2007

The Effects of Multimodal Collaboration Technology on Subjective Workload Profiles of Tactical Air Battle Management Teams

Victor S. Finomore

Benjamin A. Knott

W. Todd Nelson

Scott M. Galster

Robert S. Bolia

Repository Citation

Finomore, V. S., Knott, B. A., Nelson, W. T., Galster, S. M., & Bolia, R. S. (2007). The Effects of Multimodal Collaboration Technology on Subjective Workload Profiles of Tactical Air Battle Management Teams. *2007 International Symposium on Aviation Psychology*, 190-196.
https://corescholar.libraries.wright.edu/isap_2007/104

THE EFFECTS OF MULTIMODAL COLLABORATION TECHNOLOGY ON SUBJECTIVE WORKLOAD PROFILES OF TACTICAL AIR BATTLE MANAGEMENT TEAMS

Victor S. Finomore, Benjamin A. Knott
General Dynamics, Advanced Information Systems, Dayton, Ohio

W. Todd Nelson, Scott M. Galster, Robert S. Bolia
Air Force Research Laboratory, Dayton, Ohio

A tactical air battle management task required a team of two weapons directors, two strike operators, and a tanker operator to communicate with each other in order to coordinate offensive and defensive air attacks, and aerial refuelling. This study compared the impact of two types of communication modalities (Voice or Picture Chat) and the number of enemy targets (4 or 6) on team performance and perceived team workload. Three subjective workload scales were evaluated in their ability to characterize task difficulty, communication demands, and demands of the different team roles. The results are discussed with respects to the descriptive and discriminating abilities of the three team workload scales.

Introduction

Team Workload Assessment

The assessment and measurement of various aspects of team performance is increasingly important to many military and civil aviation domains such as tactical air battle management (ABM), air traffic control, and emergency medical response. Within these domains, modern communication and information networks support collaboration within and between widely distributed teams. As a result, the design and evaluation of collaborative technology (CT) interfaces with respect to their impact on team communication, coordination, and information sharing is receiving a great deal of recent attention among human factors researchers (Cummings, 2004; Bolstad & Endsley, 2005; Knott, Bolia, Nelson, and Galster, 2006a).

When exploring the effectiveness of CTs, the evaluation of performance measures is important. However measurement of mental workload imposed by the task is also critical. Mental workload refers to the demands on cognitive resources experienced during the performance of a task. This concept has been used to identify bottlenecks in systems where task demands exceed operators' supply of cognitive resources (O'Donnell & Eggemeier, 1986). Mental workload has also been used to evaluate the different resource demands placed upon the operator from alternative system designs in order to determine the most optimal interface (Wickens & Hollands, 2000).

Extending this concept from individual operators to teams, it is crucial for identifying bottlenecks related to team structure, processes, or system interfaces where demands of the environment exceed team resources. A better understanding of team workload would allow researchers to identify problems with perceptual, physical, and cognitive demands and aid in the development of more efficient CTs and team structures. One goal of the present study was to assess and characterize the workload demands placed upon teams engaged in an ABM task when using different CT interfaces. The experiment evaluated and compared the relative contribution of three subjective workload scales to the understanding of demands placed upon ABM teams. The three workload scales that were compared are:

(1) The NASA-Task Load Index (NASA-TLX; Hart & Staveland, 1988), which is one of the most effective measures of perceived mental workload currently available (Wickens & Hollands, 2000). It provides a global workload index on a scale of 0 to 100 and identifies the relative contributions of six sources of workload: Mental Demand, Temporal Demand, Physical Demand, Performance Effort, and Frustration. The NASA-TLX has been used in many laboratory and real-world tasks to identify situations that are cognitively demanding (Warm, Dember, & Hancock, 1996).

(2) A modified Multiple Resources Questionnaire (MRQ, Boles & Adair, 2001), in which observers are presented with a set of mental processes based upon a combination of dimensions drawn from Wickens' Multiple Resource Theory (Wickens & Hollands, 2000). The MRQ consists of the 17 resource dimensions listed in Table 1.

Using a modified scale (Finomore et al, 2006) from 0 (no usage) to 100 (extreme usage), observers are asked to rate the extent to which a task they just performed utilized each dimension. Research with the modified MRQ has indicated greater sensitivity without modifying its diagnostic profile (Finomore et al, 2006). This result supports the conclusion that a modified MRQ may be useful in identifying sources of mental workload in tasks that are not present in the NASA-TLX. In addition, the MRQ has been successful in predicting the interference between tasks based upon shard resource dimensions (Boles, Bursk, Phillips, & Perdelwitz, 2007).

Table 1

The 17 MRQ resource dimensions

Subscales	Abbreviations
Vocal	V
Tactile Figural	TF
Facial Figural	FF
Auditory Linguistic	AL
Auditory Emotional	AE
Visual Phonetic	VP
Visual Lexical	VL
Facial Motive	FM
Spatial Quantitative	SQ
STM	STM
Spatial Concentrative	S
Spatial Positional	SP
Spatial Emergent	SE
Visual Temporal	VT
Spatial Categorical	SC
Manual Process	MP
Spatial Attentive	SA

(3) The Team Workload Scale (TWS, Hildebrand, Pharmer, & Weaver, 2003) is exploratory and to date has not been psychometrically validated. However, the TWS is of interest to the current investigation because it was designed specifically to measure the demands of team processes. The TWS provides a measure on a scale of 0 to 20 of five sources of workload: (1) Communication Demand, (2) Monitoring Demand, (3) Control Demand, (4) Coordination Demand, and (5) Leadership Demand. The TWS does not provide an overall team workload score but rather provides ratings for each of the subscales.

Collaborative Technologies for Tactical Air Battle Management

Tactical ABM refers to the command and control of assets engaged in air combat operations such as strike, and defensive counter air missions. At the tactical level, this task is handled by weapons directors (WDs) who are typically located on an airborne platform such as the E-3 Airborne Warning and Control System (AWACS). The primary task of a WD is the control of air assets, which includes communication-intensive tasks such as vectoring of aircraft to intercept hostile targets, or sharing the tactical picture with other platforms. In current ABM, collaboration is accomplished primarily by means of voice communication over multiple radio channels. This may not be optimal however, as voice intelligibility is affected by platform noise, is subject to interference from multiple sources, and imposes working memory demands on the operators (Bolia, Nelson, Vidulich, Simpson, & Brungart, 2005).

The proliferation of collaboration technologies such as instant messaging (chat), data visualization, and digital whiteboards, promise alternatives or enhancements to the traditional voice communication. However, commercial-off-the-self CTs are often employed without proper evaluation or consideration of the work domain and the cognitive requirements of its operators (Scott, Cummings, Graeber, Nelson, & Bolia, 2006).

In the present study, teams conducted a simulated ABM scenario in which they communicated in the traditional manner using radio headsets, or by using a custom graphical whiteboard tool, called "Picture Chat" (PC), to augment voice communication. Picture Chat is a collaboration tool that was developed with consideration to the communication-intensive demands of ABM. This tool allows WDs to send images of a tactical display to teammates, annotated with task-relevant symbols indicating desired actions or directives. This type of visual communication is germane to our evaluation of subjective workload instruments for team research, because it presumably imposes significantly different cognitive demands on the team compared to the baseline voice communication.

Methods

The Distributed Dynamic Decision-Making (DDD) Simulator was employed to create a set of ABM scenarios conveyed to participants through a tactical display that exhibited the movement of entities within a battle space.

The ABM scenario was constructed around a five-member team. Two WDs coordinate operations by communicating with each other, two Strike Operators (SO) and a Tanker Operator (TO) to intercept threats and resupply assets as needed. To do this effectively, WDs must understand the capabilities, limitations, and resources of their operational environment. Within the simulation, three classes of friendly fighter assets (F-15, F-16, & F-18) and two classes of hostile targets (MiGs & Su-27s) were employed. The F-15s and F-16s were equipped with two missiles and could only attack the MiGs; the F18s were outfitted with four missiles to attack two MiGs and two Su-27s. Moreover, each fighter began the mission with different fuel capacities, the F15s and F16 could refuel at the Air Force tanker and the F18s at the Navel Tanker.

The WDs' role was to match friendly fighters with the appropriate enemy targets, schedule fighters for refueling and resupply, and communicate their plan of action to the SOs and TO. As such, WDs had primary decision making and leadership responsibility. The role of the SOs and TO was to maneuver assets as instructed and to provide pertinent information to WDs concerning the resources of team assets. The number of targets present throughout scenario was deliberately controlled as a manipulation of task difficulty. A more in-depth explanation of the scenario can be found in Knott, Bolia, Nelson, and Galster (2006b).

Participants

Five men and five women between the ages of 18 and 25 yrs. (Median = 22) were paid to $15 for each hour of participation in the experiment. Participants were combined to create 10 unique teams. Seven of the participants were undergraduate students and three were graduate students.

Procedure

Prior to the experiment, all participants completed two 4-hour training sessions in which they practiced on the DDD platform for all team roles and with the CT software. Team members communicated either through simulated radio headsets or by using Picture Chat (PC). An example of a PC message is shown in Figure 1, in which a WD was requesting two target intercepts and refueling of two fighters.

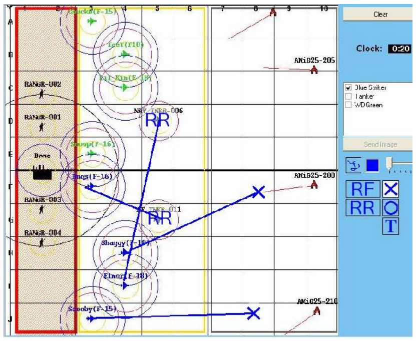


Figure 1. An example of a Picture Chat message. Directives are communicated to teammates in a visual manner by connecting assets and targets with lines and task-relevant symbols.

The trainer informed participants that the purpose of the study was to evaluate how teams used communication technology to work together to meet the ABM scenario's objectives. In addition, participants were trained on and practiced communication brevity, such that voice communication mimicked the highly structured communication used in ABM. Participants were also trained on the specific objectives and rules of the mission, and were instructed that the performance of the team would be measured for each trial based on how well they met their objectives.

The experimental session consisted of four 10-minute experimental trials per team. After each trial, participants completed the three subjective mental workload scales (NASA-TLX, MRQ, and TWS). All major simulation events (e.g., the occurrence and outcome of attacks, refueling events, etc.) were recorded in data logs for later analysis. In addition, all Voice and PC communications were recorded.

Experimental Design

There were two levels of collaboration technology (CT) (voice-only, and voice & PC), and two levels of task difficulty (4 or 6 targets). These independent variables were combined factorially, yielding a 2 × 2 within-subjects design. In the Voice & PC condition, WDs were instructed to use the PC tool for all movement related communication and could use voice to check asset status. The order of conditions was counterbalanced across trials.

Results

Team Performance

A single measure of team performance was calculated for each trial by averaging the percentage of (1) targets that penetrated friendly airspace, (2) high value assets destroyed (the air base, infantry units, and tanker aircraft), and (3) fighter assets lost. The average was then subtracted from 100, resulting in a team performance score in which 100 indicated optimal team performance in accordance with the mission objectives.

A target × collaboration technology repeated measures analysis of variance (ANOVA) was performed on the team score, which revealed a main effect for number of targets, F (1, 9) = 73.78, $p < .001$. In this and all subsequent ANOVA's, Box's epsilon was employed to correct for violations of the sphericity assumption. Performance scores were significantly higher for the 4 target condition (M = 82.7, SE = 2.6) compared to the 6 target condition (M = 69.1, SE = 2.9). All other sources of variance in this analysis lacked significance.

Subjective Workload Scales

Subscale ratings for each of the three subjective workload instruments were submitted to a Subscale × Team Role (WD/SO) × Targets (4/6) × CT (voice/voice & PC) mixed-design ANOVA with team role as a between-subjects factor. A series of analyses were conducted to determine how each of the instrument's subscales varied as a function of 1) changes in task difficulty, 2) the demands of different team roles, and 3) the demands of different collaborative interfaces. When appropriate, a Bonferoni correction was used to test post- hoc comparisons. Correlations coefficients between subscales and the team performance measure were also computed to determine how each of the workload subscales could serve as predictors of team performance. Results are reported for coefficients that are significant at the $p < .05$ level.

NASA-TLX

Weighted NASA-TLX ratings were submitted to a 5 (Subscale) × 2 (Team Role) × 2 (Targets) × 2 (CT) mixed-design ANOVA. The main effects of targets and CT indicated that the NASA-TLX instrument was indeed sensitive to a) the primary task difficulty manipulation, and b) the differential demands of the two CT interfaces. The 6-target condition ($M = 138.8$, $SE = 6.8$) resulted in higher average scores than the 4-target condition ($M = 130.6$, $SE = 7.1$). Additionally, use of PC resulted in higher ratings ($M = 138.9$, $SE = 7.3$) than voice communication alone ($M = 130.5$, $SE = 7.0$).

Post hoc comparisons of the significant targets × subscale interaction ($F(5, 90) = 2.8$, $p < .05$) indicated that the sensitivity of the NASA-TLX to targets was due to the reliable difference on the temporal demands and performance subscales for 6 and 4 targets (Figure 2). The significant CT × subscale interaction ($F(5, 90) = 2.8$, $p < .05$) indicated that the sensitivity of the NASA-TLX to CT conditions was due to the reliable difference between the voice only and PC conditions for frustration and performance subscales (Figure 3).

Temporal and performance demand subscales varied with task difficulty, such that higher task difficulty resulted in higher ratings on these scales. Frustration and performance varied with CT in that the PC condition resulted in higher ratings on these subscales. Notably the NASA-TLX was not sensitive to differences in team role.

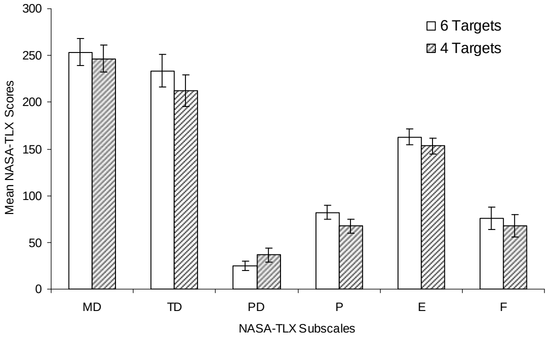


Figure 2. Mean NASA-TLX subscale scores for number of enemy targets.

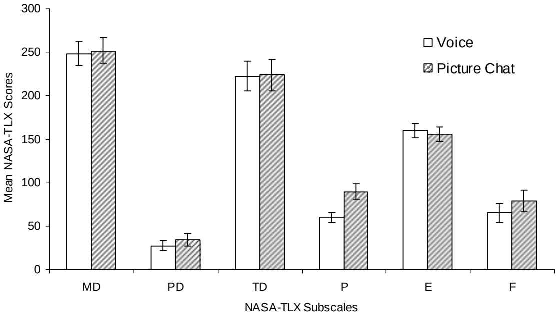
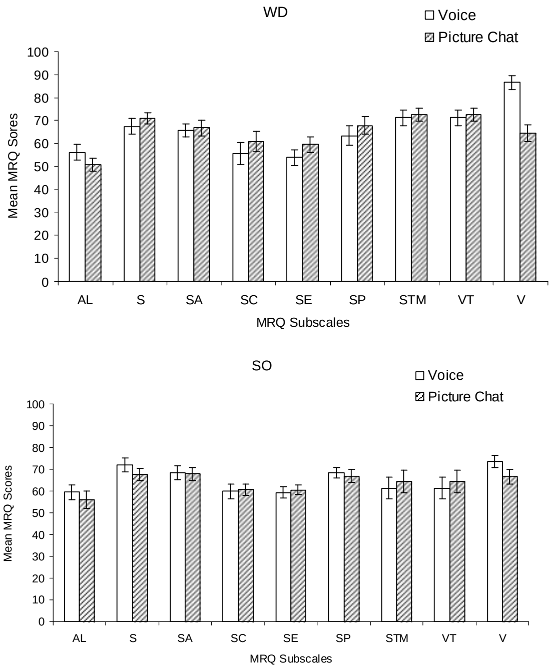


Figure 3. Mean NASA-TLX subscale scores for communication modality

MRQ

Ratings were submitted to a 17 (Subscale) × 2 (Team Role) × 2 (Targets) × 2 (CT) mixed-ANOVA, yielding main effects for subscale, $F(16, 288) = 58.7$, $p < .05$, and a two way interaction of CT × Subscale, $F(16, 288) = 4.8$, $p < .05$. The significant three-way interaction, $F(16, 288) = 1.74$, $p < .05$, indicated a different CT × subscale profile for the two team roles (Figure 4). For WDs, the voice only condition lead to greater demands on vocal processes, but the PC condition led to greater spatial emergent processes demands. In addition, there were marginal differences between CT conditions on SC, SE, and SP subscales, though these differences did not reach significance in post hoc comparisons. For SOs, only vocal process differed across CT conditions, and the trends for the spatial processing dimensions were not present. MRQ ratings were sensitive to CT and team role differences as evidenced by these profiles. However, the MRQ was not sensitive to the primary task difficulty manipulation of targets.

None of the subscale ratings for SOs were correlated with performance. However, Auditory Linguistic subscale ratings were positively related to team performance for WDs ($r = .44$). One possible explanation is that teams performed better when they were more actively listening to and comprehending messages from their teammates.



Note: Due to space limitations only subscales that exceeded the 50% mark were included in the figure

Figure 4. Mean MRQ subscale scores for team role and communication modality

TWS

Ratings were submitted to a 5 (Subscale) × 2 (Team Role) × 2 (Targets) × 2 (CT) mixed-ANOVA, yielding main effects for team role, F (1,18) = 26.1, $p < .05$, CT, F (1,18) = 6.8, $p < .05$, and subscale, F (4,72) = 22.4, $p < .05$, as well as a two-way interaction of team role × subscale, F (8,144) = 4.74, $p < .05$. Mean TWS ratings were higher for WDs (M = 12.3, SE = .63) than for SOs (M = 7.8, SE = .63), and they were slightly higher for PC collaboration (M = 10.4, SE = .45) than for Voice only (M = 9.7, SE = .48). The interaction showed that team process demands were substantially lower for the SO role on all subscales except Communication, in which ratings between roles were not significantly different (Figure 5).

None of the TWS subscale ratings for WDs correlated with performance. However, the Monitoring subscale was inversely related to team performance for SOs (r = -.31).

The Team Role × Subscale interaction showed that the TWS is diagnostic in its ability to identify the WDs as primarily responsible for coordination and leadership functions, which is consistent with the role description of the WD. As with the MRQ, the TWS was not sensitive to the primary task difficulty manipulation of targets.

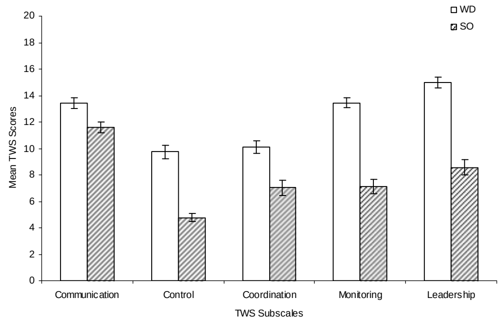


Figure 5. Mean TWS subscale scores as a function of team role

Discussion

The purpose of this study was to evaluate and compare the relative contribution of three subjective workload scales to our understanding of the demands of collaboration tools and team tasks. Ten teams were trained on an ABM scenario and tested under two communication modality conditions (Voice only and PC), and two levels of task difficulty (6 or 4 enemy targets). The latter factor was reflected in team performance scores.

Reliable workload measures are crucial to the evaluation of team tasks and the technologies employed for collaboration. Diagnostic workload instruments may be used to understand how to restructure team responsibilities, or design interfaces consistent with the communication demands of a task. A useful instrument must be sensitive to the demands of the task, teammates' responsibilities, and technology interface manipulations.

All three workload instruments were sensitive to the overall demands of the CT, however only the NASA-TLX was sensitive to the task difficulty manipulation, showing that more targets imposed greater temporal demands on the team. While the NASA-TLX was sensitive to the demands of the CT interfaces overall, it was not diagnostic in detecting differences between the team roles.

The TWS was sensitive to team roles, and indeed provided a useful profile that differentiated team process demands for the WD and SO responsibilities. Perhaps the MRQ was the most diagnostic with respect to CT and teammate responsibilities, in that the three way interaction provided descriptive profiles for both of these factors. In particular, it measured the high demands on vocal processes for WDs using voice only communication compared to the PC. The PC served to reduce vocal processing demands but resulted in higher spatial emergent processing and did not reduce workload overall. While there was no evidence that high vocal processing workload was related to poor performance, the descriptive nature of these results may still suggest a path forward for future CT development. For instance, although the PC relieved vocal process demands, the STM demands were quite high in both CT conditions. Future CT development should focus on strategies for reducing this source of workload.

The findings indicate that not one of the workload scales is sufficient in themselves at capturing the complete profile of team workload. Each scale has its own strengths and limitations in their ability to accurately discriminate or describe task demands. However, together these scales paint a descriptive picture of the demands placed upon the team.

Further research into the development of a team workload scale is necessary A reliable and valid team workload battery will be critical for the exploration of a range of collaboration technologies and design of team environments such as that used in command and control situations.

Acknowledgements

The authors gratefully acknowledge the contributions to this research of James Hyson, Brent Miller and Allen Dukes, without whose innovative and efficient software development, troubleshooting, and lab support, the experiment would not have been completed. Special thanks are also extended to Becky Brown and Dan Schwartz for their role in developing the CBT and data collection.

References

Boles, D. B., & Adair, L. P. (2001). The multiple resources questionnaire (MRQ). *Proceedings of the Human Factors and Ergonomics Society, 45*, 1790-1794.

Boles. D. B., Bursk, J. H., Phillips, J. B., & Perdelwitz, J. R., (2007). Predicting duel-task performance with the multiple resources questionnaire (MRQ). *Human Factors and Ergonomics Society, 49*, 32 – 45.

Bolia, R. S., Nelson, W. T., Vidulich, M. A., Simpson, B. D., & Brungart, D. S. (2005). Communications research for command and control: Human-machine interface technologies supporting effective air battle management. *Proceedings of the 10th International Command and Control Research and Technology Symposium.* Washington: Command and Control Research Program.

Bolstad, C.A. & Endsley, M.R. (2005). Choosing team collaboration tools: Lessons from disaster recovery efforts. *Ergonomics in Design, 13*, 7-14.

Cummings, M. L. (2004). The need for command and control instant message adaptive interfaces: Lessons learned from tactical tomahawk human-in-the-loop simulations. *CyberPsychology & Behavior, 7*, 653-661.

Finomore, V. S., Warm, J. S., Matthews, G., Riley, M. A., Dember, W. N., Shaw, T. H., Ungar, N. R., & Scerbo., M. W. (2006). Measuring the workload of sustained attention. *Proceedings of the Human Factors and Ergonomics Society 50th Annual meeting* (pp.1614-1618).

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload.* (pp. 139-183). Oxford, UK: North-Holland.

Hildebrand, G. A., Pharmer, J. A., Weaver, M. D. (2003). Performance and usability evaluation of an automated task management concept prototype design. *Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting*, 1913-1917.

Knott, B. A., Bolia, R. S., Nelson, W. T., Galster, S. M. (2006a). Effects of collaboration technology on the performance of tactical air battle management teams. *Proceedings of the Symposium on Human Factors Issues in Network-Centric Warfare,* Sydney, Australia.

Knott, B. A., Bolia, R. S., Nelson, W. T., Galster, S. M. (2006b). The impact of instant messaging on team performance, subjective workload, and situation awareness in tactical command and control. *Proceedings of the 11th International Command and Control Research and*

Technology Symposium. Washington: Command and Control Research Program.

O'Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology. In K. R. Boff, L. Kaufman & J. P. Thomas (Eds.), *Handbook of perception and human performance, vol. 2: Cognitive processes and performance*. (pp. 41-1-42-49). New York: Wiley.

Scott, S. D., Cummings, M. L., Graeber, D. A., Nelson, W. T., & Bolia, R. S. (2006). Collaboration technology in military team operations: Lessons learned from the corporate domain. *Proceedings of the 2006 Command and Control Research and Technology Symposium*. Washington: Command and Control Research Program.

Warm, J. S., Dember, W. N., & Hancock, P. A. (1996). Vigilance and workload in automated systems. In R. Parasuraman, & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (183-200). Mahwah, NJ: Erlbaum.

Wickens, C. D., & Hollands, J. G. (2000). *Engineering psychology and human performance* (3rd ed.). Upper Saddle River, NJ: Prentice-Hall.