cases. For example, it would be difficult to blind the observer collecting data in real time in a study on the effects of extended duty shifts (*i.e.*, nighttime *vs.* daytime cases) on clinical task distribution.

Materials and Methods

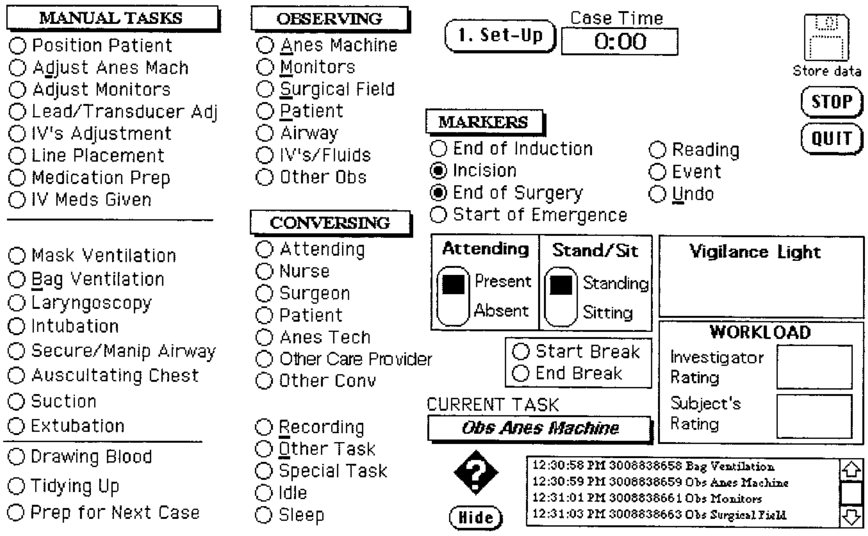
Subjects

After obtaining approval from the institutional review board at the University of California–San Diego, 20 clinical anesthesia cases involving general endotracheal anesthesia were observed and videotaped in the operating rooms of the San Diego VA Medical Center and the University of California–San Diego Medical Center. The anesthesia care providers were seven experienced certified registered nurse anesthesia providers, three CA-2 residents, and two CA-3 residents performing anesthesia for a variety of elective noncomplex surgical cases, primarily on patients who were American Society of Anesthesiologists physical status 2 (table 1). The anesthesia providers gave written informed consent and were instructed that participation in the study would not interfere with their patient care.

Experimental Design

One trained observer (OB1) videotaped and simultaneously performed task analysis in the operating room (OB1-OR) using custom software on a laptop computer. Two to three weeks later, the same observer reviewed each of the 20 videotaped cases and again performed task analysis from the videotapes (OB1-VID). The comparison of this observer's data from the cases in the operating room *versus* the same cases on videotape (OB1-OR *vs.* OB1-VID) provided one assessment of the intrarater reliability of the task analysis methodology and was also used to evaluate the validity of off-line analysis from videotape compared with real-time data collection.

A different trained observer (OB2) subsequently viewed and performed task analysis on the videotapes of the same cases on two occasions separated by at least 2 weeks (OB2-V1 and OB2-V2). It should be noted that the second observer viewed and analyzed only 19 cases because one case's videotapes were misplaced after the first observer had completed her analyses. The comparison of the second observer's two data sets, both from the same videotaped cases (OB2-V1 *vs.* OB2-V2), provided an independent assessment of the intrarater reliability of the methodology (which was not contaminated



Obs Anes Machine

```
12:30:58 PM 3008838658 Bag Ventilation
12:30:59 PM 3008838659 Obs Anes Machine
12:31:01 PM 3008838661 Obs Monitors
12:31:03 PM 3008838663 Obs Surgical Field
```

Fig. 1. Computer screen image of the task analysis data collection software used in this study showing the 38 distinct clinical task categories. Note that in the current study, workload and vigilance data were not collected.

by the issue of the relative fidelity of the videotaped cases *vs.* real-time data). Finally, interrater reliability was assessed by comparing the second observation of each case from videotape by the two observers (n = 19 matched cases; OB1-VID *vs.* OB2-V2). For analytic purposes, the cases observed in this study were treated as a random sample of cases from the population of all possible short elective routine general anesthesia cases of an average complexity (*i.e.*, a random as opposed to a fixed variable model).

Observer Training

Both observers had a college education and had no clinical experience before their comparable training in preparation for this study. The training (10–12 h/week) began with elementary reading about anesthesia care and practice[21] and the observation of clinical cases performed by anesthesia providers in the operating room (*i.e.*, similar to a medical student rotation in anesthesia). The observers then learned a "task dictionary" (see fig. 1 for the list of anesthesia tasks) that defined each task and specified how each clinical activity was to be categorized. Observers then practiced task analysis data collection on both real and videotaped cases under the close supervision of more experienced observers. Data from a minimum of eight complete real-time operating room cases were collected and reviewed before either observer was allowed to participate in the study.

Task Analysis Methodology

In the operating room, the observer was positioned adjacent to the medication cart, across from the anesthesia machine, permitting the observer to view clearly the subject's activities without interfering with patient care. Each case was simultaneously videotaped using a single camera (Sharp VL-E600U; Sharp Electronics Corp., Mahwah, NJ) that had audio-recording capabilities, mounted on a six foot-high tripod to provide a view analogous to the view of the observer collecting data in the operating room.

Data were collected using a custom Hypercard (Claris Corp., Cupertino, CA) task analysis program.[18,19] This software permitted categorization of the anesthesia provider's clinical performance into 38 distinct tasks (fig. 1), separated into broader task groupings of manual, observing, and conversing tasks. The observer used a mouse to click on the button of the task that was currently being performed. If two or more tasks were being performed simultaneously, the observer toggled between the appropriate tasks based on the frequency of each task and the extent that each task consumed the anesthesia provider's time. The software automatically logged the task and the time of its initiation.

The observers logged the beginning and end of induction, the start of surgery, and the initiation of emergence during each case. "End of induction" was defined as the time when the patient had been intubated and the endotracheal tube had been secured, or when the anesthesia provider had told the surgeons that they could begin operating, whichever occurred first. "Beginning of emergence" was defined as occurring when the anesthesia provider shut off all anesthetic agents and began delivering 100% oxygen. "Other conversation" was used to identify any conversation between the subject and the observer. "Other task" was defined as any task that did not fall into any other task category and included telephone conversations, emptying of the urimeter, removal or putting on rubber gloves, taping over the eyes, and setting up or clamping the surgical drape to intravenous poles. The subject was recorded as being "idle" when he–she was performing no other manual, observational, or conversational tasks.

Data Analysis

The analytical approach was structured to test the hypothesis that the task profile of a hypothetical surgical

Table 2. Estimated Duration of Different Phases of the Anesthetic

Phase of Case	Overall Average (min)*	Mean Difference between Observations Intrarater† (Test–Retest)	OR *versus* Video‡	Interrater§
Induction	16.59 ± 7.41	2.13 ± 1.19	1.20 ± 0.28	1.65 ± 0.45
Maintenance	66.54 ± 24.72	3.46 ± 1.71	4.25 ± 1.12	4.80 ± 1.32
Emergence	13.14 ± 6.52	1.59 ± 0.55	4.50 ± 1.29	4.57 ± 1.39‖
Total case	96.26 ± 24.74	1.21 ± 0.65	0.60 ± 0.10	0.60 ± 0.12

* Average (± standard error of the mean) across all four observations (OB1-OR, OB1-V1, OB2-V1, OB2-V2). † Average difference (± standard error of the mean) of one observer's estimates of the duration of each phase of 19 cases on two independent observations from video (OB2-V1 *vs.* OB2-V2). ‡ Average difference (± standard error of the mean) of a different observer's estimates of the duration of the phases of 20 cases comparing one observation in real time (OR) and one observation from video (OB1-OR *vs.* OB1-V1). § Average difference (± standard error of the mean) in the two observers' independent estimates of duration of each phase of 19 cases, each based on one observation from video (OB1-V1 *vs.* OB2-V2). ‖ Significant difference between observers ($P < 0.05$) for estimates of the duration of this phase of the case.

OR = operating room.

case would be evaluated in the same manner by different observers and by the same observer on different occasions. Task data from each case, automatically saved as a text file, was processed and collated using custom software written in Visual Basic (Microsoft Corp., Redmond, WA). The percentage of time spent on each task for each phase of the anesthetic (induction, maintenance, and emergence) and for the total case was calculated. The number of individual occurrences and mean duration of each occurrence (task duration) of each task were also calculated by phase of the anesthetic and over the entire case.

Intraclass correlation coefficients (ICC)[22,23] were used to assess both intraobserver and interobserver reliability in this study. A discussion of the rationale for this approach is provided in the Appendix. An ICC can be interpreted in the same manner as any reliability coefficient, with values of 0.70 or higher generally considered acceptable for basic research purposes[24] and coefficients of at least 0.80 considered necessary for high-stakes decisions (*e.g.*, professional certification exams).[25]

The design of this study allowed for three different ICCs to be computed for each surgical case: (1) a correlation between an observer's behavioral ratings of videotaped cases on two separate occasions (test–retest reliability); (2) a correlation between an observer's real-time ratings and ratings made off-line (from video); and (3) a correlation between the ratings of the same target case behavior made by two different observers (interrater reliability). Because the second analysis (real-time *vs.* off-line) confounds intrarater reliability with mode of observation, comparing that coefficient to the ICCs obtained in the first (OB2-V1 *vs.* OB2-V2) analysis provided an estimate of the unique effects of mode of observation. If minimal differences are found between the two intrarater ICCs, then there would be test–retest data for both raters and it would be possible to structure the data in a 2 (rater) × 2 (time) × 19 (surgical case) factorial design, where each factor represents a facet or condition of measurement. Using analysis of variance procedures, a generalizability coefficient[22] can be computed that estimates the reliability of measurement across all three measurement facets (see Appendix for a more detailed description of generalizability theory).

Results

Twenty cases with an average duration of 107 ± 7 min (mean ± standard error of the mean), performed by 12 different anesthesia providers (seven certified registered nurse anesthetists and five residents), were studied (table 1). There were no significant unexpected or adverse events during any of the cases. Average duration of the different phases of the cases as classified by the observers at each viewing is presented in table 2. Significant differences between the two observers occurred only for emergence, with observer 1 (OB1) reporting a 5-min longer duration for this phase than observer 2 (OB2). Mean values obtained for all four viewing occasions as well as the mean differences for all three comparisons for percent time spent on each of the clinical tasks during the entire case are shown in table 3. There were no significant differences for the percentage of time spent on any of the tasks across any of the three rater comparisons.

Comparison of Second Observer's Two Video Analyses (Intrarater Reliability)

Nineteen videotaped cases were viewed twice by a single observer (OB2-V1 *vs.* OB2-V2). ICCs were calculated for the three criterion variables (percent time, task occurrences, and task duration) for each phase of the anesthetic (induction, maintenance, and emergence), as well as for the total case. Comparison of these observations provided an estimate of within-person (intrarater) or test–retest reliability (OB2-V1 *vs.* OB2-V2). Table 4 presents the test–retest ICC for percent of time spent on tasks for each case by phase. Mean reliability across these surgical cases was high for all phases, ranging from 0.84 to 0.97. Overall reliability for the percentage of time measure was slightly lower and variability in ICC values across cases was greater for induction and emergence

Table 3. Average Percent Time Spent on Each Task Category and Mean Differences between Observation Conditions

		Mean Difference Between Observations*		
Tasks†	Overall (min)‡	Intrarater§ (Test–Retest)	OR *versus* Video‖	Interrater#
Recording	16.83 ± 0.58	1.34 ± 0.30	2.10 ± 0.37	3.90 ± 0.77
Observe monitors	14.38 ± 0.86	2.17 ± 0.50	2.54 ± 0.60	5.19 ± 1.06
Other task	12.54 ± 0.64	2.48 ± 0.51	3.45 ± 0.59	5.07 ± 0.80
Observe surgical field	5.90 ± 0.60	1.72 ± 0.33	2.11 ± 0.47	2.78 ± 0.42
Medication preparation	3.70 ± 0.31	0.89 ± 0.15	1.30 ± 0.39	2.70 ± 0.63
Bag ventilation	3.64 ± 0.55	1.28 ± 0.47	1.29 ± 0.32	3.59 ± 0.78
Position patient	3.27 ± 0.20	0.95 ± 0.17	0.74 ± 0.12	1.45 ± 0.33
Observe patient	3.12 ± 0.23	1.00 ± 0.23	1.48 ± 0.34	2.49 ± 0.43
Lead or transducer adjustment	2.92 ± 0.15	0.47 ± 0.08	0.65 ± 0.07	1.04 ± 0.22
Intravenous line adjustment	2.86 ± 0.18	0.63 ± 0.15	0.89 ± 0.25	1.21 ± 0.23
Adjust anesthesia machine	2.73 ± 0.18	0.78 ± 0.14	0.70 ± 0.14	1.60 ± 0.29
Nurse conversation	2.45 ± 0.21	1.27 ± 0.21	0.72 ± 0.09	1.93 ± 0.46
Intravenous medications given	2.45 ± 0.11	0.45 ± 0.10	0.74 ± 0.14	0.74 ± 0.15
Other conversation	2.25 ± 0.30	0.93 ± 0.26	1.46 ± 0.38	1.25 ± 0.32
Patient conversation	2.20 ± 0.30	0.65 ± 0.18	1.10 ± 0.26	1.30 ± 0.32
Attending conversation	2.03 ± 0.25	0.44 ± 0.14	0.52 ± 0.15	0.61 ± 0.11
Mask ventilation	2.00 ± 0.15	0.63 ± 0.15	0.70 ± 0.19	1.47 ± 0.38
Idle	1.73 ± 0.57	0.01 ± 0.01	1.04 ± 0.38	2.65 ± 1.30
Adjust monitors	1.73 ± 0.15	0.61 ± 0.13	0.67 ± 0.22	0.68 ± 0.11
Prepare for next case	1.64 ± 0.24	0.52 ± 0.15	1.13 ± 0.50	1.87 ± 0.48
Secure or manipulate airway	1.34 ± 0.11	0.42 ± 0.07	0.28 ± 0.07	0.56 ± 0.13
Other care provider conversation	1.04 ± 0.14	0.26 ± 0.06	0.69 ± 0.21	0.80 ± 0.19
Surgeon conversation	1.03 ± 0.17	0.41 ± 0.14	0.60 ± 0.17	0.58 ± 0.13
Other observation	0.63 ± 0.13	0.35 ± 0.08	1.04 ± 0.35	0.82 ± 0.30
Suction	0.62 ± 0.07	0.27 ± 0.12	0.21 ± 0.05	0.16 ± 0.05
Observe airway	0.42 ± 0.09	0.51 ± 0.12	0.01 ± 0.01	0.93 ± 0.20
Laryngoscopy	0.41 ± 0.06	0.17 ± 0.04	0.10 ± 0.02	0.20 ± 0.05
Intubation	0.40 ± 0.04	0.18 ± 0.05	0.16 ± 0.04	0.34 ± 0.07
Observe intravenous line or fluids	0.37 ± 0.06	0.30 ± 0.08	0.18 ± 0.06	0.32 ± 0.07
Tidying up	0.34 ± 0.04	0.28 ± 0.07	0.28 ± 0.06	0.32 ± 0.07
Line placement	0.28 ± 0.16	0.05 ± 0.02	0.14 ± 0.08	0.57 ± 0.41
Extubation	0.20 ± 0.02	0.15 ± 0.04	0.15 ± 0.04	0.17 ± 0.04
Observe anesthesia machine	0.20 ± 0.02	0.19 ± 0.04	0.16 ± 0.04	0.18 ± 0.04
Auscultate chest	0.14 ± 0.03	0.11 ± 0.04	0.08 ± 0.04	0.21 ± 0.07
Anesthesia technician conversation	0.06 ± 0.02	0.01 ± 0.01	0.09 ± 0.03	0.15 ± 0.07
Draw blood	0.03 ± 0.03	0.12 ± 0.12	0.00 ± 0.00	0.00 ± 0.00

* There were no significant differences for any of the comparisons. † Unperformed tasks have been excluded. ‡ Average (± standard error of the mean) across all four observations (OB1-OR, OB1-V1, OB2-V1, OB2-V2). § Average difference (± standard error of the mean) in one observer's estimates of the percent of time spent on each task category on two independent observations from video (OB2-V1 *vs.* OB2-V2). Only 19 cases were available for study. ‖ Average difference (± standard error of the mean) of another observer's estimates of the percent of time spent on each task category on two observations; one done in real time (OR) and one from video (OB1-OR *vs.* OB1-V1). All 20 cases were used in this comparison. # Average difference (± standard error of the mean) in estimates of the percent of time spent on each task category by the two different observers (OB1-V1 *vs.* OB2-V2). Only 19 cases were used in this comparison. OR = operating room.

than maintenance. However, on average, reliability was still acceptable (> 0.80) for induction and emergence across cases.

Table 5 presents the test–retest ICC for the number of task occurrences by case and phase. Reliability for this metric was almost as high as that for the percent time measure, with the mean ICC for the total case being 0.94. Stability of observation was high for the maintenance (mean ICC, 0.94) and emergence (0.87) phases but lower for induction (0.78). Again, there was more case-by-case variability in ICC in the induction and emergence phases than in the maintenance phase.

The test–retest reliability for the task duration measure was appreciably lower than for the other two criterion measures across all phases of the anesthetic (table 6). For example, for the total case, the task duration mean ICC (0.68) was much lower than the mean ICC values for percent time (0.97) or number of occurrences (0.94). The task duration mean ICC values ranged from 0.67 (induction) to 0.73 (maintenance). Variability across cases was also greater. It should be noted that the ICC frequency distributions are skewed in the negative direction for each phase, and thus the median ICC may provide a better estimate of central tendency (tables 4–6).

Real-time versus *Off-line (Video) Intraobserver Analysis*

Data from 20 cases collected by one observer in real time in the operating room were compared with data

Table 4. Intraclass Correlation Coefficients Assessing Intrarater and Interrater Reliability for Percent Total Time by Phases of Surgical Procedure*

	Induction			Maintenance			Emergence		
Case	OR *versus* Video	Test–Retest	Interrater	OR *versus* Video	Test–Retest	Interrater	OR *versus* Video	Test–Retest	Interrater
1. Laparoscopic cholecystectomy	0.54	0.88	0.83	0.96	0.99	0.92	0.70	0.90	0.24
2. Inguinal hernia repair	0.57	0.95	0.80	0.91	0.98	0.89	0.66	0.92	0.44
3. Diagnostic laparoscopy	0.90	0.46	0.85	0.98	0.95	0.95	0.96	0.89	0.80
4. Knee arthroscopy	0.95	0.79	0.85	0.99	0.96	0.95	0.81	0.48	0.45
5. Knee arthroscopy	0.89	0.92	0.87	0.98	0.96	0.91	0.81	0.69	0.84
6. Excision of anal warts	0.95	0.73	0.93	0.96	0.98	0.97	0.77	0.78	0.90
7. Breast biopsy	0.86	0.89	0.85	0.97	0.99	0.94	0.62	0.98	0.76
8. Inguinal herniorrhaphy	0.90	0.94	0.84	0.91	0.98	0.61	0.24	0.97	0.43
9. Inguinal herniorrhaphy	0.93	0.90	0.61	0.98	0.99	0.90	0.69	0.96	0.61
10. Sinus cystectomy	0.74	0.96	0.89	0.98	0.95	0.89	0.79	0.97	0.84
11. Panendoscopy and biopsy	0.61	0.84	0.74	0.84	0.93	0.77	0.86	0.85	0.57
12. Laparoscopic cholecystectomy	0.85	0.88	0.72	0.96	0.94	0.96	0.78	0.48	0.28
13. Toe resection	0.94	0.89	0.93	0.98	0.99	0.99	0.79	0.73	0.72
14. Hydrocelectomy	0.89	0.85	0.89	0.95	0.99	0.67	0.85	0.85	0.54
15. Neck mass biopsy	0.75	0.79	0.56	0.94	0.93	0.91	0.45	0.82	0.82
16. Panendoscopy	0.83	0.82	0.34	0.97	0.99	0.96	0.90	0.83	0.75
17. Laparoscopic cholecystectomy	0.84	0.97	0.92	0.89	0.99	0.68	0.82	0.98	0.69
18. Incisional herniorrhaphy	0.75	—	—	0.88	—	—	0.15	—	—
19. Laparoscopic cholecystectomy	0.90	0.93	0.77	0.96	0.99	0.74	0.95	0.97	0.72
20. Labial excisional biopsy	0.87	0.90	0.91	0.85	0.99	0.89	0.77	0.98	0.23
Mean	0.82	0.86	0.79	0.94	0.97	0.87	0.72	0.84	0.61
95% Confidence interval	0.77–0.88	0.81–0.91	0.73–0.86	0.93–0.96	0.96–0.98	0.82–0.92	0.67–0.82	0.77–0.91	0.52–0.71
Median	0.87	0.89	0.85	0.96	0.98	0.91	0.79	0.89	0.69

* In the OR-1 *versus* VID-2, n = 20; in the test–retest and interrater comparisons, n = 19.

OR = operating room.

from the same cases analyzed by the same observer off-line *via* videotapes (OB1-OR *vs.* OB1-VID). It is important to note that this comparison confounds occasion effects (test–retest stability) with mode of observation (real time *vs.* off-line). The ICCs for the three criterion measures are shown by case and phase in tables 4–6.

For percent of time, ratings were highly reliable across cases and phases (table 4). The mean percent time ICC for the total case was 0.94 and for maintenance was 0.97, with little variability in ICC across cases. The mean ICC was lower for induction (0.82) and emergence (0.72), with greater variability among the cases. Similar results were obtained when number of task occurrences was the criterion variable (table 5). Intraobserver reliability was high (> 0.90) for the total case and the maintenance phase. Intraobserver reliability was lower for induction and emergence (0.79 and 0.77, respectively). As was the case for test–retest reliability, intraobserver reliability for real-time *versus* off-line ratings was weaker when task duration was the criterion (table 6). Overall, the data provided strong evidence that raters are consistent in their measurements over time when the percentage of time spent on each task and the number of individual task occurrences are used as metrics. Furthermore, the use of real-time *versus* off-line ratings had very little effect on consistency of measurement.

Interrater Reliability

The estimates of interrater reliability based on the ICC (tables 4–6) provide evidence of the extent to which the two observers obtained equivalent results when watching the same cases on videotapes. Overall, interrater reliability was consistently lower than intrarater reliability, although still within the acceptable range for two of the three metrics. The mean ICC for the total case was 0.87 and 0.86, respectively, using the percentage of time (table 4) and number of task occurrences (table 5) as criterion variables. Interrater reliability was high (> 0.84) for these criteria during the maintenance phase of cases but lower during induction and emergence. Interrater reliability was notably decreased for percent of total time during emergence (mean ICC, 0.61) and for task occurrences during induction (0.69). Using task duration as the criterion variable resulted in low and unacceptable interrater reliability. Mean and median ICC values for the task duration criterion were less than 0.60 for all phases as well as the total case (table 6).

Generalizability Task Assessments

As a final assessment of the reliability of the task analysis method, we analyzed the data using a generalizability analysis.[22] Generalizability theory can be used to provide an estimate of the extent to which observed measures generalize across facets of measurement to the

Table 5. Intraclass Correlation Coefficients Assessing Intrarater and Interrater Reliability for Number of Occurrences by Phases of Surgical Procedure*

	Induction			Maintenance			Emergence		
Case	OR *versus* Video	Test–Retest	Interrater	OR *versus* Video	Test–Retest	Interrater	OR *versus* Video	Test–Retest	Interrater
1. Laparoscopic cholecystectomy	0.60	0.65	0.46	0.96	0.99	0.67	0.80	0.49	0.28
2. Inguinal hernia repair	0.63	0.87	0.45	0.97	0.94	0.95	0.57	0.93	0.53
3. Diagnostic laparoscopy	0.82	0.62	0.50	0.97	0.64	0.88	0.87	0.86	0.40
4. Knee arthroscopy	0.96	0.77	0.87	0.99	0.93	0.86	0.87	0.88	0.71
5. Knee arthroscopy	0.78	0.71	0.66	0.98	0.99	0.55	0.74	0.68	0.79
6. Excision of anal warts	0.91	0.60	0.67	0.78	0.99	0.93	0.93	0.74	0.81
7. Breast biopsy	0.86	0.91	0.63	0.96	0.97	0.86	0.86	0.95	0.72
8. Inguinal herniorrhaphy	0.90	0.95	0.61	0.89	0.99	0.78	0.15	0.89	0.48
9. Inguinal herniorrhaphy	0.76	0.93	0.64	0.92	0.92	0.68	0.85	0.94	0.75
10. Sinus cystectomy	0.78	0.95	0.70	0.91	0.96	0.86	0.85	0.97	0.93
11. Panendoscopy and biopsy	0.26	0.74	0.66	0.58	0.95	0.80	0.73	0.93	0.71
12. Laparoscopic cholecystectomy	0.84	0.68	0.79	0.97	0.89	0.89	0.72	0.77	0.68
13. Toe resection	0.73	0.81	0.79	0.72	0.93	0.91	0.67	0.91	0.77
14. Hydrocelectomy	0.92	0.84	0.91	0.90	0.98	0.94	0.96	0.94	0.93
15. Neck mass biopsy	0.71	0.48	0.82	0.84	0.92	0.89	0.65	0.91	0.94
16. Panendoscopy	0.89	0.70	0.69	0.98	0.96	0.95	0.93	0.84	0.79
17. Laparoscopic cholecystectomy	0.77	0.80	0.76	0.98	0.97	0.92	0.87	0.99	0.94
18. Incisional herniorrhaphy	0.88	—	—	0.80	—	—	0.59	—	—
19. Laparoscopic cholecystectomy	0.92	0.91	0.78	0.89	0.98	0.86	0.93	0.98	0.93
20. Labial excisional biopsy	0.85	0.93	0.70	0.91	0.98	0.87	0.94	0.94	0.90
Mean	0.79	0.78	0.69	0.90	0.94	0.84	0.77	0.87	0.74
95% Confidence interval	0.72–0.86	0.72–0.84	0.64–0.74	0.86–0.94	0.90–0.98	0.80–0.88	0.69–0.85	0.82–0.92	0.65–0.83
Median	0.83	0.80	0.69	0.92	0.96	0.87	0.85	0.91	0.77

* In the OR-1 *versus* VID-2, n = 20; in the test–retest and interrater comparisons, n = 19.

OR = operating room.

universe of admissible observations (see Appendix for a more detailed description of generalizability theory). In the current case, the generalizability coefficient (ρ^2) provides an estimate of the reliability of the observed measures across all possible (trained) raters, on all possible occasions, and for all possible surgical cases. Using a three-facet design (rater × viewing occasion × surgical case) on data from the total case, separate generalizability coefficients were computed for the percentage of time spent on each task, number of task occurrences, and task duration. The generalizability coefficient for percent time was 0.88, indicating high reliability in measurement across raters, time, and surgical case. Although not as high as for percent time, the generalizability coefficient for number of occurrences (0.78) is indicative of adequate measurement reliability across possible raters, time, and cases. In contrast, the generalizability coefficient for task duration was 0.36, indicating low reliability. Consistent with the ICC analyses reported above, measures of task duration do not generalize in a reliable manner across raters, time, and cases.

Discussion

The current study was undertaken to ascertain the intrarater and interrater reliability of a behavioral task analysis technique that has been used in anesthesiology for almost three decades.[11–13,15–19] In addition, this study sought to evaluate the validity of collecting task data from videotapes of anesthesia cases when compared with data collection in real time in the operating room. To accomplish these goals, two equivalently trained observers collected task data from senior anesthesia residents and experienced nurse anesthetists during elective routine noncomplex general anesthesia cases.

The formal reliability analysis provided evidence of reproducibility but also uncovered shortcomings and limitations of the task analysis methodology. Intrarater reliability was very high. In fact, when percent of total time and number of occurrences were used as criterion measures, 90% of the individual ICC values across all phases were 0.75 or higher. Thus, the task data from the intrarater reliability comparison (*i.e.*, OB2-V1 *vs.* OB2-V2) of the technique demonstrated what is considered acceptable levels of reliability, at least for research purposes.[24,25] Because of the inability of the methodology to accurately record concurrent tasks, as explained below, reliability was highest when percent of total time was the criterion measure. For the most part, data collected from single-view videotapes were sufficiently similar to data collected in real time in the operating room to permit off-line analysis when indicated by a study's experimental design (*e.g.*, need for observer blinding).

Table 6. Intraclass Correlation Coefficients Assessing Intrarater and Interrater Reliability for Task Duration by Phases of Surgical Procedure*

	Induction			Maintenance			Emergence		
Case	OR *versus* Video	Test–Retest	Interrater	OR *versus* Video	Test–Retest	Interrater	OR *versus* Video	Test–Retest	Interrater
1. Laparoscopic cholecystectomy	0.34	0.83	0.52	0.83	0.92	0.77	0.67	0.79	0.39
2. Inguinal hernia repair	0.53	0.78	0.54	0.48	0.79	0.43	0.60	0.63	0.26
3. Diagnostic laparoscopy	0.83	0.16	0.70	0.62	0.68	0.69	0.83	0.40	0.70
4. Knee arthroscopy	0.88	0.77	0.57	0.96	0.57	0.13	0.48	0.30	0.24
5. Knee arthroscopy	0.65	0.76	0.54	0.79	0.44	0.18	0.57	0.53	0.43
6. Excision of anal warts	0.75	0.65	0.69	0.68	0.79	0.70	0.73	0.54	0.48
7. Breast biopsy	0.63	0.82	0.61	0.62	0.88	0.60	0.82	0.84	0.67
8. Inguinal herniorrhaphy	0.87	0.79	0.63	0.65	0.93	0.59	0.44	0.81	0.40
9. Inguinal herniorrhaphy	0.86	0.51	0.50	0.77	0.76	0.45	0.56	0.77	0.23
10. Sinus cystectomy	0.64	0.60	0.40	0.54	0.72	0.51	0.68	0.90	0.76
11. Panendoscopy and biopsy	0.60	0.67	0.81	0.48	0.92	0.85	0.53	0.85	0.67
12. Laparoscopic cholecystectomy	0.67	0.52	0.68	0.63	0.57	0.69	0.31	0.38	0.21
13. Toe resection	0.85	0.78	0.53	0.63	0.56	0.72	0.27	0.56	0.58
14. Hydrocelectomy	0.85	0.66	0.66	0.88	0.53	0.51	0.62	0.72	0.57
15. Neck mass biopsy	0.50	0.37	0.34	0.51	0.75	0.59	0.75	0.73	0.70
16. Panendoscopy	0.88	0.57	0.28	0.73	0.81	0.75	0.83	0.53	0.76
17. Laparoscopic cholecystectomy	0.72	0.90	0.90	0.65	0.56	0.89	0.58	0.89	0.61
18. Incisional herniorrhaphy	0.75	—	—	0.17	—	—	0.27	—	—
19. Laparoscopic cholecystectomy	0.86	0.76	0.38	0.50	0.79	0.41	0.80	0.88	0.70
20. Labial excisional biopsy	0.64	0.75	0.55	0.75	0.95	0.48	0.79	0.89	0.57
Mean	0.72	0.67	0.57	0.64	0.73	0.57	0.61	0.68	0.52
95% Confidence interval	0.65–0.79	0.59–0.75	0.50–0.64	0.57–0.71	0.66–0.80	0.48–0.66	0.53–0.69	0.59–0.77	0.44–0.60
Median	0.74	0.75	0.55	0.64	0.76	0.59	0.61	0.73	0.57

* In the OR-1 *versus* VID-2, n = 20; in the test–retest and interrater comparisons, n = 19.

OR = operating room.

The interrater reliability of this technique was not as high as intrarater reliability; however, it still reached acceptable levels with percent time spent and number of occurrences as criterion measures. Interrater reliability was low when task duration was used as the criterion variable. Interrater reliability was also notably decreased during the emergence phase for all three dependent variables. These suboptimal results suggest the need for improvements in the underlying methodology.

Case characteristics appear to have contributed to the variance in the method's reliability. Some cases had appreciably lower ICC values than others. Because cases were captured on videotape *via* a stationary video camera for off-line analyses, some cases may have had more occasions when the observer's view was at least partially obscured. In addition, some cases may have had higher workload or more heterogeneity of tasks being performed (*i.e.*, were less "routine").

Real-time versus *Off-line (Video) Analysis*

Analysis of videotaped anesthesia cases has developed into an important technique to study decision-making and performance of clinicians, for example, in a trauma setting[26,27] or in on-call providers who are sleep deprived.[20] The use of videotaped cases for task analysis, in lieu of real-time data collection, has both logistical and experimental advantages. The operating room may be too crowded with people and equipment to permit the presence of an observer. Appropriate study cases may occur suddenly or at inopportune times (*e.g.*, in the middle of the night) when observers are less available. The use of videotaped cases facilitates the blinding of observers to study conditions (*e.g.*, time of day of the case or subjects' level of training) and also permits multiple analyses of the same case (*e.g.*, by several observers, as in the current study). On the other hand, off-line video analysis may limit assessment of clinical workload and vigilance.[18] Video-based task analysis is technically more challenging. Obstructed views and poorly localized conversations (*i.e.*, determining who is talking with whom, particularly when one of the conversants is not in view) are more common. During real-time data collection, observers are able to reposition themselves to ameliorate an obstructed view and can query the anesthesiologists as to what they are doing if it is not readily apparent. The limitations of video analysis can only partially be overcome by repeated viewing (which is, nonetheless, a distinct advantage over real-time data collection). Possible technical enhancements include individual microphones for each care provider or the use of multiple camera views.

Interrater Reliability during Video Analysis

Interobserver correlations were consistently not as strong as the intraobserver comparisons, and for some case segments there were appreciable differences between raters. One factor contributing to the comparatively low reliability during the emergence phase was that the two observers in this study disagreed on the time of "beginning of emergence." Thus, for that 5-min period of disagreement, the two observer's data were categorized into different phases of the total case, thereby more greatly affecting the results of the briefer emergence phase. Better definition and identification of event markers appear to be necessary, particularly for off-line video analysis.

Interrater reliability was particularly low for the task duration criterion. This finding may be attributed to a technical aspect of the task analysis method used in this and previous studies,[18,19] whereby observers record concurrent task performance (multitasking) by toggling between the task categories being performed at a rate proportional to the time spent on each task. Although this approach permits the measurement of concurrent task performance, the relative weight (*i.e.*, proportion of time) of two or more simultaneously performed tasks is subject to observer style and interpretation. Thus, different observers toggle between concurrent tasks at different frequencies. Toggling increases the number of task occurrences and shortens the apparent duration of commonly performed concurrent tasks (*e.g.*, observing and conversing tasks). In the current study, one observer (OB2) had a third more task occurrences overall than the other observer (OB1); this was most notable for the most common tasks performed. Therefore, the differences observed in both task duration and task occurrences between the two raters may primarily be a result of differences in individual technique.

These results suggest the need for more rigorous task definitions, more usable data collection software, and improved observer training (including prestudy validation of observer reliability). We subsequently implemented a number of methodologic enhancements and, in a future study, will need to ascertain whether the refinements have improved interrater reliability.

Methodologic and Study Limitations

An important attribute for successful clinical anesthesia care is the ability to time-share attention among several tasks (multitasking),[28–31] yet this critical anesthesia skill has not been formally studied. A technique that effectively captured the complexity and nuances of multitasking could shed appreciable light on the factors that affect clinical task performance. The results of the current study suggest that the traditional anesthesia task analysis methodology cannot address this need. In addition, the data collection software used in this study did not permit the observer to indicate when the anesthesia provider was multitasking so that it was impossible retrospectively to distinguish sequential from concurrent task performance. In an effort to explicitly identify multitasking and to improve the quality of task duration and number of occurrences data, the software has since been modified to allow the observer to explicitly indicate when the clinician is multitasking. The reliability of this new feature must be assessed in future studies. There are alternative methods to examine multitasking. For example, an observer could view a video of a clinical anesthesia case multiple times, each time focusing on the occurrence of individual tasks. Analysis software could then calculate occurrences of multitasking by identifying overlapping tasks.

This study has other limitations. The study design was constrained by the availability of only two equivalently trained observers. However, other investigators have successfully examined interrater reliability with just two raters.[32–34] Different results may have been obtained with different observers or after greater observer training. If these two raters were highly similar to one another but different from the wider pool of potential observers, then our point estimates of reliability may be inaccurate. However, we have no reason to believe that our raters were not representative of the population of all equivalently trained observers. This study focused on routine elective cases, not those of high complexity or long duration. Future studies should assess the reliability of task analysis during more complex cases.

Task analysis studies often use observers who are not domain experts. In previous research, observers have commonly been college-educated research technicians or senior undergraduate preprofessional students. The use of anesthesiologists or nurse anesthetists might increase the quality of the data obtained (although at a higher cost). However, previous results obtained using anesthesiologist observers appeared substantially equivalent to those from trained nonclinician observers. Perhaps the most notable aspect of the current study is that nonmedically trained personnel who have only received modest exposure to the anesthesia domain can obtain valid reproducible quantifiable data on the actions and behaviors of clinicians during actual patient care.

Implications and Conclusions

Task analysis includes a wide variety of ergonomic techniques designed to determine the requirements, goals, and operations of each task and its relation to the overall work process or system. Task analysis may also be used to elucidate the information, skills, knowledge, and abilities required for job-related tasks.[1,4–6] The refined task analysis methodology that has evolved from the current study allows for the quantitative and reproducible assessment of clinical behavior and task patterns in the highly complex anesthesiology work domain. The application of this methodology may provide complementary information when used in combination with other human factors techniques, such as workload assessment[18,19] and measures of situation awareness.[13,18,19,35] For example, in a prospective randomized controlled study of electronic anesthesia record keeping during cardiac surgery,[18] the use of electronic record keeping significantly decreased the actual amount of time spent recording when compared with manual record keeping. However, this modest effect was considered clinically insignificant because the use of electronic record keeping did not enhance vigilance, increase the amount of time spent directly or indirectly monitoring the patient, or decrease clinical workload. Validated measures of what anesthesia providers actually do while caring for real (or realistically simulated) patients may prove to be a valuable complement to written or oral competency assessment.

In summary, this is the first study to formally assess intraobserver and interobserver reliability of a behavioral task analysis methodology applied in the field of anesthesia. The results suggest that the current technique has good intrarater reliability and that off-line task analysis of videotaped cases is a viable approach. The results also suggest that the technique yields measures of percent time spent on tasks and number of occurrences that are reproducible across raters, viewing occasions, and surgical cases. Still, enhanced interrater reliability may be achieved with the use of explicit task definitions, effective data collection software, and a rigorous program of observer training and validation. Alternative techniques for capturing multitasking should be examined.

The authors thank Victor Tubbesing, B.S. (Research Assistant, Department of Anesthesiology, University of California, San Diego, California), and Christopher Herndon, B.S. (Medical Student, Yale University, New Haven, Connecticut), for significant contributions; Gery Schulteis, Ph.D. (Associate Professor, University of California, San Diego, California), for advice and assistance; and Frank Dexter, M.D., Ph.D. (Associate Professor, Department of Anesthesia, University of Iowa, Iowa City, Iowa), for statistical advice.

Appendix

Intraclass Correlation Coefficients

There has been considerable debate over the appropriate statistical methods and indices for assessing the reliability of direct observations, with the general conclusion being that there is no perfect or even universally preferred index.[36] We used ICCs[22,23] to assess both intraobserver and interobserver reliability in this study. In a review and conceptual critique of various reliability indices and their applications, Suen[37] argued that the intraclass correlation approach is superior to traditional indices such as percent agreement indices or Pearson r, because it identifies sources of observational error before estimating reliability and is flexible enough to accommodate different observation paradigms and multidimensional measurement conditions. Our experimental design manipulated dimensions of observation and hence is an ideal application of ICC. In addition, ICC identifies the different sources of observational error and thus can guide further improvement in a method's reliability.

An ICC provides information about the relative contributions of different sources of error comprising the observed score. Sources of error are estimated through an analysis of variance procedure. Thus, the ICC is a ratio of true variance in observations because of "targets" ($\sigma_T{}^2$) divided by the sum of true variance plus random error variance ($\sigma_e{}^2$). Symbolically, the intraclass coefficient can be expressed as:

$$\rho^2 = \sigma_T^2/(\sigma_T^2 + \sigma_e^2)$$

$\sigma_e{}^2$ in this case incorporates variance caused by "facets" of measurement, or aspects of the observation situation over which the researcher wishes to generalize. Examples of facets are judges, observers, items, and viewing occasion. Thus, when the proportion of variance caused by facets is low and the proportion of variance caused by targets is high, then the ICC approaches 1.0 and reliability is high. In our analyses, observers, time, viewing mode, and surgical case constituted "facets," and the "targets" were the estimates for the 38 individual tasks performed in each surgical case. For example, the "targets" in the ICC analysis for assessing the test–retest reliability of time spent on tasks would be the estimates for percent of time spent on the 38 clinical tasks in each case. Our ICC is essentially a ratio of the proportion of variance in observations caused by tasks divided by the sum of the proportion of variance caused by tasks plus the variance caused by time, observer, or viewing mode, plus residual error. Specifically, the ICC used to access interrater reliability can be expressed as:

$$\rho^2 = \sigma_T^2/[\sigma_T^2 + \sigma_R^2 + (\sigma_{TR}^2 + \sigma_e^2)],$$

where T is the task variance (*i.e.*, systematic differences in the extent to which tasks are performed), R is the variance caused by raters, TR is the variance caused by the interaction of raters and tasks (*i.e.*, inconsistencies caused by raters' evaluation of particular tasks), and e is the residual error. For intrarater reliability, the ICC can be expressed as:

$$\rho^2 = \sigma_T^2/[\sigma_T^2 + \sigma_O^2 + (\sigma_{TO}^2 + \sigma_e^2)],$$

where O is the variance caused by viewing occasion, and TO is the variance caused by the interaction of occasions and tasks (*i.e.*, inconsistencies caused by task evaluations from one occasion to another).

Different types of ICCs can be computed depending on the unit of reliability (single *vs.* average measures) and type of judgment (consistency *vs.* agreement).[23,38–40] The unit of reliability may be single measurements (*e.g.*, the rating of a judge) or average measurements (*e.g.*, the average of k raters), with estimates for average measures typically being higher than those for single measurements. The theoretical question of interest here relates to the reliability of a randomly selected observer. Thus, we computed ICC based on single measurements. An advantage of the ICC approach is that, if reliability for single measurement is inadequate, the variance component information can be used to determine how many raters would be needed to provide adequate reliability.

Different ICC values are also computed depending on the type of judgment to which one wishes to generalize. Shrout and Fleiss[23] distinguished between two types of judgments: one that reflects relative consistency in observations across facets (sometimes referred to as norm-referenced reliability) and one that reflects absolute agreement in observations across facets (sometimes referred to as criterion-referenced reliability). The relative consistency ICC, which does include rater variance in the error term (and hence results in a lower coefficient), applies when comparative judgments are made among targets of measurement. The difference between the two can be illustrated using the paired scores (2,4), (4,6), and (6,8). The consistency definition, which ignores the elevation (or scaling) differences between raters, yields a coefficient of 1.00. The agreement coefficient, which treats the scaling differences as error, yields a coefficient of 0.67. The absolute agreement ICC applies when decisions are to be made about the absolute level of a target's standing. In this study, the distinction is between, for example, how much time is spent on task A relative to others *versus* how much time is spent on task A in an absolute sense. We were interested in the stricter standard of how much time is actually spent on the tasks and therefore used the formula for absolute agreement. Thus, the ICCs reported here provide estimates of the criterion-reference reliability of a single observation made by a single randomly selected trained observer and can be interpreted as omnibus indicators of intraobserver stability and interobserver agreement.[37] An ICC can be interpreted in the same manner as any reliability coefficient, with values of 0.70 or higher generally considered acceptable for basic research purposes,[24] while coefficients of at least 0.80 are considered optimal for high-stakes decisions.[25]

Generalizability Coefficients

A form of intraclass coefficient that can be used when multiple sources of error are present is the generalizability coefficient. Cronbach *et al.*[22] introduced generalizability theory as a statistical approach to the dependability or reliability of measurements. A basic assumption in generalizability theory is that researchers and practitioners need to be able to generalize from their measures or observations over several facets of measurement. Users of the task analysis methodology, for example, want to be able to generalize from a particular rater's evaluations on a particular surgical case to all possible raters at all possible times for all possible surgical cases. Generalizability theory provides a way of identifying multiple sources of error in behavioral measurement. It also provides a summary coefficient—the generalizability coefficient—that reflects the level of dependability or reliability of measurement across the facets being studied. The error term for the typical ICC is expanded to include separate components for the facets of interest. A detailed description of generalizability theory is beyond the scope of the current investigation; interested readers are referred to Cronbach *et al.*[22] or Shavelson and Webb.[40] In the current study, generalizability coefficients provided an overall assessment of the extent to which task analysis ratings generalize across the three facets of time, rater, and surgical case.

References

1. Card SK, Moran TP, Newell A: The Psychology of Human-Computer Interaction. Hillsdale, New Jersey, Lawrence Erlbaum Associates, 1983, p 469

2. Gilbreth FB: Motion study in surgery. Can J Med Surg 1916; 40:22–31

3. Kearsley G, Halley R: Designing Interactive Software. La Jolla, Park Row Press, 1985, p 101

4. Kirwan B, Ainsworth LK: A Guide to Task Analysis. London, Taylor & Francis, 1992, p 417

5. McGraw K, Harbison K: User-Centered Requirements: The Scenario-Based Engineering Process. Mahwah, New Jersey, Lawrence Erlbaum Associates, 1997, p 380

6. Norman KL: The Psychology of Menu Selection: Designing Cognitive Control at the Human/Computer Interface. Norwood, New Jersey, Ablex Publishing, 1991, p 350

7. Salvendy G: Handbook of Human Factors. New York, Wiley, 1987, p 1874

8. Parasuraman R, Mouloua M: Automation and Human Performance: Theory and Applications. Mahwah, New Jersey, Lawrence Erlbaum Associates, 1996, p 514

9. Woodson WE, Tillman B, Tillman P: Human Factors Design Handbook, 2nd edition. New York, McGraw-Hill, 1992, p 846

10. Sells SB, Berry CA: Human Factors in Jet and Space Travel: A Medical-Psychological Analysis. New York, The Ronald Press, 1961, p 386

11. Boquet G, Bushman J, Davenport H: The anesthesia machine: A study of function and design. Br J Anaesth 1980; 52:61–7

12. Drui AB, Behm RJ, Martin WE: Predesign investigation of the anesthesia operational environment. Anesth Analg 1973; 52:584–91

13. Gaba D, Lee T: Measuring the workload of the anesthesiologist. Anesth Analg 1990; 71:354–61

14. Kennedy PJ, Feingold A, Wiener EL: Analysis of tasks and human factors in anesthesia for coronary artery bypass. Anesth Analg 1976; 55:374–7

15. McDonald J, Dzwonczyk R: A time and motion study of the anaesthetist's intraoperative time. Br J Anaesth 1988; 61:738–42

16. McDonald J, Dzwonczyk R, Gupta B, Dahl M: A second time-motion study of the anaesthetist's intraoperative period. Br J Anaesth 1990; 64:582–5

17. Allard J, Dzwonczyk R, Yablok D: Effect of automatic record keeping on vigilance and record keeping time. Br J Anaesth 1995; 74:619–26

18. Weinger MB, Herndon OW, Gaba DM: The effect of electronic record keeping and transesophogeal echocardiography on task distribution, workload, and vigilance during cardiac anesthesia. Anesthesiology 1997; 87:144–55

19. Weinger MB, Herndon OW, Paulus MP, Gaba D, Zornow MH, Dallen LD: Objective task analysis and workload assessment of anesthesia providers. Anesthesiology 1994; 80:77–92

20. Weinger MB, Vora S, Herndon CN, Howard SK, Smith BE, Mazzei WJ, Rosekind MR, Gaba DM: Evaluation of the effects of fatigue and sleepiness on clinical performance in on-call anesthesia residents during actual night time cases and in simulated cases, Proceedings of Enhancing Patient Safety and Reducing Errors in Health Care. Chicago, National Patient Safety Foundation, 1999, pp 306–10

21. Longnecker DE, Murphy FL: Dripps/Eckenhoff/Vandam: Introduction to Anesthesia, 9th edition. Philadelphia, Saunders, 1997, p 518

22. Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N: The Dependability of Behavioral Measurements: Theory of gerneralizability for scores and profiles. New York, Wiley, 1992, p 410

23. Shrout PE, Fleiss JL: Intraclass correlations: Uses in assessing rater reliability. Psychol Bull 1979; 86:420–8

24. Nunnally JC: Psychometric Theory, 2nd edition. New York, McGraw Hill, 1978

25. Walsh WB, Betz NE: Tests and Assessment, 2nd edition. Englewood Cliffs, New Jersey, Prentice Hall, 1990, p 469

26. Mackenzie CF, Craig GR, Parr MJ, Horst R: Video analysis of two emergency tracheal intubations identifies flawed decision-making. Anesthesiology 1994; 81:763–71

27. Mackenzie CF, Martin P, Xiao Y: Video analysis of prolonged uncorrected esophageal intubation. Anesthesiology 1996; 87:1494–503

28. Gaba DM, Howard SK, Small SD: Situation awareness in anesthesiology. Hum Factors 1995; 37:20–31

29. Jennings AE, Chiles WD: An investigation of time-sharing ability as a factor in complex performance. Hum Factors 1977; 19:535–47

30. Krasner H, Connelly N, Flack J, Weintraub A: A multidisciplinary process to improve the efficiency of cardiac operating rooms. J Cardiothor Vasc Anesth 1999; 13:661–5

31. Schneider W, Detweiler M: The role of practice in dual-task performance: Toward workload modeling in a connectionist/control architecture. Hum Factors 1988; 30:539–66

32. Rosenbaum B, Slezer MA, Valbak K, Hougaard E, Sommerlund B: The Dynamic Assessment Interview: Testing the psychodynamic assessment variables. Acta Psychiat Scand 1997; 95:531–8

33. Siegel AI, Bergman BB: A job learning approach to performance prediction. Personnel Psychol 1975; 28:325–39

34. Stoeffelmayr BE, Mavis BE, Kasim RM: The longitudinal stability of the Addiction Severity Index. J Substance Abuse Treat 1994; 11:373–8

35. Loeb RG: A measure of intraoperative attention to monitor displays. Anesth Analg 1993; 76:337–41

36. Mitchell SR: Interobserver agreement, reliability, and generalizability of data collected in observational studies. Psychol Bull 1979; 86:376–90

37. Suen HK: Agreement, reliability, accuracy, and validity: Toward a clarification. Behavioral Assess 1988; 10:343–66

38. James LR: Aggregation bias in estimates of perceptual agreement. J Appl Psychol 1982; 67:219–29

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

40. Shavelson RJ, Webb NM: Generalizability Theory: A Primer. Newbury Park, California, Sage, 1991