JOURNAL OF MANAGEMENT AND ENGINEERING INTEGRATION

Scope: The Journal of Management and Engineering Integration (JMEI) is a double-blind refereed journal dedicated to exploring the nexus of management and engineering issues of the day. JMEI publishes two issues per year, one in Summer and another in Winter. The Journal's scope is to provide a forum where engineering and management professionals can share and exchange their ideas for the collaboration and integration of Management and Engineering research and publications. The journal will aim on targeting publications and research that emphasizes the integrative nature of business, management, computers and engineering within a global context.

Analysis of the Societal Elements of Chinese Quality Management

Ronald Shehane[1], Xiaoli Huan[1], and Adel Ali[2]
[1]*Department of Computer Science, Troy University, Pensacola, FL 32506, USA*
[2]*Department of Math, Science and Technology, University of Minnesota, Crookston, Crookston, MN 56716, USA*
shehaner@troy.edu, xhuan@troy.edu, adelali@crk.umn.edu

Abstract

In this study, we extend previous research that identified the strengths and weakness of the United States Action Plan for Import Safety (APIS) that was designed to improve the quality of imported goods into the United States. With the Chinese export market projected to exceed $1.2 trillion annually, the Chinese market represents an attractive option to western businesses. However, in 2007, in what the press referred to as "the year of China recalls", the Consumer Product Safety Commission recalled over 110 million Chinese products, the largest recall effort in recent history. Western companies have found that to be successful in commerce with China, they must not only take into account Chinese society but integrate its values into their management approach. The current study identifies the Chinese societal elements (cultural values, social interaction, historical influence, organizational structures, and economic conditions) that must be considered when developing effective quality management programs for outsourced goods and services. The result of the current study fills a gap in research literature by developing a relationship model that demonstrates the elements in Chinese society that influence the core requirements of successful quality management policy and execution. The results of this study are expected to serve as a guide for future research and development of a model that can be applied to quality management strategy for a wide variety of outsourced international operations.

1. Introduction

In 2007, in what the press referred to as "the year of China recalls", the Consumer Product Safety Commission recalled over 110 million Chinese products, the largest recall effort in recent history [1]. Recall efforts were not only focused on consumer products such as toys, but involved a wide range of other controversial and emotionally charged consumables such as foods and medicines [2][10]. The recall led to an emotional reaction from the public and calls for federal government action. To answer the calls for action, the federal government formed a multiple-agency federal committee, headed by the Secretary of Human Services, to develop an action plan of policy and programs to address the crisis. The committees developed a report which was presented to the President of the United States on November 2007. The report presented 14 key recommendations and 50 action steps designed to protect consumers from unsafe imported goods within the confines of preserving an active global market economy [3]. The title of the report was the "Action Plan for Import Safety" (APIS). Since the initial emphasis on import safety, there was a follow-up APIS report, Chinese recall bill action, and some U.S. regulatory implementation [4]. However, the problem with imports from China and other countries has continued to capture national interest from time-to-time. Recent outbreaks of import safety problems have been contaminated face paint and melamine tainted dog food [5].

The import safety problem is broad and complex. It involves 13 major government entities, the food and drug administration, and a variety of protection organizations. It also involves the coordination and development of agreements that include a broad range of national interests. The various nations involved represent different expectations and interest, $2 trillion in annual imports; 800,000 importers; countless manufacturers; and 300 ports-of-entry. The economic and safety impact is of such an enormous scale that the development of public policy related to import safety should be a matter of careful study and review [4].

In this study, we extend previous research [4] which identified the strengths and weakness of the United States Action Plan for Import Safety (APIS) that was designed to improve the quality of imported goods into the United States. Identifying the strength and weaknesses in APIS was an important step in addressing potential improvements in the quality management of imports. However, to fully address the difficulties of quality management issues that cross national boundaries, cultural, historical, and economic considerations must also be addressed. The challenge for businesses and government alike in conducting trade with China resides in deeply entrenched societal elements that includes an "elaborate set of unwritten rules, practices, and customs" [6]. This challenge is further complicated by the rapid changes and growth of China's involvement in the global economy. The challenge is further complicated by the rich historical setting that permeates China [6]. Our study highlights a range of societal elements in China that must be considered in ensuring quality management when doing business with China in relation to our previous research on Q8 and APIS. Our study focuses specifically on United States-China quality management issues due to its enormity of size and the impact it represents on total import safety.

The recent nature of the import safety subject represents a gap in research literature that is addressed in this study. Figure 1 is a conceptual framework that presents the eight core requirements required to support quality management – quality programs, measurement, contracts, decision support, continuous improvement, partnering, knowledge sharing, and multi-perspective viewpoint [7].

The framework in Figure 1, called Q8, was developed from research work addressing complex quality management problems where the use of contractual agreements for the production of goods and services was involved. The framework involved multiple perspectives of resolution that covered the field of law, contracting, technology, functional domains, multiple organizations, social, political, economic, ethical, and multilayered communication requirements [7]. The Q8 framework provides three layers of increasing specificity of the concepts that are part of the eight core requirement. A comparative analysis of Q8 versus APIS conducted by Shehane et al. [4], revealed shortfalls and opportunities in the eight core requirements required to support successful quality management efforts.

The remainder of this study is organized as follows. We present a brief description of our research plan. Following this the results of our analysis of the cultural, historical, and legal aspects of China that affect quality management are presented in relation to the Q8 framework core requirements they impact. The analysis will show the challenges and opportunities available to both governments and businesses in working with Chinese imports. Finally a conclusion of our research findings is presented along with a discussion of future research implications.

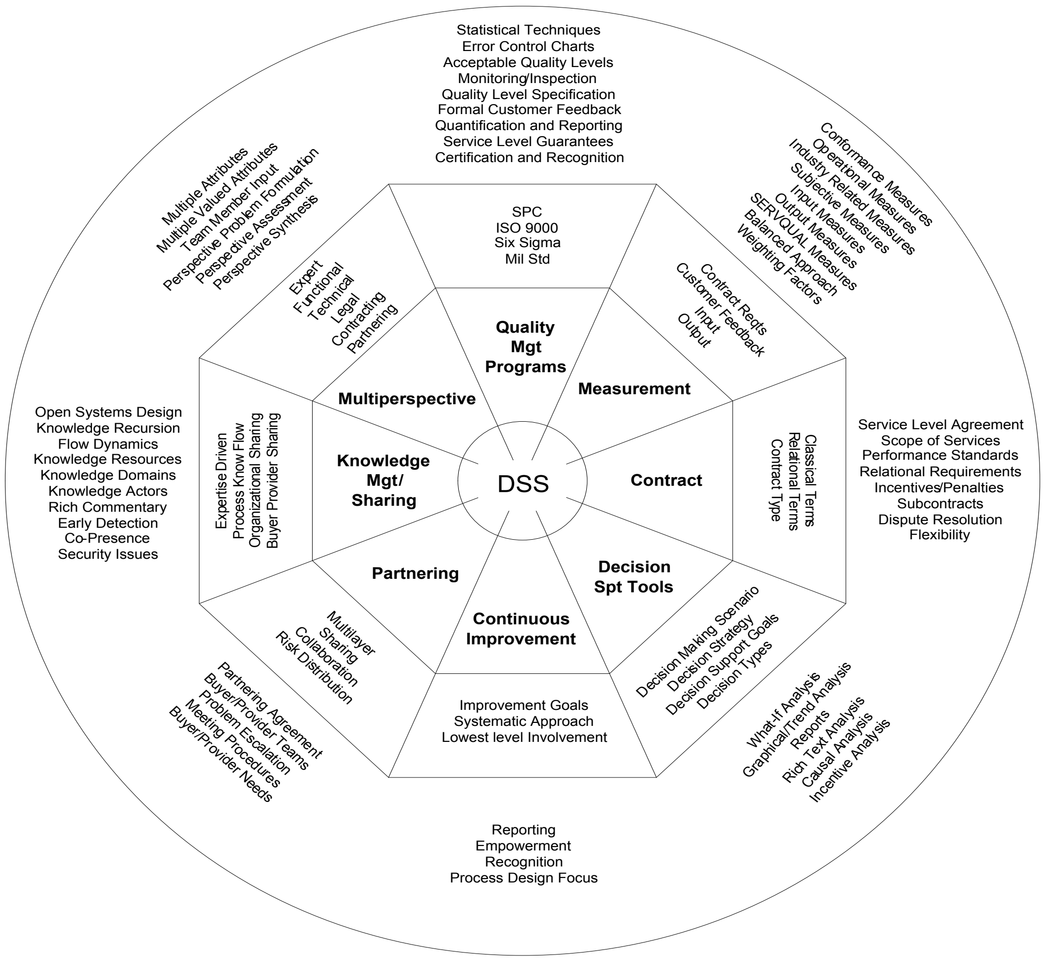


Figure 1. Q8 Conceptual Framework for Quality-Management. Note. From Outsourcing Management – Implementing Quality and Performance Decision Support, by R.F. Shehane, 2007. Saarbrücken, Germany: VDM Verlag Dr. Müller reproduced with permission.

2. Research Plan

The research question addressed in this study is “How the societal elements of China serve to enhance or detract from Quality Management efforts in the core requirements of Q8?” To identify the societal elements that impact on quality management in China, we conducted a literature research study of leading journals for the period 2000-2010. We structured our search to seek articles that discussed quality management subjects related to mainland China, Hong Kong, and Taiwan. Since research on China is somewhat immature and recent, we also broadened our search to news articles related to the unique societal elements of china and quality management issues.

3. Cultural and Historical Findings

3.1. Survival Mode (SV)

Even though China has progressed and become prosperous at a rapid pace, much of the nation's workforce is still in "survival mode". Concepts such as traceability or transparency, which are critical to quality management, do not have much meaning to those who are emerging from poverty [8]. Based on a comparison between Q8 and this aspect of Chinese society, it affects the quality management program in the Q8 framework.

3.2. Health Beliefs (HE)

While China's progress to prosperity has been an incredible journey, there are still cultural artifacts of the past that remain and affect their ability to manage quality. One example is the Chinese belief concerning human health which impacts on food products and manufactured goods. One of the beliefs in Chinese society is that excessive concerns with hygiene weaken ones immune system. This belief is especially demonstrated in the production and handling of food products [8]. The cultural climate, and the resulting personal belief produced from it, leads to complications for quality management efforts. The size of the undertaking to inspect and monitor all processes in a trillion dollar market comprising hundreds of thousands of business concerns negates any chance of identifying even a small percentage of defects. Monitoring and inspection efforts should of course continue, but other approaches such as efforts to train the workforce in areas such as hygiene, transparency, and traceability must also be involved [8]. Based on a comparison between Q8 and this aspect of Chinese society, it affects the quality management program in the Q8 framework.

3.3. Core Cultural Values (AD, HY, IN)

The research on the extent of the impact that core cultural values had on quality management was mixed. Researchers found that the Chinese core cultural values of abasement, adaptation, harmony, interdependence, and respect for authority had a positive influence on the quality management climate variables of organizational commitment, communication, unity of purpose, and management vision [9]. Noronha [9] further concluded that the success of quality management programs in Chinese companies is most dependent on how well the quality management program is infused with the accepted core cultural values. It should be noted that the Noronha's study was based on a sample of companies that already had successful quality management programs so it was more of a measure of how successful quality management efforts meshed with cultural values. Another study by Pun [11] found that "employee" acceptance or involvement in quality management was not related to core cultural values. Still other studies indicated that core cultural values influenced the way that managers looked at quality processes [12]. Based on a comparison of Q8 with these values within Chinese society, they all affect quality management in the Q8 framework.

Adaptation (AD). Chinese culture values collectivism over individualism. The reliance on collectivism leads to an emphasis on adaptive behavior rather than innovation which is contradictory to Western approaches to quality management. In addition, the preference for adaptive behavior results in more revisions to standard quality management approaches to mesh them with existing organizational culture [9]. Based on a comparison between Q8 and this aspect of Chinese society, it affects not only quality management programs, but also affects continuous improvement efforts in the Q8 framework.

Harmony (HY). The culture within China tends to refrain from acknowledging problems, especially those that may be personally disconcerting [8]. This cultural value is in opposition with basic quality management tenants where the willingness to identify and resolve problems is critical [8]. The resulting behavior from this cultural value represents a challenge for Western countries attempting to avoid the escalating problems and greater cost that such behavior creates. This behavior hampers continuous improvement efforts and the quality management precepts of transparency and traceability. The Chinese cultural emphasis on harmony and avoiding embarrassment offer a challenge that many Western negotiators often misinterpret. Chinese negotiators are eager to avoid situations that they view as antagonistic or that are aggressive in nature [13] which can hamper contracting and partnering efforts. The Chinese may become silent or attempt to withdraw temporarily from these types of situations [13]. This may often be misinterpreted by Western negotiators as a negotiating game, which is not the case, and can lead to a disruption in the successful conclusion of the contracting process [13]. The emphasis on harmony in Chinese society, affects not only quality management programs, but continuous improvement, contracting, and partnering efforts as well in the Q8 framework.

Interdependence (IN). The history of China's political structure and unequal power distribution results in a weak participatory process that extends to the factory floor and leads to the lack of transparency, traceability, and employee empowerment. Employees are not expected to exercise individual initiative. In fact, "good" supervisors are considered as those that hide problems [6]. This combination of concepts leads to a significant challenge in establishing open partnering efforts, effective quality management programs, and positive continuous improvement environments. Based on a comparison between Q8 and this aspect of Chinese society, it affects not only quality management programs, but continuous improvement, and partnering efforts as well in the Q8 framework.

3.4. Company Structures (CM)

Researchers agree that quality management in China has been slower to be implemented than in Western nations [12]. Research results have been mixed on the extent and degree of acceptance of quality management principles. China has been steadily transitioning from a centrally directed economy to a socialist economy. As a result, there are three types of companies in China – state owned, privately owned, and international joint Chinese ventures. The research revealed that state owned companies were the least advanced in implementing quality management versus privately held companies [12]. In addition, the research revealed that joint venture companies with foreign investors tended to be the most advanced in quality management initiatives [12]. Historically, the reform movement in China, where the government has been experimenting with different forms of state and private ownership has impacted on knowledge management practice used [14]. Based on a comparison between Q8 and the findings of this study, company structure affects quality management program and knowledge management/sharing in the Q8 framework.

3.5. Fragmented Structure (FR)

Although China is often thought of as being monolithic with direction only coming from the top, the development and enforcement of regulations applied to industry is significantly bureaucratic and fragmented [8]. Dong and Jensen [15] points out that there are 10 governmental ministries involved in establishing agricultural standards. The fragmented nature of the governmental organizations leads to difficulties with coordination and control, and thus quality management. As discussed in Roth et al. [8], the enforcement of regulations that provide consumer protection while encouraging economic growth represents a "delicate balancing act". The United States has had a head start of approximately a century in doing this when compared to China. Based on a comparison between Q8 and this aspect of Chinese society, it would affect the conformance measurement aspect of the Q8 framework.

3.6. Economic Pace (EP)

One of the problems affecting the Quality Management focus in China has been the rapid pace of economic growth and the resulting lack of time to transform and adjust to that growth [12]. Zhao et al. [12] discuss how China's rapid growth has resulted since the economic reforms instituted since 1978. The rapid growth has placed a strain on China's infrastructure and distribution as well as their ability to regulate [12]. The growth of exports over previous years for the period 2000 to 2008 averaged 29% with exports projected to exceed $ 1.2 trillion annually [16]. The implications for those that wish to be successful in doing business with China is that they must take into account the rapid pace of change and adjust accordingly [17]. This places a strain on businesses to obtain the latest information and to keep plans current. Based on a comparison between Q8 and this aspect of Chinese society, it would affect the quality management program in the Q8 framework.

3.7. Price Pressure (PR)

Roth et al. [8] has noted one historical trend that impacts on quality. The combination of fast economic growth experienced by China and the pressure by importers to pick the cheapest vendor has resulted in a belief among suppliers to omit quality and quote low. This historical factor has a significant impact on quality management programs and on the necessity to implement effective partnering principles. Based on a comparison between Q8 and this aspect of Chinese society, it would affect quality management program and partnering capabilities in the Q8 framework.

3.8. Long Term (LT)

Chinese culture is characterized by "long-term-orientation". This facet of their culture can be traced back 5000 years to its beginning in Confucian philosophy [6]. The Chinese long-term-orientation is exhibited in their preference for stability over change, which is somewhat counter to the tremendous economic growth and change that has occurred over the last 40 years. Their preference for stability, even in the presence of economic change, could be surmised as a precursor to their slower adjustment to newer concepts such as quality management. One manifestation of Confucianism is in their view of customers. Zhao et al. [6] reports that this cultural preference has led to a difficulty in utilizing customer views and feedback, which is an essential part of maintaining quality. Based on a comparison between Q8 and this aspect of Chinese society, it would affect quality management program implementation in the Q8 framework.

3.9. Decision Support (DS)

Tian et al. [18] discussed that the concept of decision support tools were accepted in China as early as the 1980s and had even been a state sponsored activity. The use of Decision Support Systems (DSS) both in private and government controlled businesses is reported to be operational and contributing in areas such as policy making, business development, transportation, manufacturing, and education. The university systems in China are also heavily invested in offering degrees in the decision sciences. The silence within APIS to address the need for decision support or consider decision support in the problem formulation of APIS represents a major failure of the United States in relating to China concerning import safety. Based on a comparison between Q8 and this aspect of Chinese society, it would affect the decision support effort in the Q8 framework.

3.10. Negotiations (NG)

The concept of contracting was overlooked in APIS. This was considered a major shortfall since a critical element to any import relationship is the existence of a contract that defines mutual responsibilities and requirements [19] [20]. A well written and negotiated contract establishes quality standards, defect and recall insurance, and legal recourse for failures to perform [19]. U.S. buyers in China will often make the mistake of accepting a purchase order in lieu of a formal contract [19]. These purchase orders may or may not contain terms or conditions and are not sufficient in detail to address the terms of a formal contract which leaves the buyer without legal recourse [19]. Chinese suppliers offer challenges unique to their culture. Handfield and McCormack [17] warns U.S. negotiators that "the signing of a contract is just the beginning of real negotiations." Chinese business culture has some unique views that may be missed by U.S. negotiators. Chinese negotiators believe in bringing a large number of managers to meetings so that U.S. negotiators are often outnumbered in their negotiating sessions [17]. In addition Chinese negotiators tend to stretch out meetings to last for many days and thus wear down their opposition [17]. These aspects of Chinese negotiation need to be considered with U.S. negotiators being willing to approach negotiations in a slow, careful, and deliberate manner. Zhu et al. [21] offered the recommendation that in their efforts to ensure an understanding of Chinese culture, Western negotiators should also convey their cultural values to their Chinese negotiators. Based on a comparison between Q8 and this aspect of Chinese society, it would affect contracting and partnering efforts in the Q8 framework.

3.11. Quanxi (QX)

Another aspect of contract negotiations that must be considered in China is the concept of "quanxi". Quanxi is a pervasive cultural aspect of Chinese society that is highly focused on social networks [22] [23]. The buyer-supplier business relationship is not as widely accepted in China as in Western nations. Instead, the Chinese tend to place their confidence in personal trust-based and favor-based relationships generally referred to as quanxi [22]. The reliance on quanxi is thought to have developed from Chinese culture of distrust for outsiders and the weakness of their legal system. Davison and Ou [23] expressed the view that the influence of quanxi may be considered as more important than quality or price in negotiations. In fact, Tung and Worm [24] found that the outcome of quality and price are strongly influence by the quality of quanxi between negotiators. In studies of customer-suppler pairings, Zhao et al. [12] reported that shared vision and common goals helped to improve trust which is supportive of quality improvement efforts. An additional ray of hope for those wishing to develop effective supplier-customer relationships not based on quanxi is reflected in an ancient Chinese proverb that states “there are no eternal friends and no eternal enemies, only interests” [6]. This finding provides a possible solution for Western negotiators wishing to develop a partnering arrangement. The need for careful monitoring of products and services is important in all supplier-buyer relationships. Davison and Ou discussed the special need for this in Chinese markets. Davison and Ou point out that quality specifications and expectations often are not used to judge quality, but that perceived quality is more dependent on relationship factors such as quanxi. Another aspect of Quanxi was discussed by Peng et al. [14] who found that the transfer of knowledge by small social networks and the principles of quanxi is preferred by the Chinese [14]. Based on a comparison between Q8 and the findings of this study, quanxi is the most pervasive in its impact on Q8. The various aspects of quanxi affect quality management program, measurement, knowledge management/sharing, contracting, and partnering efforts in the Q8 framework.

3.12. Regional Differences (RG)

Contrary to popular belief, China is not a monolithic political structure but contains significant regional influences. China can be viewed as containing seven different regions in that can differ significantly. The economic growth in China has been significant, but the rate of growth has been different in the various regions of China [25]. In addition, each region in China has a diverse background in their development, culture, geography, and maturity in international trade. There can be a wide variation between central and local government that is characterized by local protectionism and differences in the application of rules and regulations that impacts on quality management [12]. In addition, due to its size the Chinese government has delegated much of the legal and regulatory enforcement to local and regional authorities [8]. Added to this is the encouragement of the Chinese government of competition between the regions [6]. As a result, the regional implications offer a particular challenge to quality management that cannot be ignored. Peng et al. [14] discussed that regional differences impact on knowledge management practices, where some regions are more advanced than others or have their own particular twist on the knowledge management implementation [14]. Based on a comparison between findings of this study versus Q8, the regional difference aspect of Chinese society affects quality management program and knowledge management/sharing efforts in the Q8 framework.

4. Conclusions and Future Study

Figure 2 shows a model of the relationships between the societal elements of China and the eight core requirements identified in the Q8 framework needed to establish effective quality management. The Q8 framework was proved to be useful as a comparative tool for analyzing and categorizing the elements of Chinese society that impact on quality management. The study revealed the cultural values, social interaction, historical influence, organizational structures, and economic conditions that should be considered in developing a more robust quality policy that may help meet the overall quality management goals originally stated in APIS.

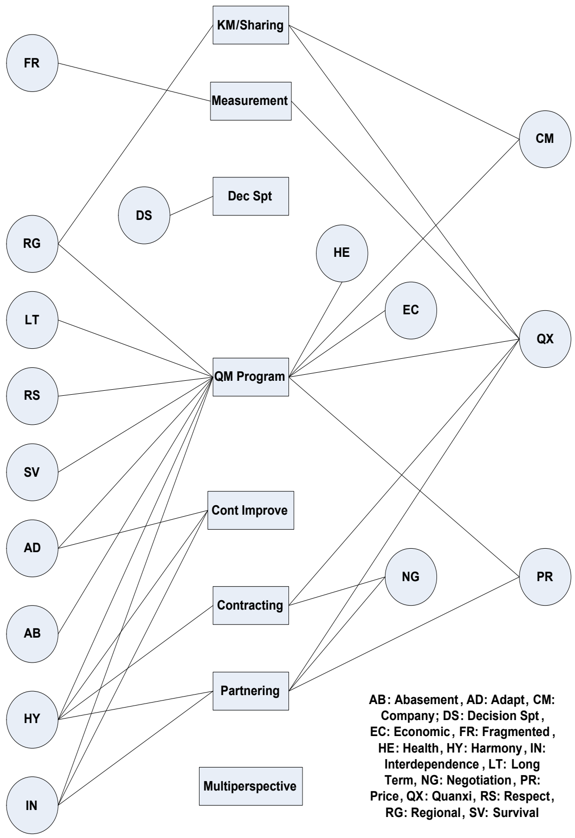


Figure 2. Societal Relationship Model

The authors' intention is to expand the Q8 framework to include the societal elements presented in this study. It is hoped that future studies will further define the interactive nature and dynamics of the relationship of the various societal elements with the individual concepts present within Q8. Future research needs to be applied to determine the quantitative degree of the relationship between each societal element and the individual core requirement present in Q8.

A question worthy of future study is whether Western or Chinese cultural practices will become dominant or whether they will become a complimentary amalgam. Also, the degree that cultural practices impact on employee acceptance of quality management programs needs to be investigated further in view of the Pun's [11] contradictory findings.

In addition, as recognized in our original study, further study is needed to assess the role that U.S. industry needs to play in ensuring import-safety. One must ask the question as to what role weak engineering design played in the Mattel toy scare and what types of U.S. company inspections were employed to avoid lead paint content and other dangerous qualities in imported goods. There are examples that suggest the need for a greater portion of "sharing of the blame" on the U.S. companies part versus the Chinese [1] [2].

5. References

[1] Goodden, R.L. Better safe than sorry. Quality Progress, 41(5), 28-34, 2008.

[2] Field, A.M. Riding out the storm. Journal of Commerce, February, 2008.

[3] Action plan for import safety, a report to president, 2007. http://archive.hhs.gov/importsafety/, retrieved on February, 6th, 2011.

[4] Shehane, R.F., Huan, X., & Ali. A. A framework analysis of the Action Plan for Import Safety (APIS). Journal of International Business and Cultural Studies, volume 5, 2010, retrieved from http://www.aabri.com/jibcs.html.

[5] Hope, F. Paint a sad face – another China product recall. China, 2009. Retrieved from http://futurenewstoday.blogspot.com/2009/05/paint-sad-face-another-china-product.html

[6] Zhao, X., Flynn, B., & Roth, A. Decision sciences research in China: a critical review and research agenda-foundations and overview. Decision Sciences, 37(4), 451-496, 2006.

[7] Shehane, R.F. Outsourcing management – implementing quality and performance decision support, Saarbrücken, Germany: VDM Verlag Dr. Müller, 2007.

[8] Roth, A., Tsay, A., Pullman, M., & Gray, J. Unraveling the food supply chain: strategic insights from China and the 2007 recalls. Journal of Supply Chain Management, 44(1), 22-39, 2008

[9] Noronha, C. Chinese cultural values and total quality climate. Managing Service Quality, 12(4), 210-223, 2002. doi: 10.1108/09604520210434820.

[10] Brockwell, I. Warning – Halloween Chocolate from China and Melamine Danger, 2008. Retrieved from American Chronicle, http://www.americanchronicle.com/articles/view/79426

[11] Pun, K.F., & Ho, K.Y. Identification of service quality attributes for restaurant operations: A Hong Kong case. Managing Service Quality, 11, 233-240, 2001.

[12] Zhao, X., Flynn, B., & Roth, A. Decision sciences research in China: current status, opportunities, and propositions for research in supply chain management, logistics, and quality management. Decision Sciences,41(1), 2007.

[13] Ma, Z. Negotiating into China: the impact of individual perception on Chinese negotiation styles. International Journal of Emerging Markets, 1(1), 64-83, 2008. Doi: 10.1108/17468800610645013

[14] Peng, J., Li-Hua, R., & Moffet, S. Trend of knowledge management in China: challenges and opportunities. Journal of Technology Management in China, 2(3), 198-211, 2007. doi: 10.1108/17468770710825142

[15] Dong, F., & Jensen, H. Challenges for China's agricultural exports: compliance with sanitary and phytosanitary measures. Choices, 22, 19-24, 2007.

[16] China Exports. CIA World Factbook, 2010. Retrieved from https://www.cia.gov/library/publications/the-world-factbook/rankorder/2078rank.html

[17] Handfield, R, & McCormack, K. What you need to know about sourcing from China. Supply Chain Management Review, 9(6), 28-36, 2005.

[18] Tian, J., Wang, Y., Li, H., Li, L., & Wang, K., DSS development and applications in China. Decision Support Systems, 42, 2060-2070, 2007. doi: 10.1016/j.dss.2004.11.009.

[19] Fremlin, G.P. Careful contracts reduce risk. The China Business Review, 35(1), 34-39, 2008.

[20] Jin, E., & Lucy, G. Market on the move. Supply Management, 13(9), 34-35, 2008.

[21] Zhu, Y., McKenna, B., & Sun, Z. Negotiating with the Chinese: success of initial meetings is the key. Cross Cultural Management: An International Journal, 14(4), 354-364, 2007. doi: 10.1108/13527600710830368

[22] Anderson, A., & Lee, E. From tradition to modern attitudes and applications of quanxi in Chinese entrepreneurship. Journal of Small Business and Enterprise Development, 15(4), 775-787, 2008. doi: 10.1108/14626000810917852

[23] Davison, R., & Ou, C. Quanxi knowledge and online intermediaries in China. Chinese Management Studies, 2(4), 281-302, 2008. doi: 10.1108/17506140810910935

[24] Tung, R.L. & Worm, V. "Network capitalism: the role of human resources in penetrating the China market", International Journal of Human Resource Management, 12(4), 517-534, 2001.

[25] Lin, J., Cai, F., & Li, Z. Social consequences of economic reform in China: Analysis of regional disparity in the transition period. In M. Renard & E. Elgar (Eds.), China and its regions. Northhampton, MA: Cheltenham, 33-86, 2002.

Ethics and Project Management: A Journal Publication Analysis

Stephen L. Allen
School of Business, Truman State University, Kirksville, MO 63501, USA
sallen@truman.edu

Abstract

The ethical role of a manager is particularly essential in project management given the typical complexities of scope, considerable budgets and schedule constraints that face project managers today. The Project Management Institute (PMI) has developed and adopted a Code of Ethics and Professional Conduct to support the work of project managers. Ethical challenges faced by project managers, as identified in the literature, are examined by an analysis of top project management journals to answer questions regarding the study of ethical challenges to project managers in the context of the PMI Code of Ethics and Professional Conduct. This article examines the link between ongoing ethics research in project management and that of the guiding principles of the Code of Ethics and Professional Conduct as outlined by the Project Management Institute. The results indicate a variety of research connections with the Code of Ethics and Professional Conduct, which confirms the idea that the code is helpful in supporting the actions of project managers.

1. Introduction and Background

In the complicated world of project management, the project manager is likely to face many ethical issues and decisions within the project's scope. In an effort to support project managers' ethical complexities, a guide to ethics has been revised and adopted by the Project Management Institute (PMI). This guide is the PMI Code of Ethics and Professional Conduct which provides an ethical framework to assist project managers to maintain high ethical standards in their work. The PMI Code of Ethics and Professional Conduct has been in existence in some form since the 1980's with its current form ratified by the organization in 2006. The PMI Code of Ethics and Professional Conduct is intended as a tool to guide project management practitioners in the execution of their management efforts within accepted standards of ethics and with high integrity.

Current ethical challenges that project managers face exhibit a variety of situations that may be categorized through an analysis of top project management journals to answer several questions regarding the study of project management ethics in the context of the PMI Code of Ethics and Professional Conduct. Current project management research efforts containing ethical elements in recent journal articles were benchmarked to the key ethics values of responsibility, respect, fairness and honesty in the PMI Code of Ethics and Professional Conduct.

While ethics elements in project management situations have been studied over time, benchmarking current research efforts to the latest version of the PMI Code of Ethics and Professional Conduct is lacking in the literature.

This research begins to fill this acknowledged void within the field of project management.

1.1 Historical perspective

A review of history often proves to be beneficial for individuals and groups when attempting to examine the status of a current situation. Throughout history, a variety of project efforts have been in place worldwide. Historically project outcomes have been both successful and tragic over time. History is rife with examples of poor ethical frameworks regarding project efforts. Consider the massive construction project for any ancient pyramid—loss of life was predictably substantial. For these project efforts, established ethics of the time period likely accepted loss of life with little concern for the human loss except from the point of view of utility.

On July 17, 1981, the suspended flooring in the Kansas City Hyatt Regency Crown Center collapsed during a gala. The tragedy killed 114 people and injuring 216 others [1]. While this failure was ultimately deemed to have been an engineering design flaw, better ethical communications during both design and construction phases might have, in the end, uncovered this flaw, preventing injuries and loss of life. [2]

The collapse of the Enron Corporation in 2001 is another example of how poor ethical standards may contribute to the failure of project efforts. As Enron moved from being an energy producer to being an integrated energy corporation, deregulation enabled the opportunity for greed in these new markets. This opportunity came about with the means of mark to market accounting practices which pushed reporting limits that ultimately were approved by the Enron Board [3]. Interestingly, Enron's stated corporate values were Respect, Integrity, Communication and Excellence [4], which are similar to the key values of the PMI Code of Ethics and Professional Conduct.

Most recently, the multi-billion dollar Ponzi scheme perpetrated by former NASDAQ Chairman Bernard Madoff in 2008 adversely affected many lives, creating a loss of trust with regard to financial stock market control in the United States. While the complete extent of this fraud may never be known, its magnitude indicates the existence of ethical miscues by regulatory personnel during the preceding years throughout which Madoff was investigated [5].

Just what drives such extremes of unethical behavior? Has training in ethics or business ethics failed in the business and engineering curriculums? Much of the work in corporations is project-centered which makes up a major portion of a firm's efforts. Unethical choices made during the execution of these efforts may lead to disastrous results later for the firm. Are researchers attempting to uncover a response to these ethical situations? What can be done to improve matters?

1.2 Ethics background

Ethics is an element that sets humans apart. Failure to maintain high ethical standards may yield extreme misfortunes for an individual or organization. Sadly, these extremes of misfortune may range from significant financial loss to the most extreme, loss of life. Yet while there may be no single agreed upon definition of ethics, the collection of ethics theories leads to somewhat dissimilar notions. As a consequence, perspectives of ethics theories, the use of ethics and business ethics will be reviewed. Considerable ongoing project management research exists regarding the optimization of effective scope, schedule and budget management as a consequence of cost and resource use efficiencies. While considerable research effort has been expended with regard to the three key project management tenets, the successful accomplishment of a project would expect to be executed within an ethical framework. It is of interest to note that the Guide to the Project Management Body of Knowledge (PMBOK) outlines, defines, and aids project managers in the practice of project management, yet a discussion of ethics is noticeably absent from the PMBOK [6].

The origins of modern ethics theories include utilitarian, deontological and virtue components. The basic tenet of utilitarian ethics theory is to act in such a way as to aid the general good while avoiding acts that do not help the general good. Deontological ethics theory has an expectation of duty—the duty to do what is expected despite the possibility of causing a conflict with promoting the general good. Virtue ethics theory suggests that a person should keep in mind that their actions should have their basis in what a selfless person would do as opposed to the action of self promotion. These fundamental ethics theories generate some notion of what ethics is and how it might be defined for business.

One definition of ethics may be defined as, "the study of what constitutes right or wrong behavior….Ethics has to do with the fairness, justice, rightness or wrongness of an action" [7]. Another perspective holds that "Ethics…encompasses the system of beliefs that supports a particular view of morality" [8]. Another interpretation indicates that ethics may be defined as "inquiry into the nature and grounds of morality where the term morality is taken to mean moral judgments, standards and rules of conduct" [9]. A more concise definition of ethics might be "the systematic attempt to make sense of individual, group, organizational, professional, social, market and global moral experience in such a way as to determine the desirable, prioritized ends that are worth pursuing, the right rules and obligations that ought to govern human conduct, the virtuous intentions and character traits that deserve development in life, and to act accordingly" [10].

A view of business ethics in particular might be defined as "the principles and standards that determine acceptable conduct in business organizations" [11]. Another view defining business ethics is that business ethics "refers to values, standards, and principles that operate within business" [12]. Another alternative defines business ethics as the action that "comprises the principles and standards that guide behavior in the world of business" [13].

Today's society demands the practice of high ethical standards in business, business practices and engineering efforts. Given the variety of ethics theories, definitions and business ethics definitions, a PMI Code of Ethics and Professional Conduct has been adopted. The purpose of the Code is to serve and guide the practitioners of project management efforts in today's society.

1.3 Literature background

When reviewing the available project management literature, a considerable number of research articles regarding the many project management process groups—initiating, planning, executing, monitoring and controlling, and terminating—have a primary focus on the technical aspects of the process element studied.

Current project management textbooks also focus on these topics of scope, budget, and schedule [14-16]. This is a natural result of project management research as outcomes such as cost have a somewhat clearer standard of measure through the investigation of cost variance analysis.

When examining current project management journal articles, fewer journal articles are found regarding the less technical aspects of the project management processes. Even fewer articles deal with broader aspects of project management such as the overlay of ethics within the context of project management. Several earlier research works investigated particular project management research topics and themes studied and reported over time. While the particular research methodology varied between articles, the conclusions of these articles reported the shifting research emphasis areas as the project management field has changed and developed over time [17-22]. In a recent article reviewing ten years of project management journal research, the authors confirmed that the bulk of the project management research focus areas and topics dealt with scope, budget and scheduled elements and to a lesser extent with minority coverage of less technical project elements and topics. The results of this comprehensive work indicated very little research coverage of ethics within the project management research journals [23]. As a consequence the current body of research makes few connections to the PMI Code of Ethics and Professional Conduct.

This research effort reviews recent studies regarding the potential for ethical missteps within the context of project management published in journals specific to project management. While other journals exist that study the context of ethics from a larger perspective or application area, these other journals were not utilized in order to focus this investigation on the narrow application area of project management.

2. Research Questions

The PMI Code of Ethics and Professional Conduct outlines four specific values that support the code: responsibility, respect, fairness, and honesty. Responsibility is defined as "…our duty to take ownership for the decisions we make or fail to make, the actions we take or fail to take, and the consequences that result." Respect is defined as "…our duty to show high regard for ourselves, others and the resources entrusted to us. Resources entrusted to us may include people, money, reputation, the safety of others, and natural or environmental resources." Fairness is defined as "…our duty to make decisions and act impartially and objectively. Our conduct must be free from competing self-interest, prejudice, and favoritism." Honesty is defined as "…our duty to understand the truth and act in a truthful manner both in our communications and our conduct."

These four key PMI Code of Ethics and Professional Conduct values are in place to guide practitioners of the profession. One key research question is, "Is there any connection of ethics research within the context of project management and, if so, how does it relate to the PMI Code of Ethics and Professional Conduct four key values?

In an effort to establish a current benchmark, research questions were developed to investigate the current study of ethics, ethical miscues, and misstep potentials and challenges identified in recent project management journal articles. The research questions investigated the following areas: To what extent was ethics studied or investigated within project management within the benchmark duration? What implications, if any, existed with regard to the authors' cultural and geographic area? What was the benchmark relationship of ongoing ethics research within project management and how did the research context link to the PMI Code of Ethics and Professional Conduct values of responsibility, respect, fairness, and honesty?

3. Methodology

The overall design of this research is similar to Littau's methodology [24] of identifying a set of high quality project management research journals with which to establish a benchmark of data to respond to identified research questions. Initially, the benchmarking process would identify recent project management research articles potentially involving ethics within the context of project management. Next, a process screen was developed to extract potential journal articles that might provide a fit for the questions posed by this research. Once articles were selected as potential research fits, these articles were then read in their entirety to clarify each article's suitability and category to respond to the research questions. Finally the data was summarized to provide meaningful results leading to conclusions regarding the status of ongoing research efforts regarding ethics within the context of project management. The project management ethics context was then examined for its linkage supporting the values stated in the PMI Code of Ethics and Professional Conduct.

3.1 Benchmarking process

Littau's [24] process of journal identification was used to benchmark the study of ethics as applied within the context of project management. The first step was to identify a representative group of project management journals from several journals that study projects and their implications. Four project management related journals were selected for this research—*Project Management Journal (PMJ), International Journal of Project Management (IJPM), International Journal of Managing Projects in Business, (IJMPB) and International Journal of Project Organisation & Management (IJPOM)* as representative of top global outlets for project management research. This study was interested in recently investigated results and thus limited itself to the preceding five years of 2005-2009 to establish the benchmark. In this time interval 152 journal articles were published in the *PMJ,* 387 articles were published in the *IJPM,* 70 articles were published in the *IJMPB,* and 23 articles were published in the *IJPOM*. Thus in total 632 journal articles were initially investigated for this article.

3.2 Screen development/article selection

The next step was to develop a useable screen to identify articles that may have a primary or secondary relationship to ethics study within the project management context. A set of screening terms was chosen using definitions and explanations found in the PMI Code of Ethics and Professional Conduct. The titles, abstracts, key words and introductions for all 632 journal articles were read to ascertain the likelihood of identifying the research in question for this study. Once journal articles were identified, those identified articles were read in their entirety subject to data extraction as defined in the next section. Sample keywords included ethics, leadership, organization, culture, cultural, trust, strategy, strategic, relations, relationships, teams, social, behavior, society, personality, code of ethics, code of conduct, knowledge, values, competencies and the like.

3.3 Article examination and data extraction

Once all 632 journal articles were initially examined, a subset of journal articles was read completely. These articles were read first to ensure suitability to the study of ethics with the context of project management. Once identified as a qualified journal article, that article was reread to link its content to the PMI Code of Ethics and Professional Conduct values of responsibility, respect, fairness, and honesty as outlined by the PMI and listed in section two of this article.

4. Results

The results of this study were analyzed in three ways. First there was a basic investigation into each article's background. The cultural origination of the articles by authorship and by country of authorship was investigated. And finally, any ethics related coverage was linked to the PMI Code of Ethics and Professional Conduct four key values.

4.1 Article Background Information

There were a total of 632 journal articles initially screened for a related ethics component. Of the total number of journal articles screened, twenty-six articles were deemed to contain an ethics element that could be linked to a component of the PMI Code of Ethics and Professional Conduct. Thus four percent of the published project management journal articles examined dealt with an ethics element suitable for this study. The linkage rate of ethics in project management journals found in this study is comparable to that of another significant project management journal analysis study [24].

4.2 Cultural Origination

Of the twenty-six qualifying journal articles, there was a combined authorship of 57 different authors. The majority of the authors were researchers (50) combined with practitioners (7). There were many countries of origin represented by the authors: Australia, Austria, Canada, England, France, Germany, Iceland, Israel, Japan, Norway, Singapore, Sweden, Turkey and the United States. The geographical areas were predominantly of Asian, European and North American. It is noteworthy that there were no authors from South America or Africa.

The particular study geographic varied. There were twelve country specific articles and fourteen other various situations. The specific country or country comparisons included: Australia/United Kingdom/United States; Germany/Canada; Germany/Sweden, Hong Kong, Iceland; Israel; Japan; Norway, Turkey; the United Kingdom and the United States. Naturally, the geographic study areas tied closely to those of the articles' authors.

4.3 Relationship to PMI Code of Ethics and Professional Conduct

The qualified journal articles for the study were read specifically for linkage to the four PMI Code of Ethics and Professional Conduct areas of responsibility, respect, fairness and honesty. The qualified articles were judged to be connected to one of the four defined ethic areas by article content. If two or more of the areas were potentially connected, the article was assigned to the primary context as offered in the article. Article area versus study year is shown in Table 1 below.

Table 1, Study Year Results

Area	2005	2006	2007	2008	2009
Respon.	3	1	2	4	2
Respect	0	0	1	2	0
Fairness	0	1	1	0	1
Honesty	1	1	0	1	5
Total==>	4	3	4	7	8

The average number of articles identified with linkages to the PMI Code of Ethics per year during the study period was 5.2 with a standard deviation of 2.2 articles. A times series regression of articles with linkages versus year yielded a measurable growth in articles published during the study interval supported by an r-square value of + 0.766.

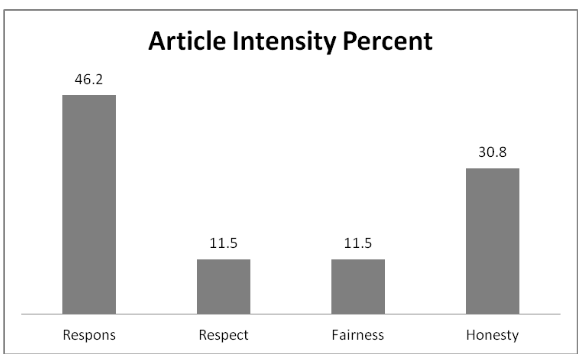


Figure 1. Article Intensity Percent

The ethics area intensity during the study period is shown in Figure 1 above. However, there are varying results by article intensity versus the four PMI Code of Ethics and Professional Conduct areas. The combined areas of responsibility and honesty garnered the overwhelming majority of article investigations representing a combined 77% of the research efforts. This preponderance of work can be explained as the area of responsibility and to a lesser extent honesty can be more easily tied to the primary project management research areas of scope, budget and schedule. The ethics areas of respect and fairness received less research consideration with a combined 23% of the research work in these areas.

5. Conclusions

A large number of project management research articles were initially investigated to address the research issue of ethics within the context of project management. Twenty-six articles were identified by the research methodology that exhibited connections to an ethics context and related to the PMI Code of Ethics and Professional Conduct.

The authorship origin of the qualified articles varied greatly across North America, Europe, the Near East and Asia. There were noteworthy absences of authorship from the areas of South America and Africa. The authors were predominately researchers as opposed to project management practitioners. The research geographical study areas tended to be within the confines of the authorship. The geographical study area of an article tended to tie to the author's origin, thus there was a lack of cross-cultural examination within the context of the connections of ethics and the PMI Code of Ethics and Professional Conduct. This paper establishes a benchmark regarding the cultural identity of research studies and researchers in this area that will benefit researchers studying the evolution of ethics within project management in coming years.

Throughout the span of this benchmarking study interval, a growth trend was supported over time with respect to the frequency of journal articles related to ethics within project management. Of the journal articles analyzed in this study, the ongoing research emphasis overwhelmingly emphasized the PMI Code of Ethics and Professional Conduct values of Responsibility and Honesty. The values of Respect and Fairness were examined to a lesser extent with regard to the PMI Code of Ethics and Professional Conduct. One can conclude that the scarcity of research in the area of Respect and Fairness may be a result of either a lack of interest in this particular research area or perhaps is due to a more challenging effort in assessing the areas of Respect and Fairness.

A research void was identified in the project management literature in the area of ethics and its linkage to the Project Management Code of Ethics and Professional Conduct. This research effort begins to fill that void. A growing trend in the research of specific ethical issues within project management was identified. The intensity of research, however, was imbalanced with respect to the values suggested by the PMI Code of Ethics and Professional Conduct. Further research could be expected in the area of assessing the effectiveness of the Code of Ethics and Professional Conduct for the project management profession.

6. References

[1] S. Woodward and R.J. Pessek, The Hyatt Horror, *Kansas City Star*, July 19, 1981, pp. 1HR-10HR.

[2] H. Petroski, *Design Paradigms: Case Histories of Error and Judgment in Engineering*, Cambridge University Press, Cambridge, UK, 1994, p. 61.

[3] L. Fox, *Enron: The Rise and Fall*, John Wiley and Sons, Hoboken, NJ, 2003, p. 158.

[4] B. Cruver, *Anatomy of Greed: The Unshredded Truth from an Enron Insider*, Carroll & Graf, New York, NY, 2003, pp. 42-43.

[5] D.B. Henriques, *The Wizard of Lies: Bernie Madoff and the Death of Trust*, Henry Holt and Company, New York, NY, 2011, p 210.

[6] *A Guide to the Project Management Body of Knowledge*, Project Management Institute, Newtown Square, Pennsylvania, 2003.

[7] Clarkson, K.W., R.L. Miller, and F.B. Cross, *Business Law*, South-Western, Mason, OH, 2012, p. 93.

[8] L.T. Hosmer, *The Ethics of Management*, McGraw-Hill Irwin, New York, NY, 2006, p. 92.

[9] P.W. Taylor, *Principles of Ethics: An Introduction to Ethics*, Dickenson Series of Philosophy, Encino, California, 1975, p. 1.

[10] Petrick, J.A. and J.F. Quinn, *Management Ethics: Integrity at Work*, Sage Publications Inc., Newbury Park, California, 1997, p. 42.

[11] O.C. Ferrell, G.A. Hirt, and L. Ferrell, *Business Essentials,* McGraw Hill Irwin, New York, NY, 2011, p. 36.

[12] J. DesJardines, *An Introduction to Business Ethics*, McGraw-Hill, New York , NY, 2009, p. 9.

[13] O.C. Ferrell, J. Fraedrich, & L. Ferrell, *Business Ethics*, Houghton Mifflin, Boston, Massachusetts, 2008, p. 6.

[14] J.R. Meredith and S.J. Mantell, *Project Management: A Managerial Approach*, Wiley, Hoboken, NJ, 2009.

[15] H. Kerzner, *Project Management: A Systems Approach to Planning Scheduling and Controlling*, Wiley, Hoboken, NJ, 2009.

[16] T.J. Kloppenborg, *Contemporary Project Management*, Cengage, Mason, OH, 2012.

[17] M. Betts and P. Lansley, International Journal of Project Management: A Review of the First Ten Years, *International Journal of Project Management*, Vol. 13, No. 4, 1995, pp.207-217.

[18] T.J. Kloppenborg and W.A. Opfer, *Forty Years of Project Management Research: Trends Interpretations and Predictions*, PMI Research Conference, PMI, 2000, pp. 41-59.

[19] P.W.G. Morris, M.B. Patel and S.H. Wearne, Research into Revising the APM Project Management Body of Knowledge, *International Journal of Project Management*, Vol. 18, No. 3, 2000, pp. 155-164.

[20] G. Thermistocleous and S.H. Wearne, Project Management Topic Coverage in Journals, *International Journal of Project Management*, Vol. 18, No. 1, 2000, pp. 7-11.

[21] B. Urli and D. Urli, Project Management in North America, Stability of the Concepts, *Project Management Journal*, Vol. 31, No. 3, 2000, pp. 33-43.

[22] A.M. Zobel and S.H. Wearne, Project Management Topic Coverage in Recent Conferences, *Project Management Journal*, Vol. 31, No. 2, 2000, pp. 32-37.

[23] L. Crawford, J. Pollack and D. England, Uncovering Trends in Project Management: Journal Emphases Over the Last 10 Years, *International Journal of Project Management*, Vol. 24, 2006, pp. 175-184.

[24] P. Littau, N.J. Jujagiri, and G. Aldbrecht, "25 years of Stakeholder Theory in Project Management Literature," *Project Management Journal*, Wiley, Vol. 41, No. 4, pp. 17-29.

Real Time Object Tracking: Using Graph Edit Plus and Video Analyst

Anim Adrian Amarsingh, Kofi Nyarko and Cynthia Brown-LaVeist
Department of Industrial and Systems Engineering, Morgan State University,
Baltimore, MD 21251, USA
anama1@morgan.edu, Kofi.Nyarko@morgan.edu, Cynthia.BrownLaveist@morgan.edu

Abstract

Unmanned Aerial Vehicles (UAVs) have the potential to be utilized for a wide variety of applications, such as military, law enforcement and search and rescue. UAVs typically have cameras as well as other vital instrumentation for measuring and recording data. In order to obtain maximum benefit from video captured on UAVs, it is important that this video be able to be processed in real time. Real Time Object tracking enables more efficient use of incoming data. This paper will focus on performing real time object target tracking on UAV video footage through the use of an Intergraph Video Analyst filter. Previous research has explored the powerful filters available in Intergraph Video Analyst. In particular, the ICentroid filter has been identified as a filter, which can be implemented to enable real time object tracking. This paper seeks to utilize the ICentroid filter for the purpose of object tracking from streaming video.

1. Introduction

Real Time Object Tracking Technology has the capability to be used for a plethora of activities. In addition to facilitating military reconnaissance through UAV use, real time object tracking technology has the capability to perform many civilian functions. It is with hope that the real time tracking technology discussed in this paper will expand the effectiveness of this technology for civilian use. For instance, Intergraph Video Analyst is a forensic analyst tool that has been thoroughly used and tested by NASA and the U.S Military. Intergraph Video Analyst contains a wide variety of filters that can be used for several purposes such as video stabilization and color tracking. One of the filters available in the Video Analyst package is the ICentroid Filter. The ICentroid filter has the capability to differentiate an object from the background and focus on this object. The goal of this paper is to explain the procedure for the transfer of this military proven technology for civilian use. For instance, this Real Time Object Tracking technology can be implemented on security cameras which would have the capability to automatically adjust their positioning without the use of a motion sensor. Eliminating the need for a motion detector can make active security cameras more cost effective.

Graph Edit Plus and Intergraph Video Analyst will be used collectively to achieve a software implemented Real Time Object Tracking Solution. Essentially, Graph Edit Plus is a software package that facilitates the creation and modification of Direct Show filter graphs.

Graph Edit Plus facilitates the rapid prototyping of these created filter graphs to both C# and C++ DirectShow applications. This paper will methodically describe the steps taken to build a filter graph which activates a fixed camera and streams live video. Additionally, the implementation of Video Analyst's ICentroid filter will be addressed. Finally, the steps taken to produce the C++ DirectShow application which performs real time object tracking will be discussed.

2. Background

Computer Vision is increasingly becoming important in providing information which can complement the sensors on UAVs. [1] As such there is a need for research in the video processing technology arena. According to Courtney Sharp, "The design on any real-time vision system is a daunting task: It involves a systematic integration of hardware, low level images processing (such as segmentation and feature extraction); multiple view geometry (such as pose and structure estimation) and synthesis of real-time controllers."[1]

There are problems associated with video surveillance from UAVs. This research aims to refine video footage obtained from UAVs. For example, image stabilization can greatly improve object tracking. "The goal of an effective video surveillance system is to detect objects in an area of interest... there are several issues inherent to the problem, such as rapidly changing lighting conditions (e.g. due to cloud cover), shadow, occlusion, and entry/exit of objects." [2]

This, research will seek to enable more effective use of Video footage from UAVs. Also, the aforesaid article points out the need for target location and extraction of feature points. UAV cameras need to be able to perform tasks other than landing. In order to meet the diversity of needs desired of video from UAV cameras, this research seeks to utilize a software package that can provide a wide variety of features. Therefore, the present study explores the features of Intergraph's Video Analyst in order to achieve object tracking in video footage acquired from UAVs.

This research will utilize the capabilities of Video Analyst to perform object tracking. Video Analyst is a software package that contains a variety of filters which can be utilized to assist in object tracking. Among these filters are the VARegister filter and the Color filter. For instance, the color tracker filter enables the user to select an area of interest. After this color had been selected on the particular frame of the video, the color tracker enables the user to track this particular color from frame to frame.[3] In this paper we will specifically explore the uses of the Video Analyst ICentroid filter and its integration into a real time object tracking system. Centroid tracking is also referred to as gray scale thresholding. With the ICentroid filter, target segmentation is first performed. Target segmentation involves ascertaining which pixels are associated with the target and which are associated with the background. [4] A fundamental idea on which centroid based tracking is based is that the image of interest contains intensity values that are unlikely to be found in the background. Essentially, the ICentroid filter is an algorithm that tracks objects using the centroid of the image. [5] There are other methods for object tracking for UAVs such as the COCOA system for tracking of objects on UAVs. Additional methods of object tracking can allow interested parties to compare and fine tune results

3. Objective

In order to successfully perform real time object tracking, a filter graph that can activate a fixed camera and stream live video from the camera source needs to be developed. Also, the ICentroid filter from Intergraph Video Analyst needs to be implemented and accessed via Graph Edit Plus. After a filter graph has been designed which streams video and applies the ICentroid filter to this video, Graph Edit Plus can be utilized to generate a C++ framework for this filter graph. This research paper will focus on making the ICentroid filter available to Graph Edit Plus. Graph Edit Plus will then be utilized to create a C++ Direct Show application, which facilitates object tracking.

The ICentroid Filter is a filter produced by Intergraph as part of their Video Analyst Software package. The ICentroid filter is an algorithm packaged as a DLL file. Essentially, a DLL file is a dynamically linked library. In order to make the ICentroid filter available for use in Graph Edit Plus the ICentroid DLL must first be registered in the Windows Registry by the regsvr32 Windows system command. The

ICentroid filter is principally an image based centroid algorithm. After determining the target pixels the calculated center of gravity (centroid) of these pixels can be used as the estimated target position.

A fundamental idea on which centroid based tracking is based is that the image of interest contains intensity values that are unlikely to be found in the background. Essentially the ICentroid filter is an algorithm that tracks objects using the centroid of the image. The centroid of a finite set of points can be calculated as the arithmetic mean of each coordinate of the points. In Particular, the intensity, hue, or saturation of an image could be used to calculate the centroid of a video image.

4. Methodology

In order to develop a C++ DirectShow Application, which would enable efficient real time object tracking, it was important to use the appropriate filter. Consultation with the various stakeholders combined with technical advice from Intergraph yielded that the most appropriate filter for this application was the ICentroid filter. Inquiries revealed that it was possible for the ICentroid DLL to be registered and utilized in Graph Edit Plus. The ability to use the ICentroid DLL in Graph Edit Plus is critical since Graph Edit Plus will be used as a tool to develop a system than can stream real time video and supply it to the ICentroid filter. After the ICentroid filter has been made accessible to Graph Edit Plus the next step involves building a filter graph which would stream live video from a video capture source such as would be found on a UAV. For initial purposes of demonstration a webcam was utilized as the video capture source. A comprehensive filter graph, which implements video streaming, applies the ICentroid filter, and produces viewable output, was designed. Graph Edit Plus has the ability to convert Filter Graphs into a C++ or C# framework. This feature of Graph Edit Plus was used to convert the Filter Graph into its C++ equivalent. This C++ Source code was then used as a basis to fully develop the real time object tracking DirectShow application. Microsoft Visual Studio 2008 was used to compile this C++ code and to make the desired modifications.

5. Results and Discussion

Figure 3 illustrates the registration of the ICentroid DLL using the Command Prompt on Windows 7. It is important to run the Command Prompt as an administrator otherwise the registration of the ICentroid DLL will be unsuccessful. The ICentroid DLL was made available on the Research PC when Intergraph Video Analyst Run time version was installed on the Research PC. It is important to know the file location of the ICentroid DLL in order to register the DLL correctly. The Window identified by "A" illustrates the RegSvr32 command being implemented in the Command Window. The Window indicated by "B" illustrates the location of the ICentroid DLL on the Windows 7 Operating System. The Window showcased by "C" indicates the confirmation that the ICentroid DLL was registered successfully.

Figure 5 illustrates the filter graph system being built. The block labeled "A" illustrates the WDM Streaming Capture Filter for a Microsoft webcam. This filter is located under WDM Streaming Capture Devices. It can be observed that there are three terminals labeled "Video Camera Terminal" "Capture" and "Still". For the purposes of this project the "Capture" terminal is employed.

Figure 5(B) illustrates the AVI Decompressor within the Filter Graph System. When the WDM Streaming capture device was connected directly to the Video Renderer (screen output) the video was almost unrecognizable from the camera. The AVI Decompressor filter enables Video Compression Manager (VCM) codecs to join a filter graph. The AVI Decompressor filter is located under DirectShow filters in Graph Edit Plus. Figure 5(C) illustrates the ICentroid Filter within the Filter Graph System. The ICentroid Filter has two terminals labeled "Input" and

"Output". In order to apply the filter, the WDM streaming capture device is connected to the ICentroid "Input" terminal via the AVI decompressor. Figure 5(D) illustrates the Color Space Converter filter within the Filter Graph System. Figure 5 (E) illustrates the Video Renderer Filter. The Video Renderer filter enables the Filter Graph to display the processed video to the user via a window.

Figure 6 illustrates the output of the Filter Graph when it is run within Graph Edit Plus. The Filter Graph Streams video into the Microsoft webcam applies the ICentroid Filter and returns the output to the user. For demonstration purposes an individual's hand is used as a desired object to be tracked. It can be observed that the ICentroid Filter causes the view to be focused on the sole object in the picture, while the remaining background view is almost eliminated. To get the Filter Graph to run, the user needs to select the play button.

Graph Edit Plus facilitates the rapid prototyping of a designed filter graph to a C++ framework. This C++ framework can be used as a basis to develop the desired C++ Direct Show Application. In the flowchart above (Figure 7), the blue shapes illustrate the C++ Code that has been developed. The segments of the flowchart highlighted in red illustrate ongoing work to be performed on the C++ code.

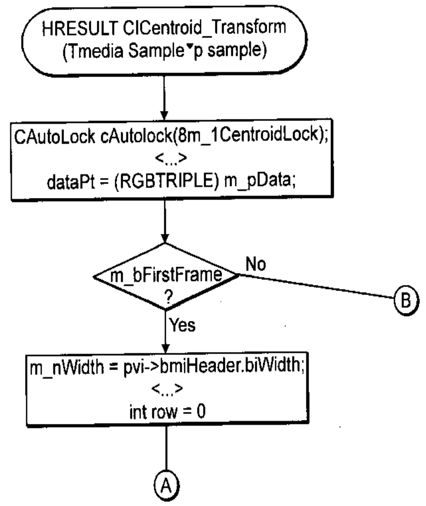


Figure 1. Flowchart illustrating operation of ICentroid filter

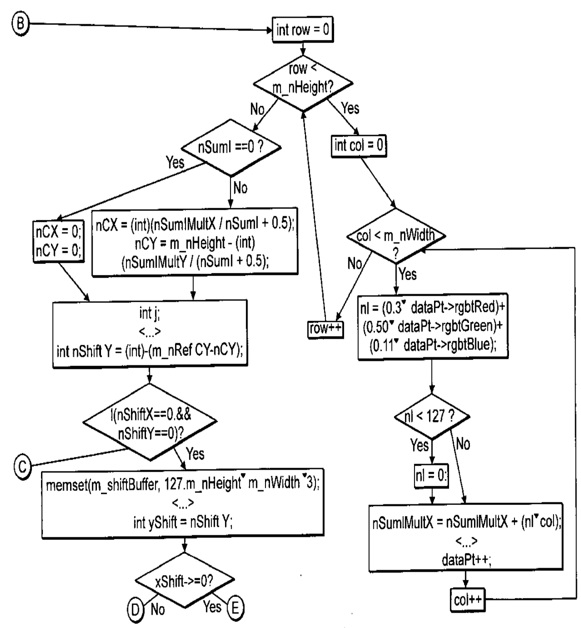


Figure 2. Flowchart illustrating operation of ICentroid Filter

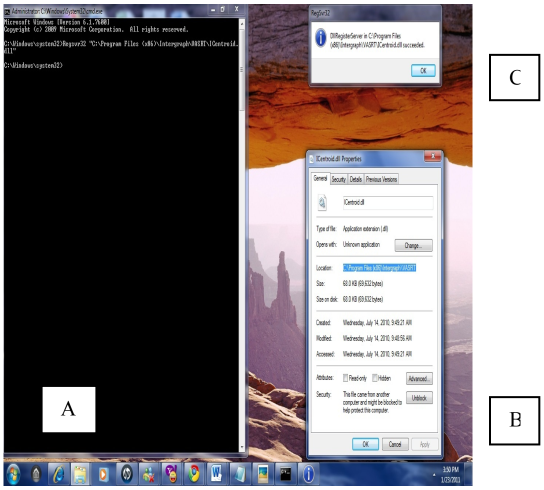


Figure3. Registration of ICentroid DLL

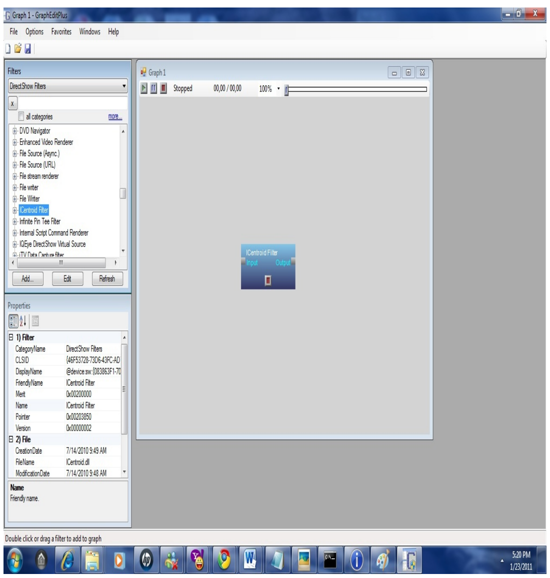


Figure 4. Graph Edit Plus Environment

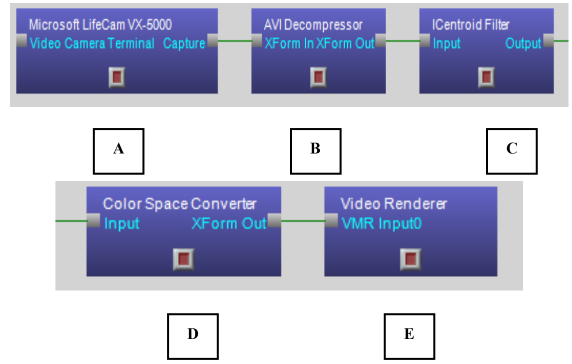


Figure 5. Filter Graph System

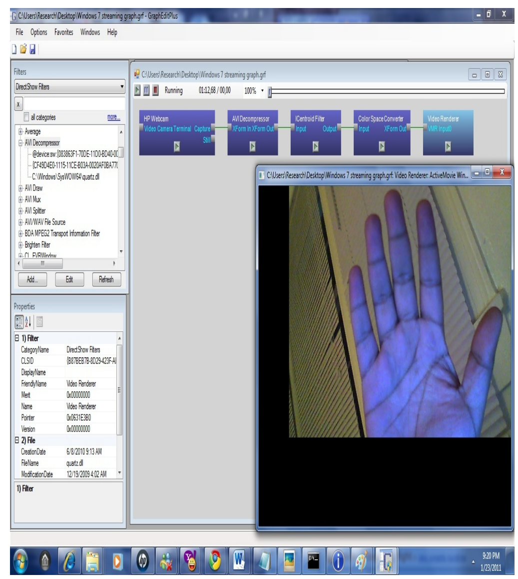


Figure 6. Filter Graph in Use

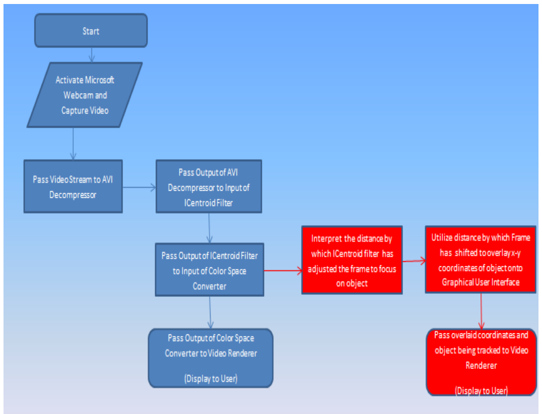


Figure 7. Flowchart for C++ Code (Real Time Obj. Tracking)

6. Conclusion

This study has ascertained that Intergraph Video Analyst's ICentroid Filter can be successfully implemented for Real Time Video Tracking through the use of Graph Edit Plus. Furthermore, it has been shown that Graph Edit Plus can generate the equivalent C++ code of a designated Filter Graph. The ability to generate C++ code is fundamental towards developing a C++ DirectShow application, which facilitates Real Time Object Tracking. This study carefully examined the procedure for building a Filter Graph to stream video and included the implementation of the ICentroid Filter. Furthermore, the conversion of the Filter Graph to C++ code was discussed and the C++ code presented.

7. Ongoing Work

Ongoing work that needs to be performed involves the development of the C++ code to enable the coordinates of the object to be displayed to the user as the object moves relative to the background. In order to accomplish this task, the standard Video Renderer currently used in the filter graph need to be replaced by a Video Renderer that has been purpose built to produce the coordinates. Since the ICentroid filter produces a black border around the object as it moves relative to the background, the extent of this black colored area can be used to ascertain the extent to which the object has moved relative to its original position. For instance the further the object moves out of view of the fixed camera, the longer the black shaded area become on the side where the object was originally located.

Ongoing work includes the construction of a custom built video renderer filter which will output the frame count along with the RGB value of the first pixel of the video in a black bar that runs across the top of the video. Continued development of this filter will seek to achieve the display of coordinates to the user by examination of the dimensions of the black bars created by the ICentroid filter. When this new Video Renderer filter has been developed in its entirety, it will be registered as an .ax file. This filter will then be available for use in GraphEdit Plus. A filter graph can then be built to incorporate this newly developed Video Renderer filter and the corresponding C++ framework generated.

8. Acknowledgements

The authors would like to extend their thanks to Almighty God for the patience and wisdom to complete this project. Furthermore, the authors give their sincere appreciation to Mr. Gene Grindstaff of the Intergraph Corporation for his technical assistance. The authors would also like to thank Mr. Gilbert Haynes, of Morgan State University (MSU), for his patience and understanding. The authors would also like to extend thanks to the Chesapeake Information Based Aeronautics Consortium and the Engineering Visualization Research Laboratory at MSU for the opportunity to pursue this project.

9. References

[1] C.Sharp, O.Shakernia, S. Sastry, "A Vision System for Landing an Unmanned Aerial Vehicle," Proceedings of the 2001 IEEE International Conference on Robotics and Automation, Seoul, Korea, May 21-26, 2001.

[2] M.Shah, A. Hakeem, A.Basharat "Detection and tracking of objects from Multiple Airborne Cameras."[Online].Available: http://spie.org/documents/Newsroom/Imported/155/2006030155.pdf [Accessed June 17th 2010]

[3] Intergraph, Intergraph Video Analyst Software for Adobe Premiere Pro User Guide, Huntsville, Alabama.2008.

[4] Albus, John E. Lewins, Lloyd J. and Schacht, Julie R. "Centroid tracking using a probability map for target segmentation." Proc. SPIE 4714, 175 (2002).

[5] Grindstaff, Gene A. Whitaker, Sheila G. "Real Time Image Stabilization." US Patent Application Publication, filed Aug. 22nd 2005.

FPGA-Based Multi-Agent System for Network Security

Jean Gourd[1], Nathan Killeen[1], Devin Stonecypher[1], Nathan Lapp[1], Maxine Sop[1] and Jeremy Kackley[2]
[1]*College of Engineering and Science, Louisiana Tech University, Ruston, LA 71272, USA*
[2] *Noetic Strategies, Inc., Huntsville, AL 35814, USA*
jgourd@latech.edu, ndk004@latech.edu, dps013@latech.edu, nvl002@latech.edu, mts018@latech.edu, jeremy.kackley@noeticstrategies.com

Abstract

The security of cyber systems is crucial and becoming an increasingly difficult problem, particularly with the growing trend and dynamic characteristics of cyber attacks. Being one step behind and merely reacting to the actions of attackers is commonplace. In this paper, we present an intelligent agent framework that is built on an FPGA-based architecture. Often we do not have the capability to alter the hardware devices such as routers and switches that reside within our network. To address this issue, we propose to develop these devices (and others) with reconfigurable and reprogrammable FPGAs (Field Programmable Gate Arrays). We then introduce a multi-agent system (MAS) that will ultimately be integrated with these FPGA-based devices and discuss how it can be used to provide intrusion detection and protection capabilities. A final discussion of related work that investigates the evolution of individual agents within the framework is also presented.

1. Introduction

2011 will see an increase in cyber war, malware creation and malicious code with the ability to counter detection mechanisms [10]. Cyber attacks have enormous negative impact on businesses, as they have to devote more resources to fight against them. Moreover, cyber criminals are stealing millions of dollars from businesses, and the intellectual property of leading corporations and governments' classified documents are continuously at risk.

Computer systems are invariably susceptible to attacks within the network. Attackers typically target data and/or computational power, both of which are located on the end nodes in the network. The plethora of defensive tools such as firewalls and anti-virus are software-based. The issue with software-based defenses is that by the time they can react to something malicious, the attack has already reached the system. In many cases, malicious software has already entrenched itself before software meant to combat attacks can even detect it.

One might suggest to look at the network hardware such as routers and switches and implement hardware-based defenses. Unfortunately the majority of reliable networking hardware is proprietary and "closed source;" thus these devices do not provide the typical user or system administrator with a useful interface. Vulnerabilities that exist on such devices can often only be mitigated through the manufacturer.

In this paper, we propose an FPGA-based network infrastructure, on which a MAS is layered for the purpose of intrusion detection and protection. The goal is to provide configurable network security at the network hardware level in order to mitigate cyber attacks before they reach the individual nodes.

2. Framework Infrastructure

The decision to utilize FPGAs as the basis for integrating IDS (Intrusion Detection System) and IPS (Intrusion Protection System) algorithms in various network devices is due to their inert reprogrammability and real-time reconfigurability. Moreover, FPGAs have become more prevalent of late due to significant breakthroughs in speed and power. Characterized by short time-to-market and low cost, FPGAs are ideal for controlling logic in embedded systems. There is also an ongoing effort to develop sound techniques that seek to thwart reverse engineers of hardware, including in FPGAs (see [5, 11, 13]).

Our proposed framework implements security mechanisms at two levels: (1) because we are designing the hardware devices for our network using FPGAs, we can incorporate security mechanisms such as intrusion detection and protection algorithms at the hardware level during the design process; and (2) we can provide security at the gateway level and throughout the network using intelligent agents. This is a software approach.

The hardware infrastructure supporting our research consists of eight rackmount servers, three freestanding servers, two switches, and two Cube PCs [1]. Each of the rackmount servers as well as the two PCs and one of the freestanding servers are equipped with NetFPGAs [9] on which our research is based. An overview of the framework infrastructure is shown in Figure 1.

Using the NetFPGAs as NICs (Network Interface Cards), each of the rackmount servers are part of a secure network that is only reachable through the head node. This network acts as the testing ground where our agents will ultimately be run through evolutionary cycles.

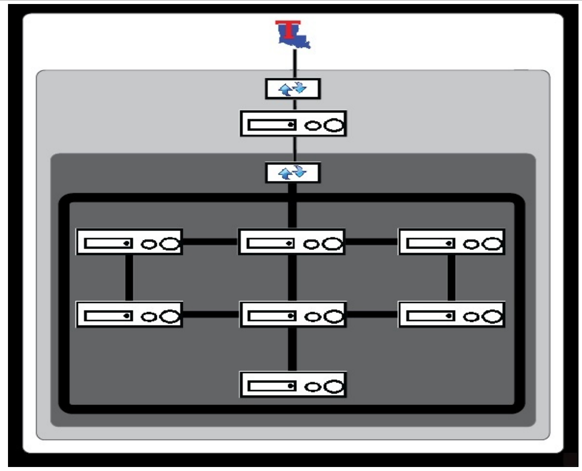


Figure 1: Framework infrastructure

2.1. FPGAs

FPGAs are an integrated circuit designed to allow configuration *after* production. FPGAs are also more economical than ASICs (Application-Specific Integrated Circuits) because they can be reprogrammed to fix bugs and typically result in shorter time-to-market.

FPGAs have recently benefited from advances in their design. As pseudo blank slates, they can be dynamically programmed to perform virtually any function. They have been used as co-processors to offload computationally intensive algorithms [6, 7], as digital signal processors [2, 8], and are being investigated for use as isolation environments for the analysis of code (including binary executables) for potential vulnerabilities. As opposed to the compilation of computer code to executables, FPGA *code* is synthesized onto the FPGA; the result is an encapsulated *chip,* similar to an ASIC, that is dynamically reconfigurable and reprogrammable. Even a remote FPGA-based sensor network may be dynamically reprogrammed by way of embedded intelligent agents or other mechanism upon processing of sensory inputs.

Depending on the development platform, the capabilities of the circuit can differ. Currently, our research utilizes several FPGA development platforms: (1) the Digilent Nexys2 FPGA [3] is a fairly basic development board consisting of a few inputs and outputs that provide testing capabilities during development; (2) the Digilent Spartan 3E-1600 [4] is a more advanced circuit and can take input from an Ethernet cable; (3) the Virtex-5 OpenSPARC platform [12] has high-performance logic fabric and many configuration options; and (4) the NetFPGA is designed specifically for network-based application and uses rapid prototyping of network devices. They are integral in our framework infrastructure, and we will be using them as substitutes for network devices that will allow the agents running on them to interact with data on the network. Combined, these development platforms provide fast design, development and testing capabilities.

2.2. FPGA-Based Network Devices

For production purposes, we will be implementing our MAS agencies on top of NetFPGA-based NICs, routers, switches, and even splices of Ethernet cable; each will function identical to their standard industrial counterparts. These designs may then be used to replace existing network devices to provide support for agent migration and evolution. In the future, integration with legacy systems is supported by simply inserting FPGA-based "cabling" between select nodes of the network.

Integrating agents and other cyber defense algorithms into existing networking hardware is difficult and typically not supported. It is often not possible to alter these devices to suit our research needs. As a result, these algorithms must typically be implemented in software on the individual computing nodes. Unfortunately, this severely restricts implementation. The development of reprogrammable FPGA-based network devices allows the embedding of novel algorithms directly onto the hardware-based network devices, thus providing access to raw packets at a lower level and providing cyber defense capabilities at the network device, prior to potential attack impact at the individual computing node.

3. Intelligent Agent Framework

The intelligent agent framework layered onto the FPGA-based network devices provides a unique capability to continually adapt to changes in the network in real-time at the hardware level. This framework is actually a multi-agent system (MAS) wherein numerous intelligent agents function throughout the system.

These agents operate throughout the LAN (local area network), but focus on perimeter and entry/exit points to the WAN (wide area network). Collectively, they form an adaptive and dynamic IDS and discover the network topology by migrating from node-to-node in the network. Some agents are motivated to migrate to nodes in the network that generate a lot of traffic; others are more interested in nodes that generate no traffic at all. Detection of potential attacks is propagated throughout the LAN to provide overall situational awareness.

The proposed intelligent agent framework is composed of two main components: (1) command, control, and communication (C3) centers; and (2) intelligent mobile agents for use as cyber detectors, defenders, and messengers. A phased C3 center is utilized to establish overall control, to organize various different classes of mobile agents and to provide secure inter-agent communications. It is partly through the C3 center that mobile agents communicate with one another and establish trust-based relationships. The framework exhibits a distributed C3 architecture much like a cellular phone system operates today. In essence, an interface to the system exists at one location that represents the primary C3 point. Numerous secondary C3 centers provide overall command, control, and communication within the network. This distribution assists in providing a coordinated system response to a potential cyber attack.

The framework's intrusion detection and protection capabilities come from a unique collection of three intelligent agent types: (1) detectors; (2) defenders; and (3) messengers. Although all classes share a similar design, each provides unique capabilities within the framework. Detectors perform the crucial task of detecting cyber attacks using various algorithms. Support for the integration of novel intrusion detection algorithms is provided in the design. Defenders act upon detected threats and attempt to mitigate them. Similar to the detectors, defenders may implement a wide variety of algorithms. Messengers distribute knowledge of detected threats throughout the network in order to alert every node in the network of an ongoing or imminent attack. They provide overall situational awareness in order to assist in real-time mitigation of attacks at any point in the network.

The combination of C3 with the three classes of mobile agents works to provide extensive measures of defense against cyber threats. Furthermore, the framework is modular in design in order to provide extensibility, particularly with the integration of new intrusion detection and protection algorithms as they are designed. In this way, agent detectors and defenders can employ the latest in cyber defense algorithms to counter ever evolving cyber attacks. The distributed nature of the framework additionally ensures that there is no single point failure on the network.

3.1. Mobile Agents

Intelligent agents have been widely used as autonomous entities that perform work on behalf of human users [14]. Consisting of encapsulated code, these software agents possess a computational intelligence that imparts them with characteristics similar to humans. Various techniques within the domain of soft computing are typically employed to allow agents to capably make decisions without the assistance of human users.

Mobile agents are a subset of intelligent agents that further possess the ability to wholly and autonomously migrate from node-to-node in the network. Such a capability introduces a variety of security issues, notwithstanding the fact that the administrative boundary of involved systems is crossed. Restricting mobile agents to FPGA-based network devices provides an intrinsic isolation environment and precludes migration to individual computing nodes, thus removing the security concern on these end-user nodes.

The framework's mobile agents build upon the typical definition of an intelligent agent that is composed of two parts: (1) a management component that is common for all agents; and (2) an action component that is unique to each agent. The management component provides typical agent functionality and characteristics such as the cloning of agents and mobile agent migration from node to node in the network. These tasks are similar for all classes of agents and all instantiations of a particular agent class. The action component provides the agent with the mechanisms necessary to perform its required tasks. This component may be unique for each agent class and even for different instances of a particular agent class (i.e., each defender agent may employ different algorithms).

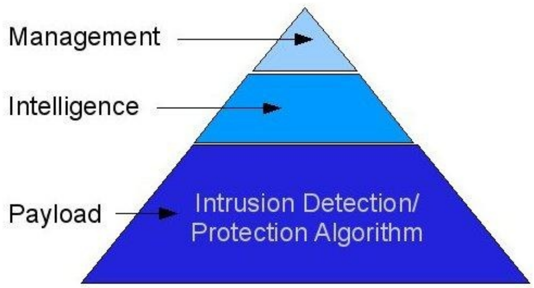


Figure 2: Agent design overview

We extend this definition by breaking down the action component into an intelligence component and a payload component as shown in Figure 2. In essence, the intelligence component serves to manage the payload; therefore it is simply a further breakdown of the agent at one more level. The intelligence of the agent can be implemented in a variety of ways and depends on the payload. For example, the use of fuzzy logic may be used to provide answers to problems that contain incomplete information, and neural networks may be used to assist in matching cyber threats to a known assortment of predefined vulnerabilities. The payload is simply a set of specific actions and algorithms that the agent performs. In the case of our proposed framework, a variety of intrusion detection and protection algorithms can be injected as the agent's payload. We envision dynamic real-time injection of these algorithms.

3.2. Command and Control

A distributed C3 center similar to Figure 3 will provide a multilevel structure to command, control, and communications. Our framework implements a design much like a cellular phone system that contains various levels of C3: a primary C3 center at a single location, a larger number of less significant C3 centers functioning as cellular phone towers, and a much larger array of very thin C3 functionality as cellular phones. The primary C3 center is located on a designated node that can further provide a user interface to the system. A variety of less saturated C3 centers are located at various other strategic nodes in the network. And similarly, thin C3 capabilities are located on the remaining nodes, all in an effort to establish C3 for the mobile agents in order to provide a thorough and coordinated response to cyber attacks. An interface to the system is provided by *plugging in* a user interface module to any C3 center in the network.

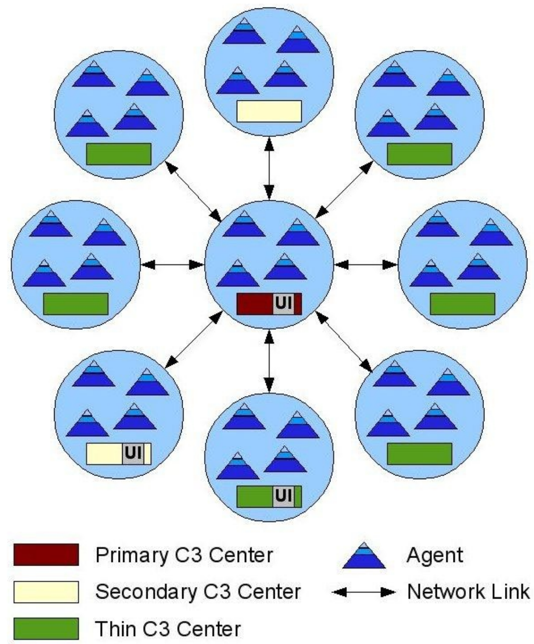


Figure 3: Agent framework C3 distribution

An important feature of the framework is its ability to cross various networked systems and topologies effectively as shown in Figure 4. In today's prevalent wireless world, many systems interact in order to provide much needed services for the public. If an attack occurs on one network, it has the potential to propagate to various other networks. If a network connection exists between two heterogeneous networks, messenger agents can communicate this threat across networks. For example, the occurrence of a threat on a LAN comprised of numerous desktop machines and routers can be propagated

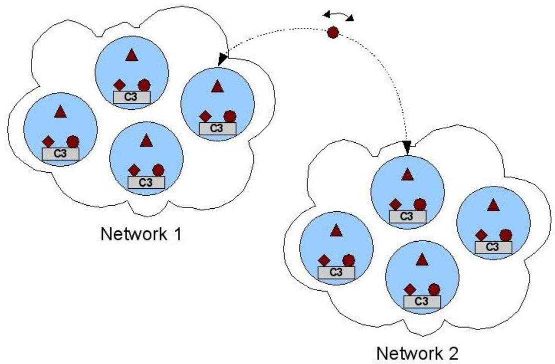


Figure 4: Framework support over heterogeneous networks

to a wireless sensor network that is connected (or reachable) in some manner.

4. Agent Evolution

The latter (and perhaps more theoretical) portion of our research aims to evolve agents into so-called experts with the ability to dynamically adapt to changes in the network. We have begun basic work to define what evolution means and how it might work in such a distributed environment.

Each agent's behavior is represented as a byte array, referred to as their genetic makeup. Additionally, agents are evaluated by their agency and assigned a point value describing how efficient and accurate each is at completing its tasks. Every generation, the most successful agents on a node get the opportunity to reproduce with other successful agents. The offspring then get portions of each parent's genetics with a chance of mutation.

Once the framework is in place to allow such reproduction and evaluation, population of random base agents throughout the network occurs, and evolution continues in order to allow them to collectively evolve to a satisfactory level of performance. After this point of evolution is achieved, a base agent may be re-specified as the minimum threshold for agent performance; continuation of both running the agents and evolving new generations of them concurrently occurs.

In classic genetic algorithms, fitness is evaluated in a centralized manner, mathematically determining both an ideal score and an individual score for each agent, and ranking the individual solutions according to this metric. This approach requires perfect knowledge of the system state which is impractical when evolving detectors and defenders in real-time or in a live system. Generalizations are possible, such as more detections or successful defenses, but might not be absolutely optimal due to unknown algorithmic failures. Having this sort of global ranking defeats the purpose of decentralization and also hampers scalability. For this reason it is better to have fitness be a *local* evaluation of agents perceived to be the best at their activity (be it offense or defense) locally. While perfect local knowledge is lacking, such as knowledge of successful, undetected attacks by opponents, it is not nearly as fragmented as an aggregated central view of the network would be. While there is still danger of evolving agents with flawed detection mechanisms on a local basis, these flaws are unlikely to be globally persistent due to differing evolutionary paths on adjacent nodes.

Evolution can be decentralized in several ways depending upon the desired granularity with implementations ranging from independent populations that occasionally exchange members to a fully distributed peer-based system with genetic exchanges propagating across the peer connections. Given the nature of network-based agents, a trend towards the latter seems more appropriate, treating each network node as essentially its own evolutionary environment, but allowing the free movement of agents between nodes.

By having each individual network node evolve agents independently of each other, we achieve an infinitely scalable evolution environment that does not saturate any centralized evolution node with new data. The cost is a delay between each node having the universal best agent, but means more branches of evolution are explored before being discarded completely. By having more open branches, we are better able to overcome local optima by evolving from different ancestors than a traditional centralized evolution scheme allows. Meanwhile, because highly successful agents will be more prolific, optimum agents will outspread and overpower their inferior counterparts, forcing evolution to continue forward across the entire system.

Successful systems will require several baseline characteristics. First, they need to be able to propagate agents successfully across all connected systems. Without this ability, even the most successful payloads cannot spread across the system, and they will not be able to even achieve a base population. Next, new parameters need to be spreadable; this way, when new threats are identified, proper fitness parameters can be utilized distributively without each node needing to make a discovery on its own. Finally, because of the large numbers of agents available on each node, and because successful agents will generally remain on a node to continue working, the mobility properties of the agents should not be stored on a per-agent basis.

To these ends, the instantiation of a single universal mobile agent should be performed by each agency, whose job it is to collect a copy of a successful local agent and transport it to another node that has less successful agents. These carrier agents, or NodeRunners, then become responsible for fulfilling all communications between the agencies. a NodeRunner knows point ranges for its home's agents and should be able to communicate with other NodeRunners in order to gather information about where its home should be sending new NodeRunners.

5. Conclusions and Future Work

We are currently developing the FPGA-based network devices for our infrastructure. Concurrently, we are developing the agent framework for operation on individual computing nodes. In the near future, our goal is to port the agent framework to the FPGA-based network devices. Moreover, we are laying out the foundation for agent evolution in our framework and plan to begin its implementation within one year.

Additionally, a few thoughts come to mind as we consider the limitations of our approach and its integration into existing systems. As an automation, the system will be limited to work on attacks that the design (i.e., the fitness function) encompasses. A well designed fitness function should allow for the detection of most varieties of attacks. Attacks for which there cannot be a rule for detection, including attacks that occur though encrypted channels, will not be detected.

The biggest practical implementation problem we foresee is that in order to maintain peak performance while continuing to evolve, the platform of execution must always be an FPGA. As the present platform's limitations are reached, we will need to port the entire project to a new platform. This implies additional costs in rewriting the self-modification portion of circuitry. While integrated circuits are common and cheap, FPGAs are still relatively rare technology; they will stay more expensive until economies of scale allow them to become cheaper. Furthermore, FPGAs have limited physical memory, and this will be a bottleneck when looking at expansion of the framework.

As we continue our research in these areas, there are many valid ideas that come to mind, some of which seem more reasoned than others. One that stands out involves facilitation of the framework's constant operation. To accomplish this, we envision the FPGAs to contain three varieties of virtual operating sections: hard, soft and volatile. The Hard Operating Section (HOS) contains both the core device functionality the FPGA replaces and the agency's available operations. The HOS only ever changes when new functionality is distributed to the agencies. The Soft Operating Section (SOS) contains all active agents within a node. The SOS is frequently duplicated to spawn mobile agents, is expandable as additional varieties of agents become active and provides a somewhat stable platform from which agents may execute agency operations. The HOS will assume traffic is valid and will process it as current hardware would, unless an agent in the SOS interrupts it. The Volatile Operating Section (VOS) is where evolution occurs. This rapidly mutating section of the FPGA provides an area for new generations of agents to be born and evaluated without giving them access back to the agency's operational commands. Agents in the VOS are isolated from the HOS until they have been verified as successful and duplicated into the SOS, where they may replace a predecessor. Each node will contain one HOS and one general SOS used to store imported agents that do not match a local specialization. Additionally they will contain a SOS/VOS pair per variety of agent specialization.

References

[1] Accent Technology. 2011. NetFPGA cube. Retrieved on 2011-03-22 from http://www.accenttechnologyinc.com/netfpga.php.

[2] Baronti, F., Lazzeri, A., Roncella, R., and Saletti, R. 2009. FPGA/DSP-based implementation of a high-performance multi-channel counter. *J. Syst. Archit.* 55, 5-6 (May. 2009), 310-316.

[3] Digilent Nexys2 FPGA Board. 2011. Digilent. Retrieved on 2011-03-22 from http://www.digilentinc.com.

[4] Digilent Spartan 3E-1600 Development Board. 2011. Digilent. Retrieved on 2011-03-22 from http://www.digilentinc.com.

[5] Eisenbarth, T., Guneysu, T., Paar, C., Sadeghi, A-R., Schellekens, D., and Wolf, M. Reconfigurable trusted computing in hardware. In *STC '07: Proceedings of the 2007 ACM Workshop on Scalable Trusted Computing,* 15-20, Alexandria, VA.

[6] Fatemi, M. R., Jebril, I., and Salleh, R. 2008. An FPGA based co-processor for elliptic curve cryptography. In Proceedings of the Fifth IASTED international Conference on Communication Systems and Networks (Palma de Mallorca, Spain, September 01 - 03, 2008). C. E. Salvador, Ed. International Association Of Science And Technology For Development. ACTA Press, Anaheim, CA, 73-77.

[7] He, H. and Guo, H. 2008. The Realization of FFT algorithm based on FPGA co-processor. In *Proceedings of the 2008 Second international Symposium on intelligent information Technology Application - Volume 03* (December 20 - 22, 2008). IITA. IEEE Computer Society, Washington, DC, 239-243.

[8] Kwan, A. K., Helaoui, M., Boumaiza, S., Smith, M. R., and Ghannouchi, F. M. 2009. Wireless communications transmitter performance enhancement using advanced

signal processing algorithms running in a hybrid DSP/FPGA platform. *J. Signal Process. Syst.* 56, 2-3 (Sep. 2009), 187-198.

[9] NetFPGA. 2011. NetFPGA. Retrieved on 2011-03-22 from http://www.netfpga.org.

[10] PandaLabs. 2010. Annual report. Retrieved on 2011-03-22 from http://press.pandasecurity.com/wp-content/uploads/2010/05/PandaLabs-Annual-Report-2010.pdf.

[11] Trimberger, S. Trusted design in fpgas. In *DAC '07: Proceedings of the 44th Annual Conference on Design Automation,* 5-8, Dan Diego, CA.

[12] Virtex-5 OpenSPARC Evaluation Platform. 2011. Digilent. Retrieved on 2011-03-22 from http://www.digilentinc.com.

[13] Wollinger, T., Guajardo, J., and Paar, C. Security on fpgas: state-of-the-art implementations and attacks. *Transactions on Embedded Computing Systems,* 3:3, 534-574, 2004.

[14] Wooldridge, M. and Jennings, N. R. Intelligent agents: Theory and practice. Knowledge Engineering Review, 10(2):115–152, 1995.

Stochastic Analysis of Pull Production Systems

Faisal Aqlan[1], Mohammad M. Hamasha[1], Osama T. Almeanazel[1], *Mohammad T. Khasawneh[1], and Tarek Al-Hawari[2]
[1]*Department of Systems Science and Industrial Engineering, State University of New York at Binghamton, Binghamton, NY 13902, USA*
[2]*Industrial Engineering Department, Jordan University of Science and Technology, Irbid 22110, Jordan*
*Corresponding author: mkhasawn@binghamton.edu

Abstract

In this study, multi-stage pull (kanban) production systems are analyzed using stochastic modeling. Analytical and simulation models are built for systems that are composed of three stations and produce a single product type. Analytical models are based on a parametric decomposition technique and two-moment approximations. Numerical experiments are conducted to check models validity and to study the effect of different parameters on system performance. Results obtained from analytical models are compared with results obtained from simulation models. Accurate estimates of system performance measures, such as throughput and work-in-process inventory, are also discussed.

1. Introduction

Performance analysis of production systems is intended to improve existing systems or design new ones. The most common method used for this purpose is modeling, either analytical or simulation. Kanban controlled (pull) and CONWIP controlled (hybrid) production systems have received much attention in recent years. In a pull system, production is driven by the flow of actual demands. The flow of jobs at various stations of the manufacturing system is regulated using a combination of card loops and authorization signals. Each card loop has a finite number of cards, which restricts the total number of jobs in the system at any given time. Jobs move from one station to another only if cards are available. Information regarding demand is passed upstream (from finished products toward raw materials) one stage at a time as product withdrawals occur. Figure 1-a shows information flow for Kanban or pull system.

To design a kanban system, the following issues should be taken into consideration: a) number of stages to be controlled by a particular kanban loop, and b) number of kanbans in each loop. CONWIP systems, on the other hand, are different in that only one kanban loop controls the product flow as shown in Figure 1-b.

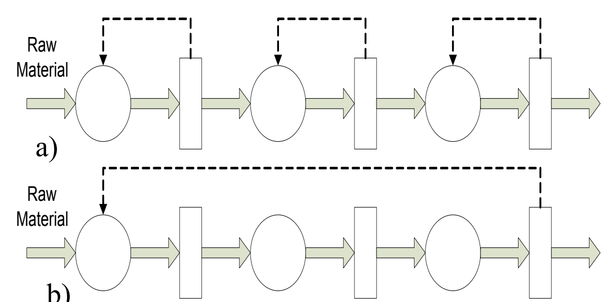


Figure 1. Information flow in Kanban (a) and CONWIP (b) systems.

Queuing models are the most popular analytical methods used to analyze kanban systems. Closed queuing networks with manufacturing stations and fork/join synchronization stations are used to design the system. The closed queuing network helps to explicitly model the work-in-process (WIP) cap, while the fork/join station models the synchronization constraints between the jobs and cards [1]. Figure 2 shows the queuing model for a multi-stage kanban system where J_{i-1} and J_i are the fork/join stations, MP_i

represents the manufacturing stations, P_i is the products queue, and F_i is the kanban queue. The fork/join station J_{i-1} models the synchronization of raw material with free stage *i* kanbans (*K*) while the fork/join station J_i models the synchronization of parts leaving the manufacturing stage MP_i with external customer demand (*D*).

2. Literature review

The literature is abundant with studies that discussed analyzing pull and hybrid production systems using mathematical modeling. A survey on different simulation, deterministic and stochastic models for analyzing pull production systems can be found in [2]. Deterministic models do not capture the impact of uncertainty and variability that exists in practical manufacturing systems. On the other hand, stochastic models take uncertainty and variability into consideration and are mostly based on Markov chains and queuing models. Markov models were built in [3] for a two-stage kanban system producing a single product. Stochastic models were developed for a kanban-based serial production line with constant [4] and stochastic [5] demands. Two-moment approximation models for analyzing multi-stage production systems were discussed in [6]. Performance analysis of single stage kanban controlled production systems using the decomposition technique were developed in [1]. They developed two-moment approximations of the inter-arrival and inter-departure times and then used the decomposition technique to build their model. The current study is based on [1] and [6], wherein the models are modified to account for the assumptions made for the systems discussed in this study.

3. Production system modeling

The system discussed in this study represents a three-stage kanban controlled production system producing a single product type. The three stages are arranged in series and kanban cards or signals are used to control the flow of products among stations. All stages are assumed to have the same number of kanbans (i.e., K_1= K_2= K_3= K_4, where K_1 represents the first stage kanbans, K_2 represents the second stage kanbans, K_3 represents the third stage kanbans, and K_4 is the demand arrivals kanbans).

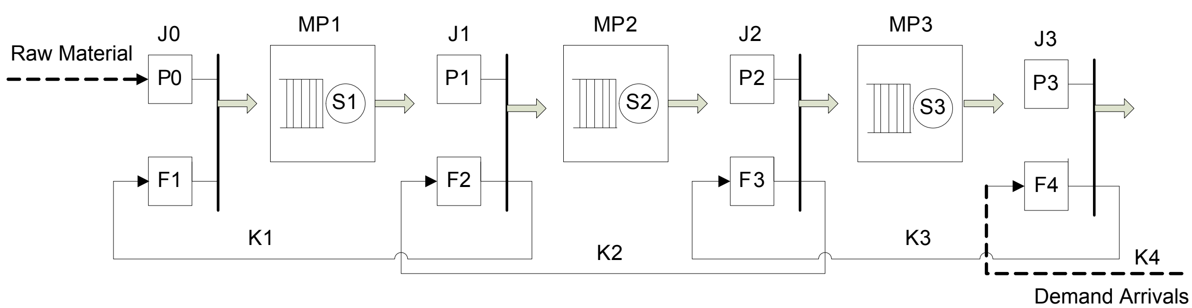


Figure 2. Queuing model of the proposed system

As stated earlier, the system is analyzed using two-moment approximation and node decomposition technique, assuming that the flow of raw materials is a continuous process and the product queue is always full of products. Demand arrivals and processing times are assumed to follow general distributions (renewal process).

3.1. Analytical modeling

As stated earlier, analytical modeling presented in this research is based on the models developed in [1] and [6] with the following modifications. First, this research extends the single-stage models developed in [1] and [6] and applies them to a three-stage system. In [1], the models were applied to a single-stage kanban system and a very briefly discussion on how the model can be used as a building block for multi-stage systems, without any modeling work or numerical experiments, is presented in [6]. Second, two cases of raw material and demand arrivals are assumed in this research, continuous and stochastic, rather than only considering stochastic demand as in [1] and [6]. For the continuous case, the two queues at P_0 and F_4 (Figure 2) will always be full and the models developed in [1] and [6] are modified accordingly. For the stochastic case, demand and raw material arrivals are assumed to be renewal processes and the models developed in [1] and [6] are applied "as is".

The overall approach consists of four main steps [6]: decomposition, characterization, linkage and solution. In the decomposition step, the three stage system is decomposed into three single stage subsystems (SUB_i), as shown in Figure 3.

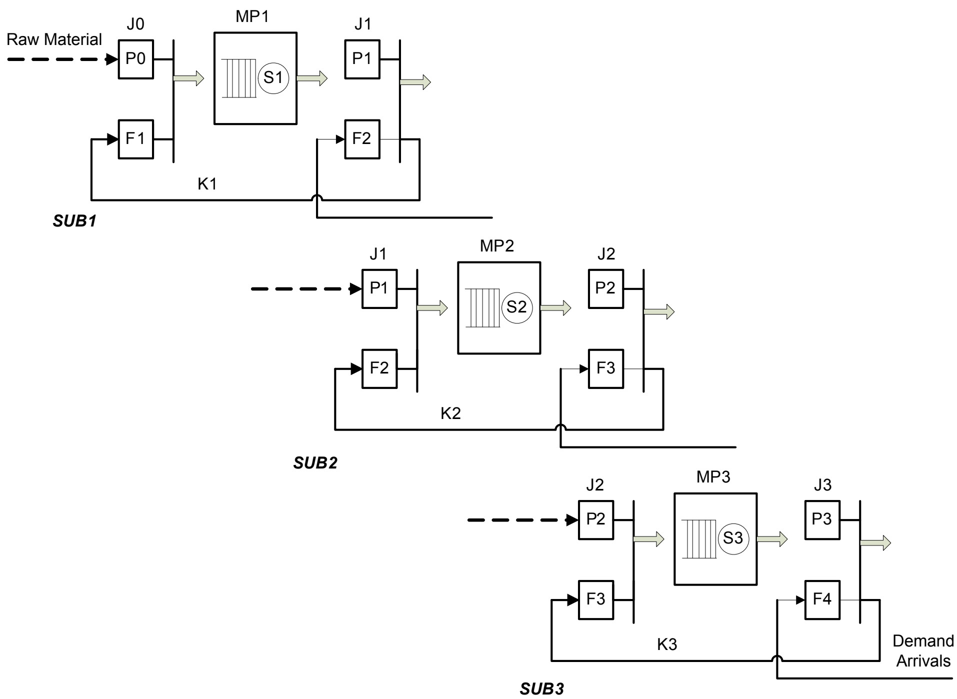


Figure 3. Decomposition of the three-stage kanban system

In the analysis step, each subsystem is modeled individually using a three-step approach: decomposition, characterization, and linking. For SUB_1, the analysis starts by decomposing the system into two fork/join stations (J_0 and J_1) and one manufacturing station (MP_1), as shown in Figure 4. The arrival or departure process from each station is characterized by the two moments: mean arrival rate and squared coefficient of variation (λ_i, C_i^2).

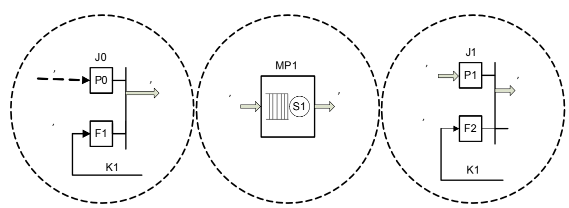


Figure 4. Decomposition of first stage

Each station is then characterized by developing a set of equations linking output to input. For the first fork/join station, assuming that the flow of raw material is continuous, then:

$$\bar{L}_{F1} = 0 \quad (1)$$

$$\lambda_{D0} = \lambda_{F1} \quad (2)$$

$$C^2_{D0} = C^2_{F1} \quad (3)$$

If raw material arrives according to a general distribution characterized by λ_{P0} and C^2_P, which is the case discussed in [1] for a single-stage kanban system, the values of the parameters in Equations 1-3 will be different as will be discussed in the analysis of the second stage. For *MP₁*, the parameter values are as follows:

$$\lambda_{d1} = \lambda_{m1} \quad (4)$$

$$C^2_{d1} = (1-\rho^2_{S1})C^2_{m1} + \rho^2_{S1}C^2_{S1} \quad (5)$$

where $\rho_{S1} = \lambda_{m1}\mu^{-1}_{S1}$

$$L_{S1} = L_{q1} + \rho_{S1} \quad (6)$$

where $L_{q1} = \lambda_{m1} \times Factor\ A$, where Factors A-G can be found in [1], as also shown in the Appendix. For the second fork/join station, J_1, define $R_1 = \lambda_{P1}/\lambda_{F2}$ and $C^2_1 = (C^2_{P1}+C^2_{F2})/2$.

If $R_1 \neq 1$,

$$\lambda_{D1} = \lambda_{P1} \times Factor\ B \quad (7)$$

$$\bar{L}_{P1} = Factor\ C \quad (8)$$

$$\bar{L}_{F2} = Factor\ D \quad (9)$$

If $R_1 = 1$,

$$\lambda_{D1} = \lambda_{P1} \times Factor\ E \quad (10)$$

$$\bar{L}_{P1} = K_1(K_1+1) \times Factor\ F \quad (11)$$

$$\bar{L}_{F2} = K_2(K_2+1) \times Factor\ F \quad (12)$$

$$C^2_{D1} = Factor\ G \quad (13)$$

The departure process from each node is linked to the arrival process of the following node. There are three nodes that require three linkages (Figure 5). For the direct linkage, the departure from node is linked directly to arrivals to next node (i.e., $\lambda_i = \lambda_{i-1}$, $C^2_i = C^2_{i-1}$). For the stochastic transformation:

$$\lambda_{P1} = \left[\frac{\lambda_{d1}}{1-\pi_{P1}}\right] \quad (14)$$

$$C^2_{P1} = \frac{C^2_{d1}}{(1-\pi_{P1})^2} - \left(\frac{\pi_{P1}}{(1-\pi_{P1})^2}\right)\left(\frac{\lambda_{d1}}{\lambda_{F2}}\right)\left(\frac{2C^2_{F2}}{1+C^2_{F2}}\right) \quad (15)$$

where π_{P1} is the long-run proportion of time that the arrivals to P_1 shutdown. The set of nonlinear equations is then solved using an iterative algorithm.

The algorithm is initialized with an estimate of the parameters characterizing the departure process from the first fork/join station (J_0). Therefore, initial values are given to λ_{D0} and C^2_{D0} and then the iterative procedure updates these estimates until they are consistent with the input parameters (μ_{S1}, C^2_{S1}, K_1, λ_{F2}, C^2_{F1}, K_2).

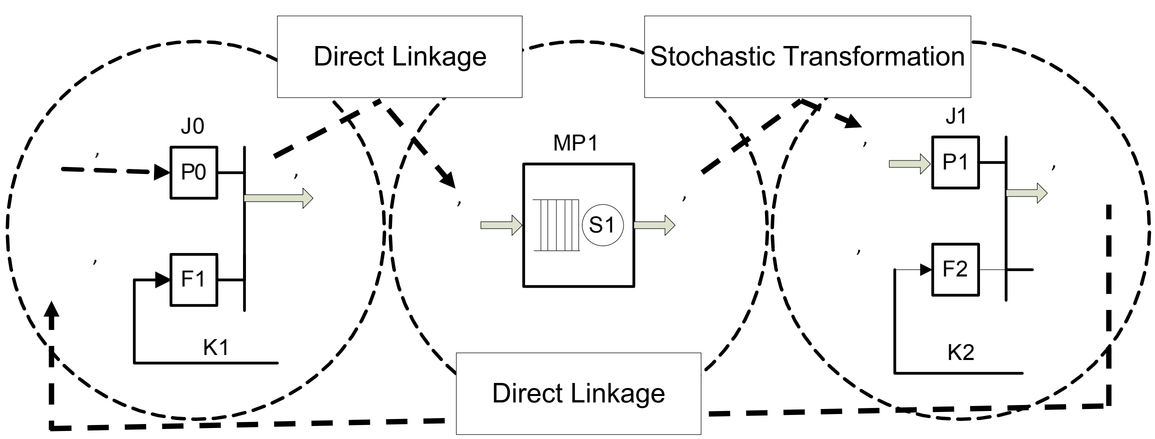


Figure 5. Linkage of first stage nodes

Analysis of the second stage will be similar to that in [1], with the exception of the continuous demand following a renewal arrival process assumption, which does not apply to the second stage. To link the three nodes, there will be two stochastic transformations and one direct linkage (Figure 6). The fork/join station J_1 is shared between the first and second stages and the governing equations will be modified in the second stage considering that the departure from first stage is the arrival to the second stage.

Analysis of the third stage is similar to that of the second stage given that the demand arrival process is renewal. However, if demand arrivals are assumed to be continuous, the analysis will be different in that the stochastic transformation that links the output from J_3 to the input to J_2 will be a direct linkage. Therefore, there will be two direct linkages and one stochastic transformation (Figure 7). The fork/join station J_2 is shared between the first and second stages and the governing equations will be modified accordingly.

The final step is linking, in which single-stage subsystems obtained from decomposition are linked together and solved for the unknown parameters using an iterative algorithm. Performance measures for the entire network are then derived. The linkage of the three stages is shown in Figure 8.

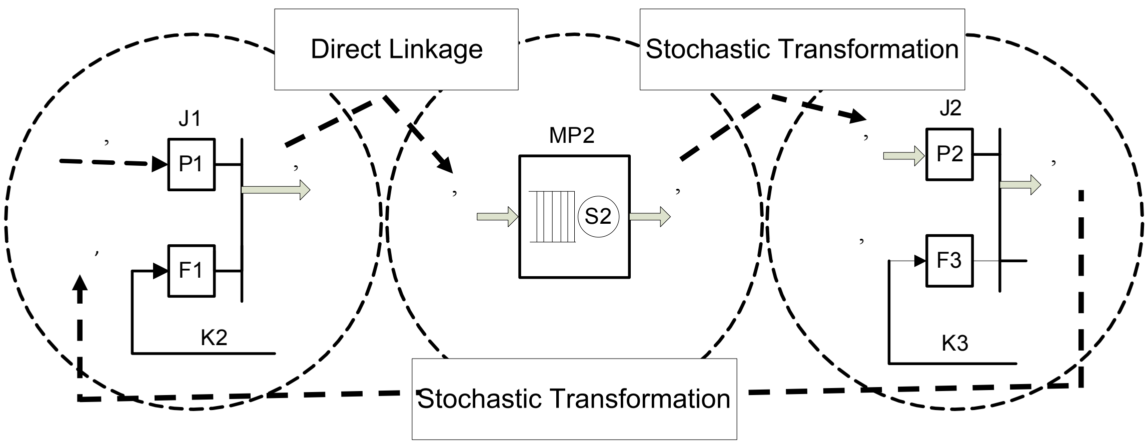


Figure 6. Linkage of second stage nodes

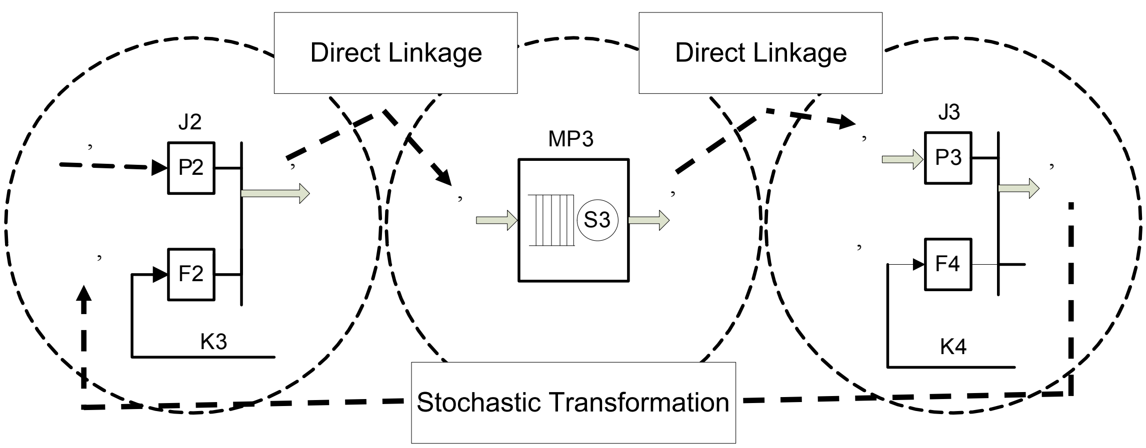


Figure 7. Linkage of nodes in the third stages

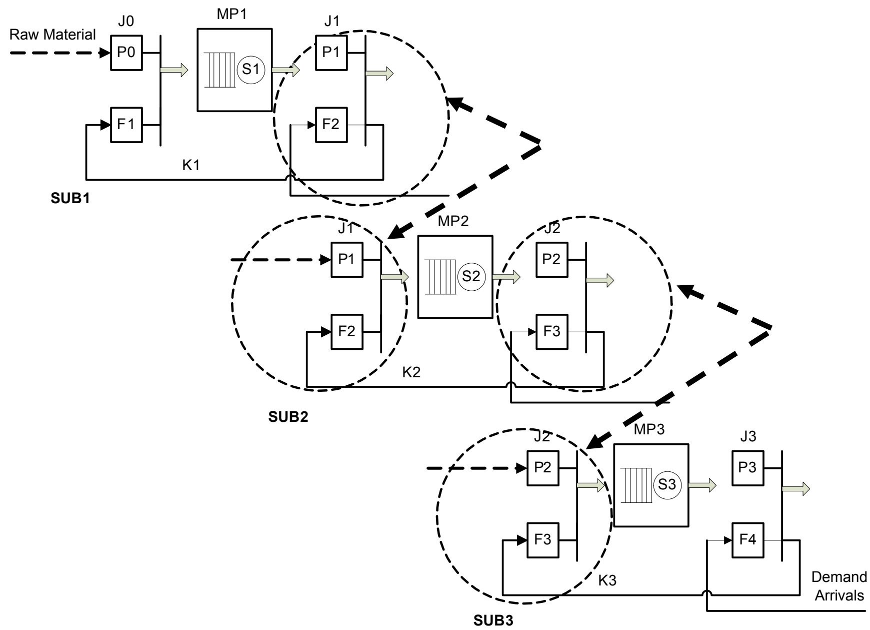


Figure 8. Linkage of the three subsystems

3.2. Simulation modeling

A simulation model is built using the Arena software (Version 13.5). Kanbans are modeled as entities and flow through the system along with the parts (also represented as entities). Parts and kanbans wait in separate queues until a part and a kanban are available. The kanban is then merged with the part and are both transferred to the manufacturing station as one entity. After processing is completed, the entity waits in the output queue for a downstream stage kanban. When it is available, the upstream stage kanban is first detached and returned to the first stage while the downstream kanban is attached to the part and are both transferred to the next manufacturing station as one entity. A comparison of the analytical and simulation approaches is shown in the following section.

4. Results and discussion

Numerical experiments are conducted to assess model validity and study system performance. The iterative algorithm used in the analytical approach is implemented on a Pentium IV PC (CPU: 2.4 GHz, RAM: 2 GB). The associated set of nonlinear equations is solved using MATLAB (version 7). Results from the analytical approach are compared with the simulation results. Both models were run on the same PC. Simulation results represent the average of fifty independent runs, each of which representing the production of about 1,000,000 parts or customer orders. For each run, statistics corresponding to the first 100,000 parts are discarded (warm-up period) to account for transient start-up effects. For each experimental observation, 95% confidence intervals are recorded and used to verify the validity of the insights.

The computational time required by the analytical method to obtain performance estimates is very low (less than 1 minute for $\varepsilon = 0.001$) when compared to that needed for simulation runs. The main performance parameters recorded in all the experiments are: 1) system throughput and 2) WIP. Figure 9 shows the effect of increasing demand variability (0.5, 0.7, 1, 2) on system throughput. It can be seen that increasing demand variability will decrease the system throughput, a less significant effect for higher numbers of kanbans (Figure 10).

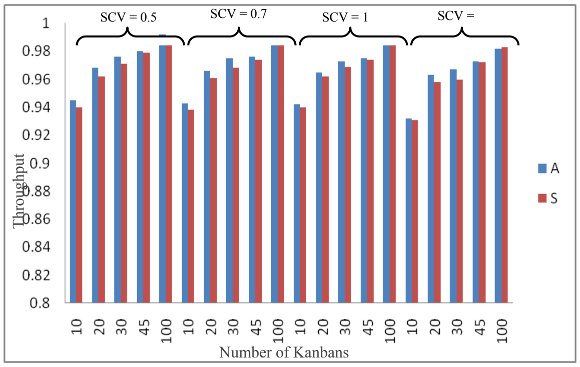


Figure 9. Impact of demand variability on system throughput (K's = 10, 20, 30, 45, 100)

The accuracy of the analytical approach, compared to simulation, is estimated by computing its percentage error, as shown in Table 1. The percentage error for the system throughput (λ_{D3}) is computed as follows:

$$(\lambda_{D3}) = \frac{\lambda_{D3}^{(Analytical)} - \lambda_{D3}^{(Simulation)}}{\lambda_{D3}^{(Simulation)}} \times 100$$

When demand variability is held constant and the number of kanbans increased (the three stages are assumed to have the same number of kanbans), the system throughput will also increase. However, system throughput is limited by the demand arrival rates and the processing rates of the manufacturing stations (Figure 10).

For WIP, the number of kanbans for each station increased while the other stations' kanbans were held constant at 30. In general, increasing the number of kanbans will increase the system's WIP, wherein the maximum impact occure in the case of the demand arrivals kanbans (*K4*), followed by the first stage kanbans (*K1*), as shown in Figure 11.

Table 1. Comparing analytical and simulation results

C_{F4}^2	K's	λ_{D3}		
Input parameters		A	S	E %
0.5	10	0.945	0.94	0.53
0.5	20	0.968	0.962	0.62
0.5	30	0.976	0.971	0.51
0.5	45	0.98	0.979	0.10
0.5	100	0.992	0.989	0.30
0.7	10	0.943	0.938	0.53
0.7	20	0.966	0.961	0.52
0.7	30	0.975	0.968	0.72
0.7	45	0.976	0.974	0.21
0.7	100	0.99	0.988	0.20
1	10	0.942	0.94	0.21
1	20	0.965	0.962	0.31
1	30	0.973	0.969	0.41
1	45	0.975	0.974	0.10
1	100	0.989	0.988	0.10
2	10	0.932	0.931	0.11
2	20	0.963	0.958	0.52
2	30	0.967	0.96	0.73
2	45	0.973	0.972	0.10
2	100	0.982	0.983	0.10

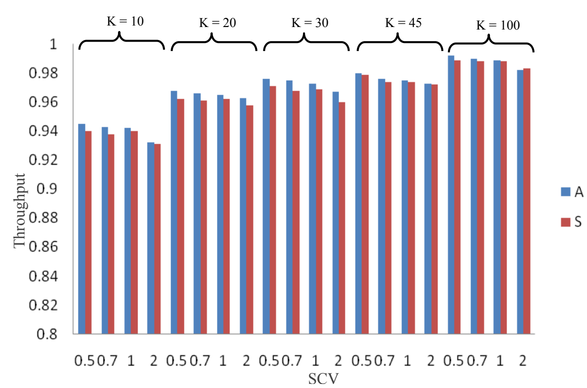


Figure 10. Impact of increasing the number of kanbans on throughput for SCV = 0.5, 0.7, 1, 2 respectively

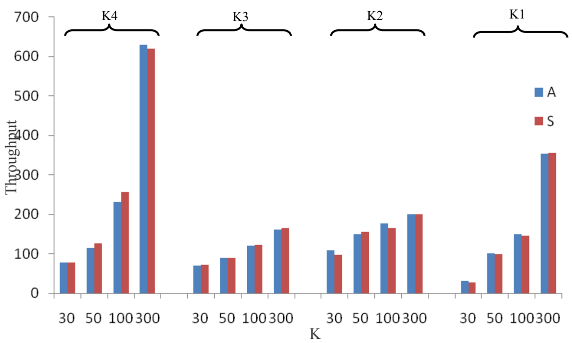


Figure 11. Impact of increasing the number of kanbans on total WIP

5. Conclusions and future work

Analytical and simulation models can be used effectively to analyze pull and hybrid production systems. Results from numerical experiments showed that variability in demand has a direct impact on system performance. Throughput of kanban systems is an increasing function of the number of kanbans but is limited by the arrival rates of demands and the processing rates of the manufacturing stations. For total WIP, increasing the value of customer orders' kanbans has no effect, while increasing the value of other stages' kanbans increases the total WIP. When there is a continuous demand arrival process, in general, increasing the number of kanbans increases the system throughput with limitations due to the service rate of the manufacturing stations. The difference between simulation and analytical indicates that the analytical approach is reasonably accurate.

The stochastic approach used in this study presents several opportunities for extensions and future work, especially to multi-stage systems. Furthermore, the approach can be extended for CONWIP systems, unlike [1] and [6], which only consider a kanban system. Since a CONWIP system can be treated as a single-stage kanban system with multiple manufacturing stations and only two fork/join stations, the models can be applied with slight modification. Additionally, the models can be modified easily to be used for analyzing systems producing several products. The optimization of system parameters, such as the number of kanbans, can be found using Genetic Algorithms or other methods. Finally, although the analytical and simulation models discussed above are used to analyze a three-stage kanban system, they can be extended easily to analyze systems with more than three stages. Also, the following four states of a multi-stage pull production system can be studied using these models:

- Stochastic raw material arrivals and stochastic demand arrivals.
- Continuous raw material arrivals and stochastic demand arrivals.
- Stochastic raw material arrivals and continuous demand arrivals.
- Continuous raw material arrivals and continuous demand arrivals.

6. References

[1] A. Krishnamurthy and R. Suri, "Performance analysis of single stage Kanban controlled production systems using parametric decomposition", Queuing Systems: Theory and Applications, Vol. 54, No. 2, 2006, pp. 141-162.

[2] R. Uzsoy and L.A. Martin-Vega, "Modeling Kanban-based demand-pull systems: a survey and critique", Manufacturing Review, Vol. 3, No. 3, 1990, pp. 155-160.

[3] U. Karmarkar, and S. Kekre, "Batching policies in kanban systems", Journal of Manufacturing Systems, Vol. 8, No. 4. 1989, pp. 317-328

[4] D. Mitra and I. Mitrani, "Analysis of a kanban discipline for cell coordination in production lines", Management Science, Vol. 36, No. 12, 1990, pp. 1548-1566.

[5] D. Mitra and I. Mitrani, "Analysis of a kanban discipline for cell coordination in production lines, II, Stochastic Demands", Operations Research 1991, 39: 807–823.

[6] A. Krishnamurthy, R. Suri, and M. Vernon, "A new approach for analyzing queuing models of material control strategies in manufacturing systems", Proceedings of the Fourth International Workshop on Queuing Networks with Finite Capacity QNETs, West Yorkshire, U.K., 2000.

Appendix [1]

Factor A:

$$\left[\frac{(K_1-1)\rho_{s1}(C_{m1}^2+C_{s1}^2)}{2K_1\mu_{s1}(1-\rho_{s1})+\lambda_{m1}(C_{m1}^2+C_{s1}^2)}\right]$$

Factor B:

$$\left[\frac{1-R_1^{K_1+K_2}}{1-R_1^{K_1+K_2+1}}\right]\left[1-0.5(C_1^2-1)\left(\frac{(1-R_1)R_1^{K_1+K_2+1}}{1-R_1^{2(K_1+K_2)+1}}\right)\right]$$

Factor C:

$$\left[\left(\frac{R_1^{K_2+1}}{1-R_1}\right)\left(\frac{1-R_1^{K_1}}{1-R_1^{K_1+K_2+1}}\right)-\left(\frac{K_1R_1^{K_1+K_2+1}}{1-R_1^{K_1+K_2+1}}\right)\right]\left[1+\left(\frac{1-R_1}{1+R_1}\right)\left(\frac{R_1^{t}}{1+R_1^{t}}\right)(C_1^2-1)\right]$$

Factor D:

$$\left[\left(\frac{K_2}{1-R_1^{K_1+K_2+1}}\right)-\left(\frac{R_1}{1-R_1}\right)\left(\frac{1-R_1^{K_2}}{1-R_1^{K_1+K_2+1}}\right)\right]\left[1+\left(\frac{1-R_1}{1+R_1}\right)\left(\frac{R_1^{t}}{1+R_1^{t}}\right)(C_1^2-1)\right]$$

Factor E:

$$\left[\frac{K_1+K_2}{K_1+K_2+1}\right]\left[1-\frac{0.5(C_1^2-1)}{2(K_1+K_2)+1}\right]$$

Factor F:

$$\frac{1}{2(K_1+K_2+1)}$$

Factor G:

$$\left[\left(\frac{\lambda_{p1}^2}{\lambda_{p1}^2+\lambda_{p1}^2}\right)C_{p2}^2+\left(\frac{\lambda_{p1}^2}{\lambda_{p1}^2+\lambda_{p1}^2}\right)C_{p2}^2\right]\left[1-\frac{1}{K_1+K_2+1}-\frac{1}{(K_1+K_2+1)^2}\right]$$

Secure Mobile Agent System and the Gold Miner Paradigm

James Jacobs
Center for Defense Integrated Data, Jackson State University, Jackson, MS 39204, USA
jjacobs@c-did.com

Abstract

Mobile agents have been a research topic for multiple years; however, their usage has been limited by the security concerns related to their application. These security concerns have lead to other invocations, such as web services and other M2M communications, being utilized for M2M applications that may have been better suited for the mobile agent environment. The purpose of the Gold Miner Paradigm (GMP) is to address the security concerns while maintaining the features that make mobile agents unique and valuable. The paradigm applies security mechanisms analogous to that of a gold mine for each of the key components of the data mining task of the mobile agent and mobile agent system. The resulting Secure Mobile Agent System using the Gold Miner Paradigm (SMASGMP) provides the ability to mitigate the security concerns to a level consistent with those posed by using a web service.

1. Introduction

Mobile agents [1] are coded objects of encapsulated instructions and possibly data with the ability to move from machine to machine with the intent of gathering and processing information and moving to another location related to a task from a dispatcher. Each machine in a mobile agent system (MAS) provides an agency for the mobile agent to facilitate these tasks by providing a place to operate, resources to consume and a network for mobility and communications. The resulting system gives the dispatcher the ability to send a mobile agent on a task and upon completion return with the results.

Web-services [2] are a set of machine to machine (M2M) communication protocols which allow a remote machine to request processing on a local host machine and/or transfer of information back to the remote machine which requested the service. The web-service paradigm requires much of the same mechanisms as mobile agents with the client and server of a web-service requiring common communication and hosting of resources.

The key contrast between mobile agents and web-services lies in functionality defined for the data control, the instruction's origin and the relationship between these two components in acquiring the response. This contrast in functionality gives the mobile agent (client) the added abilities of movement, processing local to the data, and transporting the data when compared to web-services. The added functionality is acquired by relaxing the restrictions related to the control of the data and the instructions being processed by the server hosting the information. The reduction of these constraints increases the security concerns associated with mobile agents in respect to web-services. In some respects the added functionality allows the mobile agent to be compared to a virus with the only definable difference being the intent of the code.

The purpose of the gold miner paradigm is to use the mechanisms and responsibilities of a gold mining operation and apply them to mobile agents. The paradigm will address the threat classifications: agent-to-platform, platform-to-agent, agent-to-agent defined by [3]. In addition, the paradigm protects the data serviced by the platform as it is communicated back to the dispatcher. The end result is a mobile agent system with security on the level of a web-service while allowing the added functionality of the mobile agent.

2. Related works

Security of mobile agents is the diverse topic as security must be provided from various fronts. Security must protect the agent, the platform, the communication and the data. Encryption and reputation systems are two methods used to protect the agent against attack. AES encryption is used in [4] to secure the agent and the agent communications. The theory here is to make the information the agent is carrying and the agent harder to interfere with.

A reputation based concept of trusted third party hosts (TTP) is proposed in [5]. TTP creates a reputation mechanism to determine the viability of a mobile agent moving to an agency based on the agency's repetitions. The approach is effective if the mobile agent always stays on friendly (trusted) agencies, however this approach losses some of it autonomy of the mobile agent. In this work the proposal removes the reason for interfering with the agent to enhance the protection rather than adding to the complexity of the agent system.

Protection of the system mechanisms using sandbox approaches proposed in [6] is the general accepted approach to protect the platform. This approach walls of the area of operation for the mobile agent reducing the ability of the code to do damage to the system. Although this is sound in principle, in practice it is a little more difficult to ensure. Conflict between the work involved in configuration management for the sandbox and the robustness for the service make the concept difficult to manage. This work proposes the isolation of the critical aspects of the mobile agent operation and the resource with operation available to the mobile agent reduced to only those necessary for operation.

Secure Socket Layers (SSL) has become the standard security mechanism communication. SSL allows two parties communicate using confidentiality, message integrity and endpoint authentication [7]. This work assumes that all external communication, including the transportation of the mobile agent, is accomplished in SSL.

Several data security mechanisms for mobile agents including: Partial Result Authentication Codes, Hash Chaining, Set Authentication code and Modified Set Authentication Code are evaluated in [8]. The results indicate that there is a trade-off between efficiency and security and the choice is dependent on the application. This work uses end-to-end chain-of-custody and point-to-point communications to protect the communication of the data.

3. Gold Miner Paradigm (GMP)

Gold mining was chosen as paradigm for mobile agent security because the operation has a direct correspondence between the components comprising and interacting in the mining operations and a mobile agent system. The gold mining also has many of the same security priorities to protect as the mobile agent environment. The analogy follows as the gold mining operation has the priority of extracting and protecting the gold (resource or data) in the mine (agent platform), while being mined (processed) and in transit (communication) to a remote location (client) from the mine. The gold mining operation has a miner (agent) who works for the client in the process of extracting the gold using tools (instructions) and transporting the gold via cart from the mine to the client. The mine owner (agent platform owner) has the concerns of security of the mine (agent platform) and the miner (mobile agent) to maintain operations (services). The application of the gold miner analogy allows the application of the security procedures and limitations of the gold miner operations as the base line for the mobile agent security. The GMP block diagram is shown in Fig. 1.

3.1. Description of the mining operation

The gold mine operation consists of four main components: the mine, the miner, the tools (shovel, pick, etc…) and the cart. The mine contains an aggregate of both the desired material by the host and other material that is unimportant and can be discarded. The mine for SMASGMP is the agent platform and the aggregate is the resource (data or service). As with the mine, the agent platform's task is to provide a place for the mobile agent to work and the resource to work.

The miner acts as an agent for an organization to extract the aggregate from the mine using a limited set of tasks. These tasks include: entering the mine, using a pick and other tools to separate the aggregate from the mine creating ore, using a shovel to place the ore in a cart, and exiting the mine. The gold miner for the SMASGMP is the mobile agent dispatched from a home agency to the local agency to use the resources available. The agent has a similar task list: entering the agent platform via the agency, using coded instructions to separate the information from the agent platform, storing the information in a cart and exiting the agency.

The tools the miner uses to perform the mining task maybe a pick, a shovel, heavy machinery, explosives, etc… and the tools are provided by the organization administrating the mine. The tools provided are limited to those required by the task the miner is performing and may have additional security, if the tools can accomplish missions outside of the intended purpose. For example, explosives would be secured as the intended purpose maybe to accelerate the extraction process in regards to the pick; however the explosive may also be utilized for a more malicious purpose like collapsing the mine. A pick would most likely be unsecured are the unintended results are not sever enough to the mining operations to warrant security.

The tools for the mobile agent are the instructions the agent platform provides or allows to accomplish the agent task list. The tools are provided by the agent platform and the mobile agents internal functions are not allowed within the secure section of the agent platform. Some of the tools may have addition security requirements to add robustness to certain agents, however using the ability should be done with caution as it leads back in the direction of relying on human management to insure security. The same could be applied to mobile agent's internal functions with the same security concerns. The limited functions using the GMP have the security concerns on par with the web-service with the only difference in the way the function is executed. The web-service call is made from remote calls were the mobile agent calls are local to the machine.

Note that even though the mobile agent may have more instructions available in the mobile agent object encapsulation than the agent platform provides, the instructions are still limited if the agent platform does not support them. This observation is one of the key attributes providing the ability to secure the agent platform and the data from the mobile agent without having to explicitly block the instructions.

The cart is where the ore the miner separates from mine is loaded using a shovel or authorized tool. Once loaded to the miner's satisfaction the cart ferries the ore to the exit of the mine. The miner does not control the cart after the dispatch to the exit and will no longer be in contact with the ore. This ensures that the gold is always in the control of the mining operation and the mine owner from the point of extraction to the point of use.

The same cart concept is used for SMASGMP to supply this same level of control of the data. All data extracted by the mobile agent must be placed in the cart and SMASGMP will retain this control to the point of delivery to the requesting agent's dispatcher. In contrast, a standard mobile agent system would store the retrieved information with the agent in the agent's brief case. This difference allows for additional layers security to be enacted as explained in the next section.

4. Gold mine owner and security

The main purpose of the gold mine is to extract benefit from the mine for the owner. Naturally, the mine owner wants to protect the operation and the gold from activities not in his interest. The information server owner has the same concerns in protecting the interests and operation of the system. In both cases the design described in Sec. 3 provides only the framework for the operation with some security interweaved through the design, however to insure the system and components adhere to the restrictions additional security is required. The focus of the additional security is the host organization responsibilities to protect the mine from the miner, the ore from the miner and the miner from the mine. Protection of a miner from another miner is generally left to outside agencies; however this concern is lessened by the paradigm as well.

4.1. Protection of the mine

The gold mine is general thought of as a resource without vulnerability; however the mine is vulnerable to many types of attacks from malicious forces (disgruntled employees, competition, etc…). A server containing data is even less resilient and has even more malicious forces and possible attacks to protect against. In both cases the service provided has to take place and additional security needs to be applied to secure them from the hazardous forces.

A mine has limited access points or entry shafts with security at each entry point. The security guards at these points inspect the miners as they enter and leave the miner. A miner requesting entry will be searched in several ways to insure contraband is not brought into the mine. Security may also get an entry metric, such as weight, image or metallic content, of the miners to be used for comparisons of the miner at time of exit. Anything not allowed can be stored in lockers to be reclaimed when the miner exits. After the miner has completed the work required security inspects him again for difference between the entry and exit which may find stolen gold.

Applying this portion of the gold miner paradigm the agent platform will have limited access with security at the entry (Security In) and exit (Security Out) point. Mechanisms like access lists [9] and credential verification [10] can be used to restrict specific agents from entry. The allowed agents must then be searched for malicious functions that may conflict with the operation supported in the Limited Access Area. The removal of added functions insures the limited tools protection described in Sec. 3 is valid. Any contraband material or functions maybe checked into the Baggage Check to allow a successful search. In addition, any state information that the agent wants to retain should be stored in the Baggage Check as the agent will be required to be at its original state of inspection upon leaving the mine or at minimum a difference that does not indicate data has been removed. A modularization of the mobile agent as depicted in Fig. 2 allows the agent to quickly check this information without effecting other operations. The agent has the base agent code separate from the task methods, state information and the brief case to allow the code to be removed and checked. The brief case can remain encrypted to prevent snooping if required. An image will be taken of the agent after contraband has been relinquished and state information stored. The purpose of this image is to give a metric for differencing the entering agent from the existing miner. The image will be used in the protection of the gold (data).

4.2. Protection of the gold

The gold mine's main purpose for existence is the gold as without it the operation described would not take place. Protection of the gold requires restrictions to exist in order to insure the security of the material is maintained. Some of the restrictions are embedded in the operation itself as described in Sec. 3. The limitation of the functions supplied to the mobile agent and the cart define a separation between the ore and the agent. The functions supplied are unidirectional with the output (ore) supplied to the cart and not returned to the agent. In addition, security elements of the gold mine are added to retain the chain of custody of the ore from data source to the point of delivery.

The chain of custody is the key in insuring the ore never has a chance be out of the control of the mine operator. A web-service implementation follows this rule as the data is always in control of the host until the point of delivery to the client. The chain of custody for a mobile agent is typically more agent oriented which relies on the agent's retention of the data that is mined and violates the chain.

In order to add the chain of custody into the SMASGMP while maintaining the functionality of mobile agents, we compliment the "Security In" inspection with a "Security Out" inspection. The inspection uses the image taken upon entry to determine the difference between the mobile agent entering and exiting. The image and differencing can be in the form of a check sum [11], code differencing [12] or other differencing techniques with exact details dependent on implementation. Regardless of the implementation the purpose is to exam what changes occurred between the entry and exit of the mobile agent from the Limited Access Area. The changes found, if any, must be determined to be within tolerances set by the security.

In addition, GMP uses a courier agent service to maintain the chain of custody all the way back to the dispatcher of the agent or an agent specified location. The courier is another mobile agent with the responsibility of transporting the ore in the cart from the host server to the client mobile agent's dispatcher or another location defined by the client mobile agent. The destination information and any required authorization credentials are exchanged at "Security In". The use of the courier agent insures that the chain of custody is always in the hands of the host until the actual point of delivery. The end-to-end data exchange of data helps the host acquire a record of who is using the information and can be used to track anomalies related to security.

The complete chain of custody allows the ability to record and track what information is being used by the agents. Tracking the ore is similar to how the source mine of gold can be tracked using the trace elements contained within the sample. The data service adds meta-data tags to the source data and by maintaining these tags we can track the ore or determine its origin and relevance to the area in which it was discovered. Data without these source tags or with source tags not conforming to the security will not be allowed to exit the server and the mobile agent can be detained.

4.3. Protection of the miner

Protection of the miner requires defense from the mine and other miners. In a real gold mine the mine needs to be maintained and safety guidelines followed to insure the safety of the miner. In the virtual world of GMP most of these are the responsibility of the server admin and are not covered here. What is important is the description of how the separation of the data and the methods using GMP helps in mitigating the risk of agent to agent and platform to agent attacks.

Most of the concerns or motivating factors of agent to agent and platform to agent attack are in regards to the data they would carry or the modification of the functions they need for their work. Removing the data couriering option in GMP from the agent reduces the data threat entirely. Since the data (gold) is no longer transported by the agent there is nothing to attack. Without the data the value of interfering with the agent is reduced to only interfering with it tasks, however the key tasks in GMP are secured on the platform and separate from other agents. The separation of function from the mobile agent proposed by GMP reduces the functional attacks as well.

Since data attacks are eliminated and function attacks are lessened the remaining attacks are platform against agent. The dispatcher could be used to hoarse the courier agent; however that would be counterproductive as the dispatcher wanted the information. The data host platform could also attack the agent; however this is also counterproductive to the mission of the platform in facilitating this exchange of information. Other platforms and agencies could hoarse the agent as well but without the data there is little point at least in regard to the service provided by the dispatching and hosting agencies. In short, the isolation of data and functions from the mobile agent as it moves is very efficient at removing any benefit of agent to agent and platform to agent attacks.

5. GMP and FPGAs

The common implementation of mobile agents is using virtual machine architecture such as Java Virtual Machine (JVM) [13] or mobile C [14]. The virtual machine is hosted by an OS on a physical server which provides network service and data access the VM. This provides a semi-separation of the VM and the data and provides the basis for the sandbox approaches typically used for mobile agents. The main problem with this approach is the VMs are typically used for multiple purposes which lead to conflicts between needed functionality by the application and the security required when trying to secure the system. GMP using an alternative approach to separates the general mobile agent architecture used from secure area. The mobile agent area can then be as robust as required while maintain the principles described out by GMP.

In order to accomplish the separation of the mobile agent and the system hosting the information the Limited Access Area in fig. 1 should be a separate entity. Field Programmable Arrays allow for this possibility by providing a hardware solution that is reprogrammable. A set function and interface to the service and data elements provided by the host can be easily implemented and changed as needed while maintaining the security set by GMP. The hardware solution also limits the ability to exploit the system by following the GMP of limiting the tools to only what is needed. Security In and Out could also be implemented in hardware to maintain there integrity. The JVM and Mobile C options can be utilized as the initial interface if need to minimize development effort required. However, the there needs to be a strict separation between those environments and the date and service provided as described by SMASGMP.

6. Viability of SMASGMP

The purpose of GMP was to mitigate the security concerns of MAS to that of a web-service. In order to accomplish this feat, the SMASGMP system limits the functions available reducing some of the value of mobile agents. However, SMASGMP is just as functional and secure as a web-service and a metric is needed to determine when to use a web-service and when to use SMASGMP.

The two main components to web-services and mobile agents are processing and communication. Processing in time and space for both mobile agents and web-services should be equivalent. The reasoning follows that although mobile agents and web-services process in distinct locations, the end result will be equivalent as either method is given the same set of instructions and data to produce the same result. The only difference is the location at which the processing occurs.

Communication differs drastically between the paradigms as MAS only requires the results required, where as a web-service would required the entire data set to be transferred. The advantage of mobile agents is the local processing to the data allows the communication to only contain needed information reducing the load compared to the web-service. The overhead is in the MAS and the security required.

Assuming that both the web-server and the SMASGMP are implemented we can make a comparison between the two based on time.

$$T_{ma} \leq T_{ws},$$

where T_{ma} is the total time for the mobile agent and T_{ws} is the total time for web-service to complete a task.

Expanding the equation to separate the respective overheads obtains:

$$Tcomm_{ma} + Tsec_{ma} \leq Tcomm_{ws} + Tsec_{ws},$$

where $Tcomm_{ma}$ and $Tsec_{ma}$ are the respective communication and security times for the mobile agent and $Tcomm_{ws}$ and $Tsec_{ws}$ are the respective communication and security times for the web-service.

Since both the mobile agent and web-service can communicate on the same communication backhaul the difference in $Tcomm$ between the two will translate directly to the size of the data. The values for $Tsec$ depends on the implementation of the security mechanisms chosen for specific implementation. However, both could be close to equal if the local mechanism is a localized web-service. Supposing the transfer time of the mobile agent is minimal, the entire operation is primarily dependent on data size and transfer rate.

In short, the reduction states that the larger the data transfer the more mobile agents are viable and web-services less desirable. Applications with larger data sets, like Geographic information systems, would be better served by SMASGMP and limited data transfers, like on-time weather data, are better serviced by a web-service.

7. Conclusion

SMASGMP main contribution is to allow mobile agents to be applied with security concerns to that are on par with web-services while allowing the mobile agent to maintain its autonomy and mobility. SMASGMP addresses the security of M2M operations by protecting the platform from the mobile agent, protecting the data from the mobile agent and protecting the mobile agent from the platform. The protection is accomplished by limiting the access of the agent to the secure area, limiting the tools of the operation available and inspecting the agent entering/leaving the secure area while maintaining control of the data throughout the process. A set of equations is provided to help determine when mobile agents have a benefit over web-services based on each services predominate factors. The application of the gold miner analogy allows the application of the security procedures and limitations of the gold miner operations as the base line for the mobile agent security.

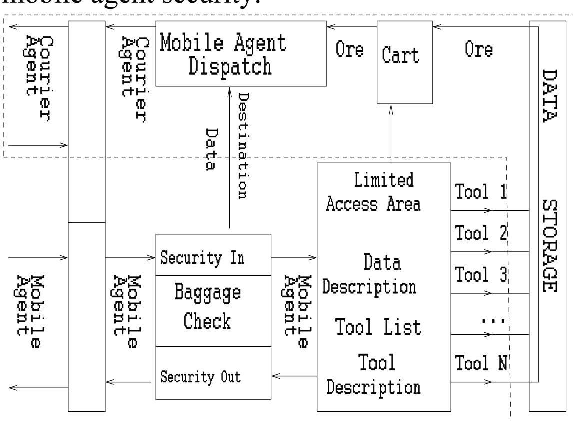


Figure 1. Gold Miner Paradigm (GMP)

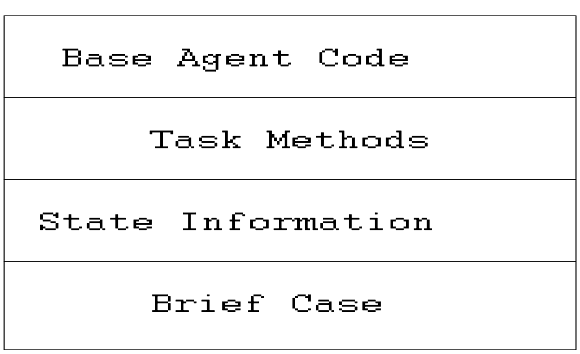


Figure 2. Agent segmentation into major tasks for easy baggage check in GMP

8. References

[1] J. White, "Mobile Agents White Paper," General Magic,1998. http://www.klynch.com/documents/agents/.

[2] W3C, "Web Services Architecture," W3C, 2004, http://www.w3.org/TR/2004/NOTE-ws-arch-20040211/.

[3] W. Jansen and T. Karygiannis, "NIST special publication 800-19 – mobile agent security. Technical report", National Institute of Standards and Technology, Computer Security Division, Gaithersburg, MD 20899. U.S., 2000, http://csrc.nist.gov/publications/nistpubs/800-19/sp800-19.pdf.

[4] D. Lin and T. Huang, "A Mobile-Agent Security Architecture," e-Business and Information System Security (EBISS), 2010 2nd International Conference, May 2010, pp.1-4.

[5] O. Bamasak and N. Zhang, "A distributed reputation management scheme for mobile agent based e-commerce applications", in Proc. of 5th IEEE International Conference on e-Technology, e-commerce and e-Service, 2005, pp. 270- 275.

[6] A.D. Rubin, D.E. Geer Jr., "Mobile code security," Internet Computing, IEEE, 1998 vol.2, no.6, pp.30-34.

[7] L.D. Bisel, "The Role of SSL in Cybersecurity", IT Professional, March 2007, vol.9, no.2, pp.22-25.

[8] V. Gunupudi and S.R. Tate, "Performance evaluation of data integrity mechanisms for mobile agents", Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. International Conf., April 2004, vol.1, pp. 62- 69.

[9] R.S. Sandhu and P. Samarati, "Access control: principle and practice,"

Communications Magazine, IEEE, Sep 2994, vol.32, no.9, pp.40-48.

[10] N. Karnik and A. Tripathi. "Agent Server Architecture for the Ajanta Mobile-Agent System", In Proceedings of the 1998 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'98), July 1998, pp. 66–73.

[11] R. Braden, D. Borman, and C. Partridge. "Computing the internet checksum", SIGCOMM Comput. Comm. Rev. 19, 2 April 1989, pgs. 86-94.

[12] J.I. Maletic and M.L. Collard, "Supporting source code difference analysis", Software Maintenance, 2004. Proceedings. 20th IEEE International Conference, Sept. 2004, pp. 210-219.

[13] T. Lindholm and F. Yellin, *Java Virtual Machine Specification (2nd ed.)*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[14] S.S. Nestinger, Bo Chen and H.H. Cheng, "A Mobile Agent-Based Framework for Flexible Automation Systems", Mechatronics, IEEE/ASME Transactions, Dec. 2010 vol.15, no.6, pp.942-951.

Distributed Decision Support System for Network Security

James Jacobs
Center for Defense Integrated Data, Jackson State University, Jackson, MS 39204, USA
jjacobs@c-did.com

Abstract

Network and system security is paramount requirement for all organizations in today's age and the requirement to be connected to remote systems outside of the administrator's control provides a tremendous advantage to the attackers. The Distributed Decision Support System for Network Security (DDSSNS) framework connects the administrators to interconnected systems and provides the ability for an administrator to see the communication originating and/or delivered to their networks at the network that the communication was originated from / or was addressed to. The duplex view of the network allows for direct diagnoses and / or resolution of attacks without the aid of remote admins.

1. Introduction

Network and system security is a paramount task for all administrative personnel of interconnected systems. The various subtasks, like managing firewalls, networks, systems and resources, related to this endeavor and the diversity of knowledge required is daunting for any admin. In addition, the threats against the managed systems are prevalent and constantly changing, which makes the learning rate and the ability to acquire knowledge key concerns. For externally connected networks, like those connected to the internet, the security task is compounded as the administrator and support personnel are facing unknown attackers in an area that is out of their direct control. The only clue to the solution is the symptoms observable on their network which may be indicative of a variety of attacks or other problems caused by software or hardware malfunctions.

Communication between administrative personnel of the interconnected networks is often required and composite solutions are often implemented in response to the attacks. Since the network is a critical component to the business, the disruptions of service require a real-time response and resolution by the support personnel. The communication requirements and the limited details add pressure on the local admin to be efficient as possible and increase the delay in service restoration and the cost of service to the ISP in support dollars. The following scenario illustrates these points.

A standard scenario for responding to an external attack on a network usually requires two phases of operation for an administrator. The first phase is acquiring and implementing a solution to the symptom in the local network. The standard tact for acquiring the solution is to analyze the symptom for a list of possible causes within the administrator's domain. This usually requires the analysis of log files or compiled network statistics. The task's difficulty is compounded by the difficulty and the inability to look for the actual cause of the problem due to the administrator's scope of vision being limited.

Once a cause is located the admin must acquire a solution from either previous knowledge or by searching the internet (if available) for a solution and then implement the solution. If the internet is not accessible the admin may have to find a local resource or communicate with outside help using another means to acquire a solution. If external help is required, additional delay in the response to the outage an additional cost maybe incurred if fees are required for support. After the solution is found it is implemented and in most cases this implementation degrades the interconnected service until the actual cause is rectified.

The second phase is actually identifying the cause of the problem in the external network and acquiring a complete solution. The second phase always requires communication with outside sources, such as administrators of the ISP/access point, administrators of other effected networks and/or administrators of the attacking source network. This process is very timely as the remote administrators have to identify the cause on their respective networks, identify a solution and then implement it using the same process described in the first phase. Compounding this problem is the networks sizes tend to increase as the ISP escalates the problem to the supporting backbone networks through their respective access points. Meanwhile, the attacked admin waits for a response with a degraded system. Finally, with the external cause is rectified and the local admin is notified and returns his network to a nominal state. The nominal state typically results from removing the changes implemented allowing the problem to reoccur with a different attacker.

The key issues with the described approach are the difficulty in troubleshooting outside of administrator's own network and the inability to acquire an end-to-end solution efficiently and permanently. In either case the common requirement is the administrator's communication and collaboration with the other administrators to acquire information outside of their sphere of influence. The image of an interconnected networks is depicted in Fig.1 with each administrator's view limited to what is contained within the dashed circle with each connection between these networks referenced to as access points.

In addition, the scale of the Internet in both size and data transmission rates reduces the ability to add security mechanisms to the connections within the WAN / Internet structure without having a large effect on the speed and reliability. The lack of security mechanisms within the Internet places the burden of security on the networks and systems connecting to it for service and ultimately on the administrators of the connected networks and systems. The lack of other solutions and the reliance on the administrators dictates that collaboration between the administrators is the best solution for the problem and Distributed Decision Support System for Network Security (DDSSNS) provides the mechanism for this collaboration.

The objective of the DDSSNS is to provide administrators with a real-time ability to communicate and interact with other administrators to gain information and increase the vision of the root cause of the issue plaguing their network. DDSSNS provides the end-to-end vision to a connecting administrator and a common resource that allows for a coordinated defense against attack and the ability to search for solution to mitigate attacks in real-time.

Although the framework has the ability be applied to networks of various scales, the main focus is networks and systems connected to the Internet and WANs. These connections require extensive collaboration with agencies and administrators that our outside the control of the party or parties being affected in an attack.

2. Background and current solutions and approaches

The scenario of a network attack described in Sec. 1 has several common causes and attack mechanisms including: Denial-of-Service (DoS), Man-in-the-Middle (MITM) and port scans. The attackers usually use spoofing of protocols, like IP (Internet Protocol), and Bot-Nets to distribute the attack on unaware systems without revealing the true source of the attack. These problems are typically addressed with firewall and routing rules to block the attacker from the target destination.

IP spoofing [10] is the process where a malicious system uses and IP that is not genuine or registered to the system (A false IP). IP spoofing allows a malicious host to send a packet onto a network and mislead the receiving network on the IP traffics origin and intentions. IP spoofing also allows the malicious host to emulate another valid IP to reroute connections to itself creating a relay or interception of IP packets. Hackers use the spoof to gain information, and intercept and manipulate traffic to perform multiple forms of mischief.

A Bot-Net [12] is a consortium of computers connected to the Internet with malicious code controlling some portion of the host. The owners of the actual machine are usually unaware that their system is being controlled remotely. The Bot-Net provides anonymity to the attacker and also a powerful combination of distributed computing power and network resources making the Bot-net a formidable foe. The Bot-Net is used by a master controller to launch coordinated attacks like DoS.

DoS [3] is a type of network attack used to render a service inaccessible by overloading a host resources for a given service. This attack is usually done by flooding a single destination from multiple sources and overwhelming a machine's resources causing valid service requests to be denied service. An identification method for a Distributed DoS attack using queue models Gaussian Mixture Model (GMM) is proposed by Hao et al [5] and several solutions are given by Cisco Systems [3]. However, DoS attacks can take many forms and the solutions provided only address the local symptoms, not the cause of problem. DDSSNS is aimed at both sharing the knowledge of a current attack gained by [5] or other means and more importantly addressing the source(s) of the DoS.

MITM [10] is a type of attack where a server is placed in the middle of a communication channel to eavesdrop and / or manipulate the data stream. This attack is usually accomplished by using some type of spoofing technique like IP spoofing. A current solution to this issue is addressed with Cisco switches [11] using a trust based system and validation of tabulated connection information. When a violation occurs the offending post is disabled. This solution works for local networks where the networks composite information is available, however the separation of networks in WAN / internet environments inhibit this solution. DDSSNS allows the access-points to determine that a MITM exists and mitigates the hack from a net-centric view.

Port scans [13] are attacks where a remote computer attempts to identify open ports and service on a host in order to make connections and/or gain access into a secured network. The identification of a Port scan may indicate an attack on a vulnerable system [14] or they may indicate the potential of an attack. At minimum they typically indicate a potential attacker is interested in accessing a system. Common efforts to address this problem involve patching systems with the vulnerabilities and silencing ports and networks by removing functions such as ping, trace route and telnet. The problem with this approach is the tools mentioned were created to help administrators troubleshoot their networks and the networks of others. In addition, ICMP, used for ping and trace route, is required when updating to IPv6 [7]. DDSSNS allows for the IPv6 upgrade by blocking of the attackers requests at the source while allowing valid request to remain allowing for ICMP.

Firewalls are the most common solution the network security. They have various features [2] to isolate the user's network from the larger internet with hope that intruders and/or disruptions are localized to the periphery of the administrator's network. Several approaches use collaboration between administrators allowing the sharing of information between collaborators. Collaboration helps disseminate configurations like blacklists, signatures, firewall rules, access control lists, and related security policies and known solutions [1]. A Collaboration Network Security Platform [9] using a P2P network overlay allows for the communication of secure services to multiple parties. Some services may have addition analysis using inference models like Bayesian [6]. A proposed system using Collaborative Intrusion & Malware Detection (CIMD) [4] enhances this collaboration by grouping similar architectures under the assumption that the administrators would require similar information based on their hardware and objectives. Several approaches use centralized policies to try to address security concerns. A centralized approach using a Policy Based Security Manager (PBMS) [7] server is utilized to manage and maintain policies addressing concerns with IPv6 [8].

DDSSNS is designed for both sharing the solution information and the sharing the actual data related to the attack. The sharing of the data is accomplished by increasing the scope of the administrator's view. The change in scope allows the administrator to acquire real-time information on the problem and see what is happening throughout the network to make an informed decision.

3. Connection and communication

The main task of DSSSNS is to create connections between all the key administrators interconnected within a WAN to provide the ability to share information. In order to illustrate this point a change in focus from an Internet view of the network in Fig. 1 to a single set of interconnect networks on a WAN in Fig. 2. It is noted that any or ALL of the interconnected networks could be another WAN without loss to the concept. This ability provides the ability to scale the proposed framework from a single network ISP with interconnected customer networks to ISP to ISP or any other interconnection scheme.

In Fig. 2 the network is reduced to the more common net-centric perspective of the collection of external links to a WAN. The figure illustrates the global impact of the two party's view of the collective domains and an interconnected network. All of the admins of the connected networks are on one side of the dashed line and the WAN's admin is on the other. The diagram also indicates the direction the admin will troubleshoot an issue as he/she must call the cloud / WAN admin when it is most likely the issue is not in the cloud but started by one of the other connected networks outside of the ring. Ideally the affected admin and could see through the cloud and to the other side at the source of the problem.

In order to accomplish this task the DDSSNS nodes must be located in an area accessible to all parties. The proposed location is between the customers firewall/external connection router and the WAN's Access point. The location is illustrated in Fig. 3 with added interconnect from both the ISP's access point and the connected networks. The placement of the DDSSNS nodes in this location makes each node dual-homed giving the node the ability to be visible and accessible by the both the connecting networks admin and the WAN admin. This ability is not possible with a standard ISP access point as its purpose is only to provide access functions to the network, such as authentication and network connectivity.

In addition to the dual access provided, the additional links facilitate the ability to maintain a connection to the WAN when the primary connection is being degraded by attack. The admins will still have access to the Internet for troubleshooting purposes and the DDSSNS for communication with other members. The communication mechanism will stay active as attacks on access points do not exist. Also, the connection and routing for the access points are all controlled by a common owner organization which will severely limit the ability to attack the access points directly in the future.

The DDSSNS nodes ability to maintain connection to the WAN allows for the ability to create and maintain the distributed natural of the DDSSNS and facilitate communication between all the admins connected to the WAN during the time of disruption. The distributed approach also provides a consolidated front at the access points similar to the consolidated firewalls proposed except in this proposal the access point can stop traffic going from the network to the WAN and from the WAN to the network. A consolidated firewall only blocks traffic into the network.

It is proposed that a P2P networking scheme is used to create the interconnected DDSSNS. P2P's key benefit is the ability to be dynamic which is critical in WAN environments with networks connecting and disconnecting at-will. P2P also provides the ability to distribute the intelligence and organize into like groups [15].

4. Sharing information and decision

Other distributed systems simply provide the results of their analysis, i.e. blacklists, white lists. The purpose here is not only to share the results of complex analysis but to share raw information as well. The primary premise here is that an attacker is not only going to attack one location, the attacker will attack multiple locations. The secondary premise is that the attacker will use the same pattern in attack from one location to another. The dissemination of raw logging or statistics at the access point provides the ability to look outside the scope of an admin's domain and confirm or deny if the suspect is an attacker or just a standard user.

As with every DSS, DDSSNS has AI in the form of inference engines, statistical packages and other supporting intelligence tools. The ability to use AI in the friend versus foe identification also provides that software ability to match and generate the patterns or signatures of attack. The patterns generated by the AI can subsequently be passed to other members of the DDSSNS. The advantage of using patterns versus using specifics is that the pattern is not as easily modifiable by the attacker, like blacklists and other source identification techniques. In essence the code the attacker uses needs to change where blacklists only need the source to change.

Information security is maintained by only allowing the information that was sent from a network to a receiving network to be accessible by the connected network admins and the WAN admins. This level of access is the same as if the networks were working properly. No information will be monitored or relayed that was not destined for the system and released from the sending server. The exclusivity of ownership between the sending and receiving parties is actually more secure as the WAN admin did not have to review the information when the problem is resolved by the sending and receiving admins.

5. Application toward network attacks

The ability to see end-to-end allows for a trace to be enacted for any traffic entering the network through an access point destined for target system. This includes access points inter-connecting back hauls and other networks external to the access-points controlling entity. The trace allows for the identification of the source by identifying the access point by which the offending traffic originates. The admin would then have the ability to contact the administrator of the system using the access point and if necessary block the traffic at the source until a resolution can be found.

5.1. DDSSNS applied to DoS Attacks

Applying the framework to a DoS attack allows for location of the actual source within the system and contact of the admin of the source of the attack. The sources or WAN admin can then disable the access point of the offender and /or block the traffic and contact the offending system admin using the access point through the local DSS node. The details of why the access was denied can also be supplied and the history of offense maintained. If the attack was a genuine mistake, such as a runaway loop or script sending spam, the offending admin can fix the issue and contact the access point admin for reinstatement. This process can also be automated to allow easy re-instatement.

The use of DDSSNS for DoS is a major shift from the current paradigm of the affected admin simply blocking the offending IP in the firewall and attempting to contact the admin of the offending system. The offending system and access point would not be affected and the excess traffic would still exist on the WAN. The firewall would have to constantly check the IP and block accordingly until the offending system was corrected, which may never occur.

Mitigating DoS attacks is even worse if the IP was spoofed as the stated IP is not actual sending the traffic. The firewall block could be affecting a system that is not actually part of the attack if the IP being spoofed is a user of the blocking system. Contacting the admin of the spoofed IP would only result in negative results as the traffic was not from there. At this point the firewall rule would have to remain indefinitely and the admin would need to supply and alternate route to the effected customer.

Blocking incorrect IP addresses is a large concern for e-commerce sites that do not have a constant customer base. The firewall could block a potential customer, so the firewall solutions are typically time based, meaning the rules expire after a period of time. The expiration allows for the attack to resume and the resolution process is started over again.

The ability to jump from the affected system to the origin provided by DDSSNS is a major advantage for this form of attack. The end-to-end approach allows for the affected admin to verify that the IP was spoofed and act accordingly saving considerable time. Also, since all the input nodes to DDSSNS are connected the admin has the ability to request all of the DDSSNS nodes on the network for any traffic sent to affected system and locate the actual access point that sent the spoofed traffic. That access point can then drop the spoofed traffic at the source. The only way to get this same result with current methods is for each intermediate admins to analyze the network for the incoming traffic in his/her network and subsequently contact the next admin until the offending access point is located. This takes a considerable amount of time and money and is typically not done unless large infrastructure is affected.

If a BotNet is responsible for a DoS attack the originating IP would not know unless for some other reason the affected system had to communicate with the blocking firewall. DDSSNS allows for the admin of the system with the BotNet to acquire why the network was blocked and acquire information from the affected network to determine where the Bot-Net is located in his network. Today's firewall blocking technique does not supply these abilities.

5.2. DDSSNS applied to Man-in-the-Middle Attacks

Unlike DoS the MITM attackers does not actually transmit information which makes this attacker stealthier. In addition, most man in the MITM attacks occur on separated networks as two identical IPs on the same routed network would cause conflict in routing tables without the adequate dispersion. In order to address the problem of stealth and the dispersion, the DDSSNS is scaled up to encompass the entire network. It is then possible to request the originating node to validate the destination is correct for the first or intercepting server by taking to the DDSSNS node. In addition, a verification of the IP / DDSSNS node combination could also be used to determine if the MITM exists using reply requests. If a BotNet or a malicious network is responsible for this action the machine can be removed from the network much the same as for the DoS example.

5.3. DDSSNS applied to Port Scans

Applying the framework to a Port scans is much the same as DoS except the DDSSNS has to first determine the intention of the Port scan. DDSSNS can use the analytics of the DSS to determine if the Port scan is malicious. The DDSSNS node can verify the origin is not spoofed and can talk to the admin of the access point without alerting the suspected host's admin. If the Port scan is determined to be malicious the access point can disable access for the malicious host. If a BotNet or virus is responsible the originating IP can be fixed and the network access restored.

6. Conclusion

The DDSSNS framework proposed in this work connects the administrators of the interconnected systems of a WAN or Internet to provide the ability to communicate, interact, and share information with other agencies commonly connected on a network. The main advantage is the ability to see a transmission from end-to-end by placing decision support systems between the connecting networks and the ISP access points. The dual-homed DDSSNS nodes and the added connectivity provides the administrator a more holistic view of his network's communications, which is a tremendous advantage over current methods that only allow an admin to see half of their network's communication. The ability to see the end-to-end transmission on a WAN provides a greater ability to mitigate attacks like DoS, MITM and Port scans. The foreseeable disadvantages are added security and cost of architecture for the system. However, these disadvantages are minimal compared to the added work and security loop holes given the current environment. Proposals are currently pending for the development of the system.

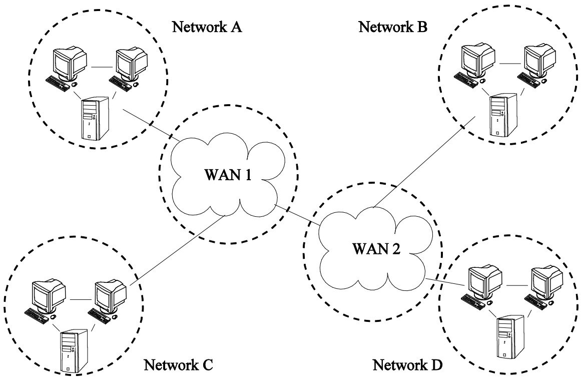


Figure 1. WAN connections for WAN 1 connecting Networks A and C to WAN 2 connecting Networks B and D

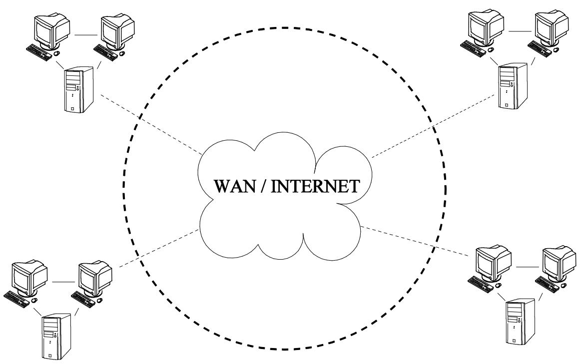


Figure 2. The current view and model of WAN/Internet connections with the dashed ring representing the combined firewall

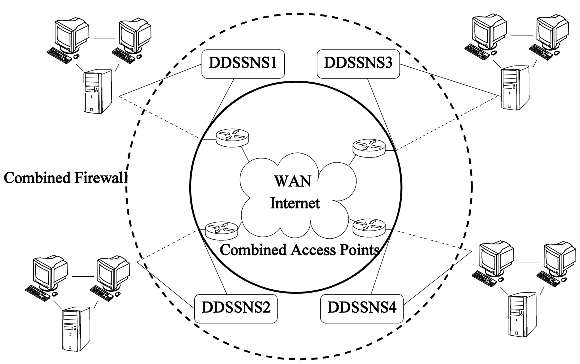


Figure 3. Distributed DSS for Network Security nodes 1-4 with the inner ring depicting the combination of access points. The ring contains the first hop routers and the WAN cloud

7. References

[1] J.L. Thames and R. Abler, "Implementing distributed internet security using a firewall collaboration framework," SoutheastCon, 2007. Proceedings. IEEE, March 2007, pp.680-685.

[2] L. Yeo, “Choosing a Personal Firewall,” InformIT, June 2003, http://www.informit.com/articles/article.aspx?p=31945.

[3] Cisco, “Understanding the Basics of DDoS Attacks Document ID:13634”, Cisco, San Jose, CA 95134. U.S., April 2008, http://www.cisco.com/en/US/tech/tk59/technologies_white_paper09186a0080174a5b.shtml.

[4] R. Bye, S.A. Camtepe and S. Albayrak, "Design and modeling of collaboration architecture for security," Collaborative Technologies and Systems, 2009. CTS '09. International Symposium on , 18-22 May 2009, pp.330-341.

[5] S. Hao, H. Song, W. Jiang and Y. Dai; , "A queue model to detect DDos attacks," Collaborative Technologies and Systems, 2005. Proceedings of the 2005 International Symposium on , vol., no., pp.106-112, 20-20 May 2005.

[6] C.J. Fung, Z. Quanyan, R. Boutaba and T. Basar, "Bayesian decision aggregation in collaborative intrusion detection networks," Network Operations and Management Symposium (NOMS), 2010 IEEE, 19-23 April 2010, pp.349-356.

[7] A.R. Choudhary and A. Sekelsky, "Securing IPv6 network infrastructure: A new security model," Technologies for Homeland Security (HST), 2010 IEEE International Conference on , 8-10 Nov. 2010, pp.500-506.

[8] IETF RFC 2460, “Internet Protocol, Version 6 (IPv6) Specification”, December 1998.

[9] C. Wang and C. Huang, "A Collaborative Network Security Platform in P2P Networks," New Trends in Information and Service Science, 2009. NISS '09. International Conference on, June 30 2009-July 2 2009, pp.1251-1256.

[10] Toolbox.com, “Man-in-the-Middle Attack,” InformIT, Dec 2008, http://it.toolbox.com/wiki/index.php/Man-in-the-Middle_Attack.

[11] Cisco.com, “Cisco Catalyst Integrated Security-Enabling the Self-Defending Network,” Cisco, 2004, http://www.cisco.com/en/US/prod/collateral/switches/ps5718/ps708/prod_white_paper0900aecd8015f0ae.html.

[12] techtarget.com, “Botnet,” techtarget.com, Dec 2004, http:// http://searchsecurity.techtarget.com/definition/botnet.

[13] L. Teo, “Network Probes Explained: Understanding Port Scans and Ping Sweeps,” Linux Journal Online, Dec 2000, http://www.linuxjournal.com/article/4234.

[14] Cisco Systems, " Cisco Security Advisory: Cisco Security Agent for Linux Port Scan Denial of Service ", Cisco Systems, Oct. 2006, Document ID: 71902.

[15] J. Jacobs. “Distributed Intelligence Framework for P2P File Sharing.” *Journal of Management and Engineering Integration*, Vol. 3, No. 1:33-40, Summer 2010

Learning Curve Analysis of a Haptic Controller

Carmen Franklin, *Zongliang Jiang, Xiaochun Jiang
Department of Industrial and Systems Engineering, North Carolina A&T State University, Greensboro, NC 27411
*Corresponding author: zjiang@ncat.edu

Abstract

An aging boomer generation continues to reduce the size of the qualified working population, especially in the construction industry. To deal with this challenge, contractors often incorporate larger proportions of female workers to the predominantly male construction labor force. To alleviate concerns from management regarding short-term negative impact on productivity when a new technology such as the haptic controller is introduced to the workplace, especially on the growing female work force, this empirical study aims to investigate the effects of gender on the learning trends in performing a simulated excavator task using a haptic controller. Twenty university subjects were recruited. Each subject performed ten repetitive simulated excavator tasks using the haptic Phantom controller. Times and deviations between each deposit and the target were measured to calculate the learning percentage and error measures. Results show that learning patterns are similar among males and females.

1. Introduction

The retiring boomer generation is creating a shortage of young qualified workers, especially in the construction industry. In response to this challenge, contractors often hire larger number of unconventional groups, such as female workers, into the predominately male labor force [1]. Accompanying the changing composition of the workforce is the evolution of construction tools and equipment with advanced technologies in recent years. One such example occurs in hydraulic excavating machineries, where design engineers are in the process of better understanding the benefit of applying control mechanisms that utilize haptic feedback to replace the traditional mechanical controls/joysticks. As both the construction workforce and technology evolve, it is important to understand the dynamics of the interaction between these two trends.

Haptic technology is a type of human-machine interface that interacts with users through the sense of touch and force feedback. Controllers that employ this technology allow users to experience tasks in real or virtual environments. Haptic controllers are utilized in many fields to enhance usability and operator performance, such as rehabilitation, medical training [2-4], tele-operation and industrial employee training [5] (Lim, 2004). The construction industry is a new frontier for haptic technology with a goal to increase operator performance, comfort level and productivity [1].

The performance outcome of haptic controller operations has been found to largely correlate with individual hand-eye coordination abilities and overall motor skills [4]. The abilities of performing certain tasks and processing information are reflected in different hemispheres of the brain. [6] stated the left hemisphere of male's brains is larger than in female brains, giving men better spatial abilities. Therefore, men have an advantage over women in hitting targets, mentally manipulating maps, computing mathematical problems, and breaking items into their component parts [7]. [8] stated that females score lower when completing cognitive tasks such as hitting targets or computing mathematical problems. As females analyze figures or break down components, they have the tendency to change the orientation of the illustration rather than processing a perception mentally [8]. Females were also found to understand how mechanical gadgets work on a more theoretical rather than technical level, where they favor smaller and simpler methods in problem solving. Those theoretical methods are more likely to be obtained from prior knowledge or experience [9].

When a new technology such as the haptic feedback controller is introduced into the workplace, a short-term negative impact on the productivity is usually inevitable. Management often has concerns about its short-term impact on productivity and, in this case, its effect on the growing female workforce. To understand the initial interaction and fit between the operator and the machine, the progression of an operator's performance can be analyzed and predicted by the learning curve approach. The learning curve theory is essentially a structured mathematical system to determine how tasks time decreases during the repetition of the same tasks. It is able to predict how long it will take for a person to reach a given level of performance on a task. With the expansion of female workers in a predominately male workforce, understanding the performance of all users over a period of time will help alleviate concerns at the management level, provide recommendations for equipment design strategies catered towards diversified user population and assist in developing a more effective employee training program. The equations used to obtain the learning percentage for a task are as follows:

$$Y_x = KX^N \text{, or} \quad (1)$$

$$N = \log(Y_x/K) / \log(X) \text{, and} \quad (2)$$

$$\text{Learning Percentage} = 100*2^N \quad (3)$$

where Y_x is the production time for Xth unit in sequence, K is the time required to produce the first unit in the measured task, X represents the total units produced, and N is an exponent leading to the learning rate, which is the slope of the line depicting change in task completion time (y) as a function of repetition number (x) in log-log space.

Learning in general can be categorized into cognitive learning and motor learning. Cognitive learning percentages have been found to be about 70%, whereas physical motor learning percentages are approximately 90%. Tasks that involve both cognitive and physical learning have learning percentages between 70% and 90% [10].

Even though the learning curve theory is a well-known method, the literature review reveals few studies have used this method in measuring operator performance in industries [11, 12]. This research uses the learning curve approach to investigate the effects of gender on the learning process (operator performance in terms of productivity and quality) during the initial adaption to the haptic controller in a simulated excavation task. The expected results are to benefit designs and worker training programs to accommodate both genders.

2. Methods

2.1. Subjects

Ten male and ten female subjects were recruited from the university community for this study. The average (and standard deviation) of age for male and female groups were 25.2 (2.6) years and 22.4 (1.7) years, respectively. The average (and standard deviation) of age for all subjects were 23.8 (2.6) years. Two subjects were left-handed and the rest were right-handed. All subjects had 20/20 corrected vision. None of them had discomfort or history of injury in the upper extremity, and all had given informed consent at the time of participation.

2.2. Apparatus

The main apparatus was the Phantom haptic device (SensAble, Woburn MA) (Figure 1). In addition, three computer workstations (Gateway, Irvine CA) were employed to construct, enable and display a simulated excavation work environment involving the haptic device. The first computer was equipped with Matlab and Simulink. Matlab code and Simulink models that simulated the mechanical dynamics of an excavator were composed and provided by the Department of Mechanical Engineering at Georgia Institute of Technology and Bobcat. The second computer was responsible for the communication and synchronization between the Matlab/Simulink output and the mechanics of the Phantom controller. A third computer rendered graphics and displayed the simulated task environment to the operator (Figure 1).

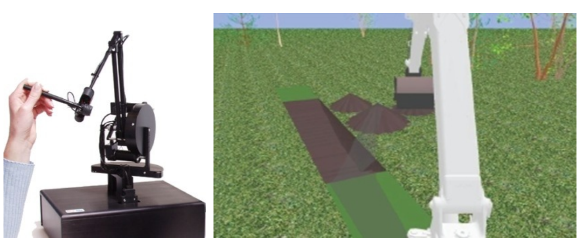


Figure 1. Simulated work environment for an excavation task using the Phanton device

2.3. Experimental design

Gender was the only independent variable for this study and had two levels: male and female. There were four dependent variables which included learning percentage, minimum error, median error, and maximum error. A between subject design was used to investigate the effects of gender. Within the same gender group, subject was a random factor.

2.4. Experimental procedures

Each subject was first introduced to the experimental procedure that had been approved by the Internal Review Board at the North Carolina Agricultural and Technical State University. If subjects agreed to participate in the experiment, they would provide informed consent in written form. Prior to an experiment, each subject went through a five to ten minutes' training session to get familiarized with the basic mechanics of the Phantom device. They received hands-on instructions on how the movement of the device corresponded to the motions of the segments of the excavator (i.e. the boom, stick and bucket). This training session lasted until the subject was comfortable with operating the device to achieve the basic motions of the boom, stick and bucket (although they were not allowed to practice the task in its entirety in order for them to remain fresh before the first task/trial). Following the training session, the trials began. Each subject was asked to perform the same simulated excavation task continuously for ten times, constituting ten trials. For each trial, the subjects needed to position the boom, stick and bucket to reach into a trench, gather, transport and deposit the soil as close to a pre-specified target as possible, and then return the bucket to the trench. The target location for soil deposit was explicitly shown on the screen and remained at the same spot throughout the ten trials. It was repeatedly emphasized to the subject that speed and accuracy were of the same importance. For each trial, time to completion and the deviation (in inches) of actual soil deposit location from the target were recorded.

2.5. Data processing

Recorded data were exported to EXCEL for processing. Since time was recorded continuously, the actual time for the completion of each trial was obtained by subtracting the timestamp at the end of the trial by the timestamp at the start of the trial. For each subject, Equations 1-3 were adapted as follows in calculating the learning percentage.

$$\text{Learning Percentage} = 100 \times 2^{\log_{10}\left(\frac{t_{10}}{t_1}\right)} \quad (4)$$

where t_1 and t_{10} are the completion times for the first and the tenth trials. Deviations of actual soil deposit location from the target were recorded as (x, y) coordinates (in inches). Therefore, the overall deviation was calculated as $\sqrt{x^2 + y^2}$. Of the ten trials, the minimum, median and maximum deviations were selected as the three error-related dependent variables.

2.6. Data analysis

Pearson correlation analyses were conducted to examine potential correlations between the learning percentage and the error measures, as well as within the error measures. An unpaired Hotelling's T^2 test was used to examine gender effects on the set of all four dependent variables. If Hotelling's T^2 yielded significant gender effect, an analysis of variance between groups would be used to investigate gender effects on individual dependent variables. Both analyses were carried out using SAS software (Cary, NC).

3. Results

Table 1 shows the learning percentages, median errors, minimum errors and maximum errors computed for each subject. These columns of data were used to perform the Pearson correlation analyses between learning percentages and the three error measures, as well as within the error measures. Tables 2-4 illustrate the results from the Pearson correlation analyses. Only those correlation coefficients with P<0.05 were shown.

Table 1. Subject-specific learning percentages and error measures (M: Male; F: Female; L.P.: Learning Percentage)

		L.P.	Error		
			Med	Min	Max
M	1	73.5	27.46	2.13	34.08
	2	77.4	8.80	1.67	25.82
	3	85.1	28.26	13.24	61.64
	4	63.3	28.97	27.03	53.12
	5	83.5	28.25	13.67	32.15
	6	92.8	13.12	3.05	28.07
	7	94.4	15.84	4.46	28.64
	8	86.3	25.48	9.66	52.59
	9	89.8	30.49	6.05	96.69
	10	85.3	14.64	6.40	28.11
F	1	80.0	18.14	2.75	29.71
	2	58.8	40.81	10.56	118.42
	3	98.5	8.43	1.22	27.42
	4	96.7	24.00	4.23	31.69
	5	72.9	31.34	14.33	45.15
	6	79.6	28.41	6.69	39.76
	7	92.6	23.60	4.79	89.29
	8	69.5	29.30	20.89	65.32
	9	91.7	27.28	24.28	116.81
	10	71.9	31.20	27.93	86.52

The results from the Hotelling's T^2 test showed no significant gender effect ($\alpha = 0.05$) when all four dependent variables were analyzed simultaneously. Statistics showed that Wilks' lambda=0.8786, F (4, 15) = 0.52, and p=0.7237. Where each dependent variable was analyzed separately, there was no significant gender effect (Table 5).

Table 2. Pearson correlation coefficients for all subjects

	Minimum Error	Median Error	Maximum Error
Learning Percentage	-0.5136	-0.5791	-
Minimum Error		0.5673	0.5066
Median Error			0.6678

Table 3. Pearson correlation coefficients for male subjects

	Minimum Error	Median Error	Maximum Error
Learning Percentage	-	-	-
Minimum Error		-	-
Median Error			0.6830

Table 4. Pearson correlation coefficients for female subjects

	Minimum Error	Median Error	Maximum Error
Learning Percentage	-	-0.7720	-
Minimum Error		-	-
Median Error			0.6319

Table 5. ANOVA results

	F(1,18)	Pr > F
Learning Percentage	0.16	0.6919
Minimum Error	0.60	0.4474
Median Error	1.21	0.2852
Maximum Error	2.47	0.1334

Figure 2 shows the averaged task completion times plotted by trial and gender. The averages (and standard deviations) of learning percentage for male and female subjects were 83.1% (9.4%) and 81.1% (13.2%), respectively (Figure 3). For all three error measures, male subjects had lower averages and standard deviations than female subjects (Figure 4). For minimum, median and maximum errors, the differences between genders in averages (and standard deviations) were 3.03” (1.90”), 4.12” (0.58”) and 20.92” (13.08”), respectively. (Note that in both Figures 3 and 4, the error bars represent standard errors.)

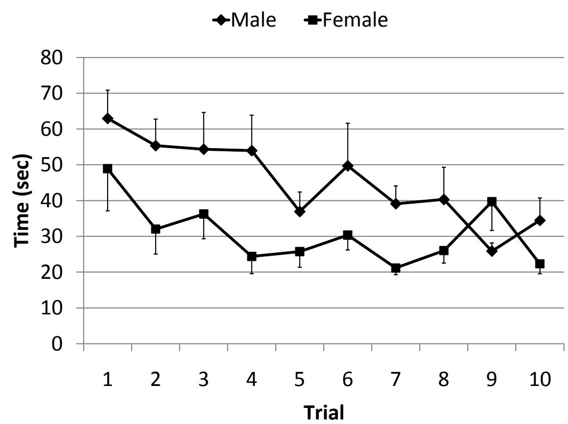


Figure 2. Averaged task completion times by trial and gender (The error bars represent *standard errors* where *n*=10.)

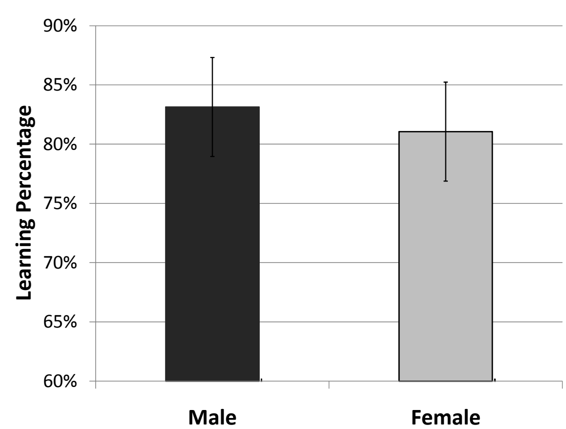


Figure 3. Learning percentage averages (The error bars represent *standard errors* where *n*=10.)

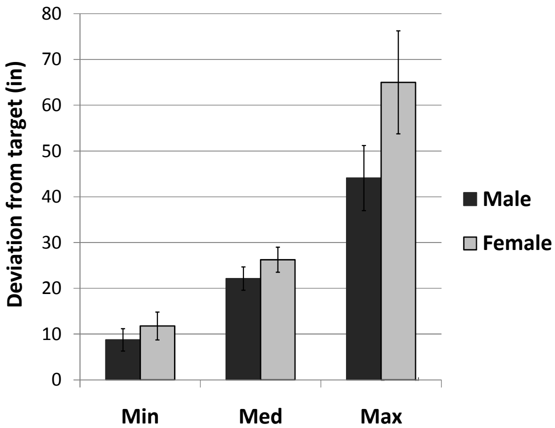


Figure 4. Minimum, median and maximum errors (The error bars represent *standard errors* where *n*=10.)

4. Discussion

In the construction industry, there has been a shortage in qualified workers within the traditional male work population. To battle this challenge, contractors have started to increase hiring within female population, resulting in an increasing number of female workers in the construction industry [1]. In the meantime, new technologies have been developed for implementation in construction machinery such as replacing hydraulic controllers with haptic controllers in an excavator. As an operator is introduced to a new piece of technology, negative impact on short-term productivity is usually inevitable. In this case, management in the construction industry may be concerned with how the new haptic technology may affect their workers especially the growing female work force. Quantification of this learning process can help assuage those concerns. The goal of this research is therefore to apply the learning curve theory to investigate the difference between genders in the learning process of adapting to a haptic controller in completing a simulated excavation task. The findings from this study suggested differences in the learning patterns between males and females.

Studies regarding gender and brain function revealed that male brains are generally in a more active state than female brains during working memory tasks [13]. Due to the stronger association between its spatial ability and the left brain hemisphere [6], men are generally better at tasks such as hitting targets and manipulating maps mentally. Supported by the literature, it was expected that male subjects would make greater progress (i.e. lower learning percentage) and smaller overall errors in the process of learning to use the haptic controller.

A Hotelling's T^2 test was conducted to examine gender effects on the learning percentage, minimum error, median error and maximum error. Contrary to what was expected, the inferential statistics results showed that there was no significant difference between male and female subject groups although descriptive results indicated less errors from male subjects (Figure 4). One of the reasons behind this may be the limited sample size within each gender (n=10). It is possible that if there was a much larger pool of subjects, statistical significance between genders would be detected especially in error related results. The Pearson correlation analysis showed that male subjects had a higher overall correlation level between the learning percentage and errors than female subjects. That means with slower learning progress (or higher learning percentage), males tend to make greater errors than females.

Some interesting findings were also obtained from observing subjects, listening to their feedback as well as further investigation on the data set. [10] found that the learning percentages for motor (physical) learning and cognitive learning are about 90% and 70% respectively, and tasks that involve both has a learning percentage between 70% and 90%. For the current study, the learning percentages for male and female subjects were both averaged at about 82% (Figure 3). This confirms that the task performed in this study employs both motor and cognitive learning. Over the course of the ten trials, there was an overall decrease in time (or increase in productivity) across all subjects. This confirmed the assumption behind the learning curve theory. However, errors were much more unpredictable and did not have a clear pattern to follow. It was observed from the raw data that near the fifth and sixth trials, male subjects tended to have smaller or more stable errors than female subjects. For most of the female subjects, error values tend to fluctuate throughout the entire experiment. Even though the results showed that there was not any significant difference between males and females, while observing the subjects, they shared some similar approaches while differ in others when completing the tasks. Both genders seemed to be more concerned with accuracy instead of speed although they were told to focus on both in equal measure. Male subjects appeared to assume a more competitive attitude than female subjects.

There were a number of limitations to this research that may challenge the generalizing ability of the results. First, subjects recruited for this study were in a rather homogenous age group (between 20 and 28 years). Therefore, cautions must be taken when attempting to generalize the results to individuals outside this age range. Secondly, because the coupling between the haptic feedback controller and the simulated excavation task was still in the developing stage at the time of the experiment, no force feedback was included in the simulation. This led to a complete focus on the "haptic" aspect of the controller instead of "feedback". Caution again needs to be taken when interpreting the results of this study. Thirdly, the simulated excavator seemed to be over-sensitive to the controller motions. This had made the device hard to control at times and may have caused extra variability in subject data (on both the learning percentage and errors). This in addition to the limited sample size may have caused the inability to reveal any significant difference between male and female subjects.

5. Conclusions

Finding an effective approach to work place training is important, and more so when both male and female workers are present. This experiment provides evidence that females are able to make similar progress to males in terms of productivity when trying to grasp and master training concepts involving a haptic controller in completing a simulated excavation task, although females are likely to make greater errors than males. Errors made by females are less likely to correlate with the progress made in productivity. Although gender was the only independent variable considered in this research, it should not be considered the only factor that can lead to different learning characteristics regarding haptic controllers with certain tasks. In the future, some other factors such as age, experience, and motor skills may be explored for potential effects on user learning patterns associated with haptic technologies. In addition, it would be interesting to compare the learning patterns on haptic controllers with those associated with different types of traditional joysticks and controllers.

6. References

[1] C.T. Jahren, "Transportation Construction Equipment", Committee on Construction Equipment, Chairman: Thomas B. Nelson, University of Wisconsin at Platteville, 2006.

[2] J. Broeren, M. Geogsson, M. Rydmark, and K.S. Sunnerhagen, "Virtual Reality and Associated Technologies", The Fourth

International Conference on Disability, Veszprém, Hungary, September 2002.

[3] H. Sveistrup, "Motor rehabilitation using virtual reality", Journal of Neuroengineering and Rehabiliation, 2004, pp. 1-10.

[4] L. Marchal-Crespo, "Haptic Guidance for Enhancing Human Motor Learning: Application to a Robot Assisted Powered Wheelchair Trainer", University of California, 2009.

[5] L. Lim, "Haptic and multi-modal interaction for teaching and designing basic controls", doctoral dissertation, 2004.

[6] K. Peek, "Gender differences in the brain", University of Texas Houston Medical School, 2001.

[7] Sutter Health System, "Gender differences found in brain anatomy", Tarleton State University, 2001.

[8] N. Chadwick, "Gender Differences" (www.webedlic.com/church/gendiff.htm), 2001.

[9] D. Eby, "Math and Gender differences", (www.talentdevelop.com/math.html), 2001.

[10] E.M. Dar-El, K. Ayas, and I. Gilad, "Predicting performance times for long cycle time tasks", IIE Transcations, 27, 1995, pp. 272-281.

[11] A.M. Anderson, G.A. Mirka, S.M.B. Joines, and D.B. Kaber, "Analysis of alternative keyboards using learning curves", Human Factors, 51(1), 2009, pp. 35-45.

[12] S.A. Reid and G.A. Mirka, "Learning curve analysis of a patient lift-assist device", Applied Ergonomics, 38(6), 2007, pp. 765-771.

[13] E. Bell, S. Dave, M. Willson, C. Morgan, and A. Wilman, "Males and Females differ in Brain Activation during Cognitive Tasks", NeuroImage, 30, 2005, pp. 529-538.

Content Validation of an Assessment Tool of an Individual's Perception of Lean Implementation

Sushil K. Shetty, Paul J. Componation, Sampson E. Gholston
ISEEM Department, The University of Alabama in Huntsville, Huntsville, AL 35899, USA
shettys@uah.edu, Paul.Componation@uah.edu, gholston@ise.uah.edu

Abstract

Lean is an element of many successful process improvement efforts. Reasons for its use include ease of application, quick turnaround in operation, and success in past applications. The literature documents Lean applications in manufacturing, service, and government organizations. One common problem however is how to measure effectiveness of Lean implementation in an organization. Usually an organization works with experts who provide an external perspective of the organization's Lean implementation. The approach has merit but can also result in significant time and resource requirements. Also, they may not have sufficient expertise in the specific industry to assess implementation from the employee perspective. This paper shows the first phase: content validation, in an alternative approach which is based on assessing Lean as seen by the employee. The approach centers on a tool built on the 5 principles of Lean Thinking: Define Value, Identify the Value Stream, Make Value Flow, Pull Value, and Pursue Perfection. Questions were created for each principle to describe and quantify its impact on the Lean implementation. This phase of research was conducted by analyzing qualitatively the feedback from 7 Lean experts in the field from all over the United States. There was agreement on most aspects, with some ambiguity on the need to include culture as an additional principle. The content validation was successfully completed and the resulting survey is ready for testing with Lean practitioners.

1. Introduction

Ever since the early 90's when Lean came to prominence with the publishing of the book "Machine That Changed the World", manufacturing has seen improvement in both quality and efficiency. The book has laid out studies that had been compiled from 5 years of industry research that showed ways and approaches that could be used to eliminate waste and improve processes. Many studies have followed since, but there has been less research done on how to measure Lean implementation progress. The few tools that are available are commonly used by a third-party agency that is involved in the training as well. This has led relatively few broad-based quantitative studies on Lean implementation. None of these studies has focused on a Lean assessment tool that is simple and inexpensive to use. One of the first steps in developing this tool is to identify the key principles of Lean and to develop a measurement strategy. This paper shows the first progress in this area, the content validation of the Lean principles used in the tool.

2. Theory Base for Research

The concept is simple: removal of waste from a system. In this context, waste is any non-value-added activity. Value is defined as something the customer is ready to pay for. Therefore, if the customer is not willing to pay for something, it is a non-value-added activity. Today, Lean is utilized in a wide range of industries, nonprofit organizations, government agencies, healthcare, and organizations in other areas as a means for producing goods and delivering services that create value for the customer with a minimum amount of waste and a maximum degree of quality [8]. In a traditional business approach the customers are treated as entities at the bottom of the supply chain. They would have nothing to do with the product until it actually was on the shelf. With Lean the customers are treated as part of the process. Their needs are an important part of the product development process.

The principles of Lean have been around for over 100 years. Henry Ford's concept of the assembly line was considered to be the first advent of Lean [16]. Later the Toyota Production System (TPS) laid the foundation for the modern principles of form of Lean. Toyota's development of Lean may have started at the turn of the 20th century with Sakichi Toyoda, in a textile factory with looms that stopped themselves when a thread broke; this became the seed of automation, or Jidoka which means automation with a human touch. Toyota's first use of Lean can be traced back to 1934 when it moved from textiles to produce its first car. Kiichiro Toyoda, founder of Toyota, directed the engine casting work and discovered many problems in their manufacture. He decided he must stop the repairing of poor quality by intense study of each stage of the process. In 1936, when Toyota won its first truck contract with the Japanese government, his processes hit new problems, so he developed the "Kaizen" improvement teams. Levels of demand in the postwar economy of Japan were low, so an emphasis on economics of scale was not applicable. Having visited and seen supermarkets in the USA, Taiichi Ohno recognized that scheduling of work should not be driven by production targets but by actual sales. Overproduction had to be avoided and thus the notion of Pull (build to order) rather than Push (build to market) came to underpin Toyota production scheduling. It was Taiichi Ohno at Toyota that put these themes together. He built on the already existing internal schools of thought and spread their breadth and use into what has now become TPS. Lean incorporates and extends TPS. Norman Bodek wrote the following in his foreword to a reprint of Ford's *Today and Tomorrow:* "I was first introduced to the concepts of Just-In-Time (JIT) and the TPS in 1980. Subsequently I had the opportunity to witness its actual application at Toyota on one of our numerous Japanese study missions. There I met Mr. Taiichi Ohno, the system's creator. When bombarded with questions from our group on what inspired his thinking, he just laughed and said he learned it all from Henry Ford's book [15]". TPS builds on Henry Ford's work by adding the scale, rigor and continuous learning aspects.

This study will be based on prior research on Lean to develop a Lean assessment tool. The principles will each have a list of questions which, by a self-assessment, would help in the measure of current state of Lean implementation in the organization. These questions will be based on theories from various researchers of Lean.

3. Summary of Prior Research

The review of prior research looked at both seminal works in Lean and identified the most commonly used models for assessing Lean implementations. The basis of Lean is often traced back to the original work done by Dr. Shigeo Shingo. He was one of the early contributors to TPS. Lean's current popularity is often attributed to the work done by Drs. James Womack, Daniel Jones and Daniel Roos. These works were used as the basis for the development of Lean implementation models that are commonly used today. This summary also notes other recent works that were reviewed but not used extensively in this research.

One of the most significant contributions to the development of Lean is the Shingo Prize developed by Utah State University in 1988. The Shingo Prize Model is based on the Lean management approach by Dr. Shigeo Shingo, a lean management consultant and considered to be one of the first proponents and contributors to the TPS. His model describes three levels of business improvement, which he calls levels of transformation: Principles; Systems; Tools and Techniques. This model emphasis that true innovation is not accomplished when Systems, and Tools and Techniques ('know how') are used alone. Rather, its effect is seen when the underlying principles ('know whys) are understood. Shingo proposed that the Tools and Techniques are the foundation on which all Lean transformation must be built, but further suggested that Systems level followed by an understanding of the Principles would help cement this philosophy and the lean ideology in a person's mind. This enables a deeper and more permanent lean transformation. According to the model, Lean transformation is a journey that is composed of 4 dimensions: Cultural Enablers, Continuous Process Improvement, and Consistent Lean Enterprise Culture and Business results. These dimensions overlay 5 business processes, namely, Product/Service development, Customer Relations, Operations, Supply and Management. These categories were designed to encompass all organizational activities, regardless of industry. There are weights given to each dimension in term of points. The basis for awarding of these points is unknown [9].

Lean today is based on the 90's book of by Drs. James Womack, Daniel Jones and Daniel Roos, "Machine that Changed the World". This book laid the foundation for Lean philosophy. The authors proposed the tenets of Lean thinking which formed the basis Lean research in the years to come. They were the first to talk about value to the customer and make value the goal around which everything was organized. This concept gained notice in the early 21st century when competition increased between organizations. Womack, Jones and Roos' book builds a case for Lean by showing the differences between the Japanese and American approach to automotive manufacturing. Their research talks about the differences between mass and lean production: Leadership, teamwork, communication, and continuous development. It was a qualitative study that was a result of the 5 years of observation and one of the first detailed books about lean principles in the western world [1].

In 1996, Drs. James Womack and Daniel Jones wrote a second book "Lean Thinking: Banish Waste and Create Wealth in your Corporation". It defined the term value as the foundation of Lean. This book defined the key principles of Lean: Define/Specify Value, Identify Value Stream, Make Value Flow, Pull Value and Pursue Perfection. These 5 principles are repeated continuously to achieve sustained perfection. This could be done at various levels of the organizations and should be followed by everyone in the organization. Womack and Jones define Lean in terms of value which is the starting point for every continuous improvement activity. They describe value as something the customer must value for a Lean implementation to be successful. The authors note small and international success stories using Lean. This supports the belief that Lean is not restricted to a particular culture or region but can be widely accepted and implemented successfully all around the world [11].

The Massachusetts Institute of Technology (MIT) Lean Aerospace Initiative (LAI) developed a Lean Implementation model known as the Lean Self-Assessment Tool (LESAT). LESAT assesses 54 Lean practices, so a detailed knowledge of Lean and the tool is required before the assessment is done. The tool was divided into 3 sections; Transformation/ Leadership, Lifecycle Processes, and Enabling Infrastructure. Each section is defined and certain enterprise characteristics are laid out in order to measure the construct or factor [10].

Another model used for Lean assessment was created in 2006 by the consulting firm Strategos. Their model measures 9 areas of manufacturing: inventory, team approach, processes, maintenance, layout and material handling, suppliers, setup quality and production control and scheduling. The model has 3 to 6 multiple choice questions per area. They mentioned that all the 9 areas are not of equal importance to Lean and mention using personal judgment in deciding what weighting criteria to use. Strategos' model uses an outside-in perspective, wherein a consultant makes an assessment of the Lean implementation at the workplace [17].

General Motors (GM) created a Lean assessment for their Supplier Global Manufacturing Systems (GMS) in 2006. It was aimed at assessing supplier status in implementing Lean principles and supplier's ability to identify and eliminate waste. It is meant to give a broad understanding of the supplier's Lean implementation culture. The GM Lean assessment looks at 5 principles: People Involvement, Standardization, Built-in Quality, Short Lead Time and Continuous Improvement. The assessment includes 71-questions and is an outside-in approach to Lean implementation assessment [18].

Karlsson and Ahlstrom in 1996 developed a model for Lean which formed a basis for the attributes of their Lean teachings. Their research was conducted over a 2½ year period with 3-4 days a week of observation done to collect data from each clinical field studies. They developed 7 attributes to assess lean production and also the determinants that define these attributes. These determinants were reported as increasing or decreasing in the study. Quantitative data was not collected. They did propose that quantitative measures for each determinant could be established to help put numbers behind each of the attributes. Further, they also made a significant point, that Lean is not a state but a journey; a journey of change [2].

There are several other models that have been proposed, including the Rapid Plant Assessment (RPA) Tool by Dr. Eugene Goodson in 2002. RPA uses 20 questions broken into 11 categories. Dr. Goodson emphasizes how to select a team and give the required necessary to build a team to make this assessment [3]. Dr. Hung-da Wan's also developed a model measuring Leanness using the DEA (Data Envelopment Analysis) technique, wherein he uses this operations research method to help develop a leanness frontier against which one can plot the current state of Leanness of the organization. A Slacks Based Model was used to develop a method to measure the current Lean state and the direction of potential improvement [5]. A study by Allada and Srinivasaraghavan in 2006 concluded that Lean implementation could be measured by finding the distance of the various variables in the model and comparing the standard data to the abnormal data from the information collected. There approach uses 5 attributes chosen from the LESAT. Further, it takes into consideration interrelationship of variables in the model by using the MTGS (Mahalanobis Taguchi Gram-Schmidt) system in the model [4]. Next, Achanga et al in 2006 looked at data from 10 small to medium sized enterprises and 3 large enterprises using observations and semi-structured interviews. They spent approx. 30-min at each particular point of observation [6].This was a very small sample and many assumptions were made regarding Lean implementation.

4. Content Validity

One key step in developing any survey tool is content validity. Content validity is a specific type of reliability study that verifies the veracity of a particular data set. By doing reliability studies we can find out whether the tool that is being used has questions that are asked in order to understand a particular principle, feature or category. Content validity is based on the extent to which a measurement reflects the specific intended domain of content [12]. In this research questions were created that answered the 5 principles of Lean: Specify Value, Identify Value Stream, Make Value Flow, Pull Value, and Pursue Perfection. Content validity was performed so that we achieve a practicality for the survey based on expert opinion. Also, content validity is a non-statistical type of validity that involves "the systematic examination of the test content to determine whether it covers a representative sample of the behavior domain to be measured" [13]. Content validity evidence involves the degree to which the content of the test matches a content domain associated with the construct. Content-related evidence typically involves subject matter experts (SME's) evaluating test items against the test specifications. A test has content validity built into it by careful selection of which items to include [13]. Items should be chosen so that they comply with the test specification which is drawn up through a thorough examination of the subject domain. By using a panel of experts to review the test specifications and the selection of items, the content validity of a test can be improved. The experts will be able to review the items and comment on whether the items cover a representative sample of the behavior domain [14].

Why was employee perception considered? The implementation of Lean principles has traditionally been done by external facilitators who are independent from the organization. The analogy behind that was that organizations did not have to invest directly in a training department:

- The outside-in approach was considered to be more balanced and independent.
- It requires less overhead for the organization.
- The assessment of the training effectiveness is commonly viewed from a management perspective.

But Lean must be implemented at all levels of an organization to be effective. By looking at perception of the employee, the organization will have a more complete view of the Lean implementation effectiveness. Thus creating a tool that enables this would be extremely useful for most organizations intending to implement Lean. By using the same agency to assess the implementation that they help create would in the true sense actually bias the results of the assessment. To avoid this, enabling an inside-out approach helps remove this bias and create a balanced assessment of the training undertaken by the external agency. By understanding the employee perception, one can understand how much the workforce has actually understood the training procedures and how beneficial they perceive this training for their day-to-day activities.

There is a wealth of related research is lean; however these noted works were found to be the most critical to developing this effort. The effort here is creation of a survey tool to assess Lean implementation using employee perception, thus bring the shop floor perspective to the matter.

5. Methodology

Creating a survey tool would not only help in the measurement of Lean, but also inform us of the lack of it with respect to certain attributes. This method would entail, to first research well tested and proven theory which has Lean divided in categories or principles, then using those principles as the basis for the survey tool creation. This would be followed by creating a set of questions which could be realized and understood by every person responding to the survey, be it a front-line employee or a CEO. As part of the content validation process, these questions were reviewed by a panel of 7 experts in the field of Lean. Experts from both academic and industry were used in the review to render a balanced content validation. The review was a blind process, wherein the identities of the experts were not revealed to each other. The subject matter experts were drawn from LAI Ed Net, Illinois Manufacturing Extension Center (IMEC), National Institute of Standards and Technology (NIST) and Alabama Technology Network (ATN).

These experts were asked to provide feedback regarding the questions within each Lean principle. They were encouraged to provide more descriptive feedback regarding the questions reviewed. The questions asked were;

- Is every question grammatically correct?
- Is every question required or not?
- Is every question in the correct category?
- If not, does it need to be moved/ dropped?

The objective of the research was highlighted throughout the process, which was to have a balanced survey being both complete and succinct. Sufficient number of questions was provided within each category to give the experts a variety of to choose from in the event of deletion of any, if required. The file was created in Excel and distributed electronically. A Likert scale was used for responses to the questions. The Likert scale provides the flexibility of responses to be more than just ‘yes’ and ‘no’ type answers. This provides a higher degree of fidelity since not all questions render a ‘yes’ and ‘no’ response.

Additional comments regarding each question and category were also requested. The experts were also asked to comment on the section of demographic questions asked at the end of the survey and comment on the relevance of each.

6. Results

Once the reviews were received, they were tabulated and analyzed. The analysis was performed by the panel that helped create the tool. If any question was disapproved by more than half of the reviewers it was changed or removed. Changes in wording were done to improve clarity and flow. Attention was also given to which Lean principle the questions supported. In some cases questions were moved if they better supported a different principle. This was a lengthy process as it involved comments that had to be analyzed and considered regarding different scenarios with respect to each question. All the responses were tabulated in Excel, which needed scrutiny to be analyzed in its entirety. Qualitative comments helped render a good balance to just yes and no type responses. This enabled clarity and relevance to each question in terms of which category they were placed in. The final survey was also shortened to make it easier to administer (Table 1).

Category	Initial Survey	Validated Survey
Define/Specify Value	9	7
Identify Value Stream	8	5
Make Value Flow	10	5
Pull Value	8	5
Pursue Perfection	8	7
Total	41	29

Table 1: Results of content validation phase

The above table tells us that within each category there was a reduction in questions and in 2 cases questions were relocated or moved to other categories. The 5 categories, as described earlier, are based on Womack and Jones's 5 principles of Lean. The experts rendered responses that were extremely informational and insightful. Primarily there was agreement; however, having criteria for judging the responses was important because some disagreements did surface. Some experts recommended the creation of new categories such as, Problem Solving and Culture. A decision was made to stay with Womack and Jones's principles of Lean Thinking. While these other categories were seen to be of value, the research into their relationships in Lean was not as well established. The additional categories were held for further research at a later date. Once the analysis was complete a finalized survey of 29 questions along with some demographic requests for were converted to an online survey.

An example of the review process can be seen below in Table 2. For question 4 of the 2nd category (C24), 'Identify the value stream', the table shows the original question from the survey, 'Implementing any corrective action increases customer value', which was submitted to the experts and their feedback regarding that question was assessed. Based on that feedback, using the methodology explained earlier, the finalized question, 'Corrective actions are carefully evaluated in relation to customer value', was arrived at. Each expert gave his/her opinion regarding the question with respect to the particular category in which it was originally located, and on that basis they evaluated it by the criteria given. Expert 1 opines that it was a confusing question. Expert 2 approves the question but believes that it should be in a category of its own, 'problem solving'. Experts 3 and 4 approve the question. Experts 5 and 7 approve the question but suggest rewording it. After review of all recommendations the final question was written. The 41-question survey was evaluated in this similar manner and the finalized assessment tool of 29-question survey was finalized. The finalized survey can be seen in Appendix-1. The survey provides the initial instructions and the 29 questions within the 5 categories as mentioned in the results. A Likert scale was used gather the responses. A short demographic section follows the 29 questions survey. Feedback regarding this section was also implemented. Some experts suggested that factors such as culture and problem solving needed to be considered as there was growing awareness of their importance in the Lean implementation process. Those comments were considered to be extremely helpful modifications of the questions but the scope of the research did not allow for the research in new factors, as basis of the Womack Jones model was used as foundation for the questions created. Few comments regarding specific tools, like A3 reports and others, were recommended to be added to the survey. At this time, they were not added because they tools were not always referred to using the same terminology in all organizations. Care was taken not to add recent non-universally accepted Lean elements as that would not be a fair method to evaluate every Lean workplace.

7. Conclusion

Development of a reliable and validated survey tool is not a trivial task. The survey must be built on a sound theoretical basis and show statistical validity. In application it must also show accuracy in measurement. This research strives to create a tool that would help assess employee perception of Lean implementation. The content validation phase is extremely important in realizing whether or not the tool has enough qualitative support and renders meaning to this effort. The basis for this tool helps garner a good foundation in order to be widely accepted. The expert review enabled a wide range of inputs from both academicians and practitioners.

Original Question	Implementing any corrective action increases customer value
Expert 1	Confusing question- N
Expert 2	Y- Problem solving
Expert 3	Y
Expert 4	Y
Expert 5	Corrective are carefully evaluated in relation to customer value
Expert 6	Reword: Corrective actions are implemented based on customer feedback- repetitive from previous section question -Y
Expert 7	I might revise to read: Does implementing corrective actions….
Revised Question	Corrective actions are carefully evaluated in relation to customer value

Table 2: An example of review of question C24

The phase to follow will be that of construct validation, in which Lean practitioners would respond to the online survey. This will be followed by statistical analysis which will determine from a statistical view point whether the questions are loaded correctly and if the 5 categories be justified. Once that is complete, the next step is criterion validation, which would be conducted by asking employees within a company to respond to the survey and then comparing the results to a separate survey completed by a Lean consultant or the Lean facilitator within that organization. 10 to 15 such organizations would be surveyed and results analyzed to see if there is any correlation between how the employees' perceive Lean implementation as compared to a facilitator or consultant. And, study the relation between how the practitioners perceive Lean implementation compared to front line employees. The results from these studies will be important to companies who want to realize if Lean implementation is been understood at a grass-root level, and in turn have a reliable tool which could be administered with relative ease and less financial burden.

8. References

[1] J.P. Womack, D. T. Jones and D. Roos; The Machine that Changed the World. New York: Rawson Associates, 1990.
[2] Christer Karlsson and Par Ahlstrom; "Assessing Changes Towards Lean Production", International Journal of Operations & Production Management, Vol.16, No.2, pp 24-41, 1996.
[3] Eugene R. Goodson; "Read a Plant –Fast", Harvard Business Review, Harvard Business School Publishing Corporation, 2002.
[4] Venkat Allada and Jayanth Srinivasaraghavan; "Application of Mahalanobis Distance as a Lean Assessment Metric", International Journal for Advanced Manufacturing and Technology, Vol.29 pp 1159-1168, 2006.
[5] Hung-da Wan; "Measuring Leanness of Manufacturing Systems and Identifying Leanness Target by Considering Agility". PhD dissertation, Virginia Polytechnic Institute and State University, 2006.
[6] Pius Achanga, Esam Shehab, Rajkumar Roy and Geoff Nelder; "Critical Success Factors for Lean Implementation within SMEs", Journal of Manufacturing Technology Management, Volume 17 No.4, pp 460-471, 2006.
[7] Lean Manufacturing History, Origins and History lean Manufacturing, http://www.strategosinc.com/just_in_time.htm
[8] Lean Enterprise Institute, A Brief history of Lean; http://www.lean.org/WhatsLean/History.cfm
[9] The Shingo Model; http://www.shingoprize.org/
[10] The Lean Aerospace Initiative and LESAT; http://lean.mit.edu/
[11] James Womack and Daniel Jones; *"Lean Thinking: Banish Waste and create wealth in your corporation"*, Free Press, New York, 1996.
[12] Carmines, E. G., & Zeller, R.A; *Reliability and validity assessment*. Newbury Park: Sage Publications, 1991.
[13] A. Anastasi and S. Urbina; *Psychological testing* (7th ed.), Prentice Hall, Upper Saddle River, NJ (1997).

[14] Foxcroft, C, Paterson, H, Le Roux, N, & Herbst, D; *Psychological assessment in South Africa: A needs analysis. The test use patterns and needs of psychological assessment practitioner,* 2004

[15] Ohno, Taiichi; *Toyota Production System: Beyond Large Scale Production*. Productivity Press, 1988.

[16] Ford, Henry & Crowther, Samuel; *Today and Tomorrow*, DoubleDay Page and Company, Garden City, New York, 1926

[17] Strategos Inc. *Assessment of Lean Manufacturing*; http://www.strategosinc.com/assessment.htm, 2006

[18] GM Supplier GMS; *Supplier Global Manufacturing System Lean Assessment.* 2006.

APPENDIX-1

Assessing an Individual's Perception of Lean Implementation

INSTRUCTIONS to Survey Respondents:
The person taking this survey should have been involved individually or as part of a work group in implementing lean in the organization for a period of at least 3 months. You should answer this survey based on your observations and perception of Lean in your organization. You should answer each question to the best of your knowledge, and in the case of not understanding or questioning the applicability of any particular question, you should answer "I don't know".
All respondents to this survey SHOULD be 20 years of age or older.

	DEFINE/ SPECIFY VALUE-- Described as, 'Specify value from the standpoint of the end customer by product family'. The importance is to understand the differencebetween value added and non-value added activities in each process.	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	I don't know
1	My group understands the difference between value added and non-value added activities						
2	My group is involved in the development of process metrics						
3	Quaility requirements are derived from the customer						
4	Process standards are used throughout the organization at each workstation/ workplace						
5	My group understands how the customer uses the product /service						
6	Customer requirements are clearly communicated at each stage in the journey of the product/ service; from concept to delivery to the customer						
7	A system to communicate customer feedback throughout the organization is present						

	IDENTIFY THE VALUE STREAM-- Described as, ' Identify all the steps in the value stream for each product family, eliminating whenever possible those steps that do not create value'.	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	I don't know
1	Corrective actions are carefully evaluated in relation to customer value						
2	Root Cause Analysis Tools like Pareto Analysis and Fishbone Diagrams are used to identify value in processes						

		Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	I don't know
3	Value stream mapping is undertaken to understand how activities create value for the customer						
4	My group is involved in defining value added steps for the product						
5	Improvements from the value stream mapping process are implemented as planned						
	<u>MAKE VALUE FLOW</u>-- Described as,'Make the value-creating steps occur in tight sequence so the product will flow smoothly toward the customer'.	**Strongly Disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**	**I don't know**
1	Training programs are an integral part of an employee's development at the workplace						
2	Managers encourage proactiveness from their employees						
3	Work in Process between workstations is limited and actively minimized						
4	Batch sizes have been aggresively reduced						
5	The technique 5S is used to organize workstations						
	<u>PULL VALUE</u>-- Described as,'As flow is introduced, let customers pull value from the next upstream activity'.	**Strongly Disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**	**I don't know**
1	Original contracted deadlines are met for every shipment						
2	A scheduling system decides how much is produced at each workstation						
3	The 'Just In Time' concept is a part of daily routine at the workplace						
4	Finished products are shipped immediately to the customer						
5	A Kanban system is used for product flow						

	PURSUE PERFECTION-- Described as,'As value is specified, value streams are identified, wasted steps are removed, and flow and pull are introduced, begin the process again and continue it until a state of perfection is reached in which perfect value is created with no waste'.	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	I don't know
1	My group is given authority to make decisions related to their work						
2	We are always making small improvements in our process						
3	We follow up to make sure improvements continue						
4	Procedures are updated when changes are implemented						
5	Management demonstrate dedication and active commitment to initiatives undertaken						
6	Failures are seen as an opportunity for improvement						
7	I try to improve my work each day						

Tell me about you:

Gender: Male _______ Female________

Age: Below 25 _____ 25-35 _____ 36-45 _____ 46-55 _____ 55+ _______

# of Kaizen/ Lean Events Participated in: ____

Level in teh Organization: Management _______ Supervisors ________ Front Line Workers ________

Type of Organaization: IT______, Manufacturing________, R & D_________ Other (name)__________

Organization :Government _________ Private__________

Company Size: Microenterprise_________Small ______, Medium ________ Larger ________
(# of Emplyees: < 10 – Microenterprise, < 50 -- Small , < 300 --- Medium and >300 — Large)

An Inverse Method for Determining and Controlling Inaccessible Domain Temperature of Hot Bodies

Hossein Rastgoftar and *Faissal A. Moslehy
Department of Mechanical, Materials & Aerospace Engineering, University of Central Florida, Orlando, FL 32816, USA
Corresponding author: fmoslehy@ucf.edu

Abstract

The paper discusses the important inverse heat conduction problem and presents an elegant analytical method for control of temperature distribution as demonstrated in cylindrical bodies. The inverse heat conduction problem has been widely used in different practical engineering problems such as estimation of surface conditions, initial conditions and thermal properties of a body from known information at some predefined internal positions. This is an important engineering application, especially when the boundary of the region of interest is inaccessible and/or the determination of its temperature is desired. Also, when the boundary conditions are known or measured and domain temperature needed to be controlled. To obtain the solution of this inverse problem, appropriate heat fluxes should be exerted on the system. This inverse problem needs a control technique be applied at the boundary. To control temperature distribution, Lyapunov method is used and boundary heat fluxes are assigned such that the time derivative of the Lyapunov functional becomes negative definite; desired temperature distribution is achieved. It is assumed that the desired temperature distribution is admissible, i.e., satisfies the steady-state conduction heat equation. Unlike common control methods the introduced method is exact and Partial Differential Equation (PDE)-based, thus it is not necessary to utilize numerical approach to discritize the governing PDE. It is assumed that the properties of the considered body, such as density, thermal conductivity and specific heat are spatially varying parameters.

1. Introduction

Forward and inverse problems are common approaches for analyzing thermal systems behavior. In the forward heat transfer problem, the distributed temperature distribution is known and the aim is to obtain heat fluxes (see Figure 1) On the other hand, in the inverse heat transfer problem, the known heat fluxes are utilized to obtain the unknown temperature distribution (see Figure 2). This is similar to the control problem in heat transfer, where the control inputs are heat fluxes and the outputs are temperature distributions.

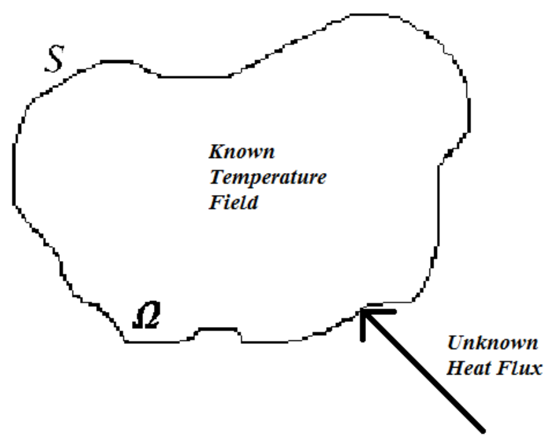


Figure 1. A Schematic of the forward heat transfer problem

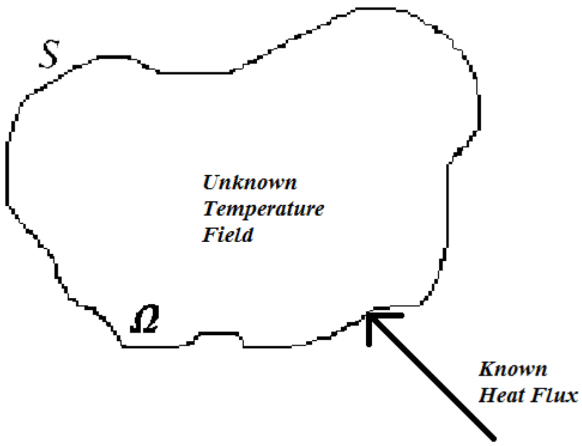


Figure 2. A Schematic of the inverse heat transfer problem

In reviewing the literature, it is evident that the researchers focus more on heat transfer analysis rather that control. The common methods they used are numerical [8] such as finite element analysis [1-4], boundary integral equation method [5], boundary element method [6], and Petrov-Galerkin method [7]. However, the work on the control of temperature distribution is almost rare [9]. In reference [8], temperature control of functionally graded material (FGM) plates using feed forward-feedback controller based on the inverse solution is presented. Also, an analytical approach for control of temperature distribution in FGM rectangular plate is presented in [9].

The control of distributed parameters systems (DPS) has found numerous applications in the past two decades. For example, control of flattering of plane wing has an importance in having a smooth and stable flight. An example of thermal application is the need for producing a desired temperature distribution in semiconductors and uniform temperature distribution in glass industries. It is known that vibration and heat transfer problems of DPS result in partial differential equations that are usually function of time and spatial variables. For example, the vibration equation of a general surface is modeled by a fourth order PDE, which its dependent variables are functions of two spatial variables and time. And the heat diffusion problem is generally modeled by a second order PDE, which considers the temperature as the dependent variable and time and one, two or three spatial variables as the independent variables.

In this paper an analytical approach for control of temperature distribution in a cylindrical shell is developed. While previous methods of control of DPS commonly rely on discritizing PDE to a set of ordinary differential equations (ODE), our approach is an exact PDE-based solution. Here we use Lyapunov theorem and introduce a positive-definite Lyapunov functional, then assign the heat fluxes such that the time derivative of the Lyapunov reverts to a negative-definite value.

We begin by presenting the governing equations of heat conduction in thin cylindrical shells. Then the control approach is introduced and control heat fluxes are assigned such that the transient temperature distribution approaches the desired one. Finally, we present the numerical analysis of this approach, with validating results.

2. The Heat Diffusion Equation in Cylindrical Coordinates

Based on Fourier law of heat conduction, the heat flux vector in cylindrical coordinate system is

$$\boldsymbol{q} = q_r\boldsymbol{e_r} + q_\theta\boldsymbol{e_\theta} + q_z\boldsymbol{e_z} = -\boldsymbol{K}.\nabla T = -\boldsymbol{K}\left(\frac{\partial T}{\partial r}\boldsymbol{e_r} + \frac{1}{r}\frac{\partial T}{\partial \theta}\boldsymbol{e_\theta} + \frac{\partial T}{\partial z}\boldsymbol{e_z}\right) \quad (1)$$

where $\boldsymbol{K}$ is the second order conductivity tensor , and $T = T(r,\theta,z)$ is the temperature field. Mathematically, $\boldsymbol{K}$ can be a non-diagonal tensor but it is diagonal for physical applications. Hence,

$$\boldsymbol{K} = \begin{bmatrix} k_r(r,\theta,z) & 0 & 0 \\ 0 & k_\theta(r,\theta,z) & 0 \\ 0 & 0 & k_z(r,\theta,z) \end{bmatrix} \quad (2)$$

where k_r, k_θ and k_z are the radial, polar and longitude conductivities of the system in cylindrical coordinates.

The heat diffusion equation in cylindrical coordinates

$$\frac{1}{r}\frac{\partial}{\partial r}\left(k_r\frac{\partial T}{\partial r}\right)+\frac{1}{r^2}\frac{\partial}{\partial \theta}\left(k_\theta\frac{\partial T}{\partial \theta}\right)+\frac{\partial}{\partial z}\left(k_z\frac{\partial T}{\partial z}\right)=\rho C\frac{\partial T}{\partial t} \qquad (3)$$

where ρ is the mass density and C is the specific heat of the thermal system represented in cylindrical coordinate system.

Since the aim is to analyze thin cylindrical shells, the variation of temperature field in the radial direction is considered zero, and the heat conduction equation reduces to

$$\frac{1}{r^2}\frac{\partial}{\partial \theta}\left(k_\theta\frac{\partial T}{\partial \theta}\right)+\frac{\partial}{\partial z}\left(k_z\frac{\partial T}{\partial z}\right)=\rho C\frac{\partial T}{\partial t} \qquad (4)$$

A schematic for considered thin cylindrical shell is depicted in the Figure 3.

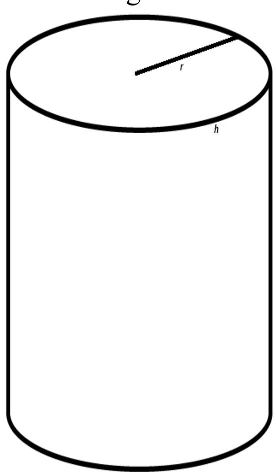


Figure 3. A schematic of the cylindrical shell

It is noted that any arbitrary distributed function of θ and z that satisfies the steady-state heat equation can be a candidate for the desired temperature distribution. Since the desired temperature represents the final distribution that we seek, it must be independent of time. Hence, the desired temperature distribution must satisfy following steady state heat equation:

$$\frac{1}{r^2}\frac{\partial}{\partial \theta}\left(k_\theta\frac{\partial T_d}{\partial \theta}\right)+\frac{\partial}{\partial z}\left(k_z\frac{\partial T_d}{\partial z}\right)=0 \qquad (5)$$

The difference between the desired and transient temperature distributions, $E(r,\theta,t)=T(r,\theta,t)-T_d(r,\theta)$, is considered as the error. Subtracting equation (5) from equation (4), the error dynamic is obtained as

$$\frac{1}{r^2}\frac{\partial}{\partial \theta}\left(k_\theta\frac{\partial E}{\partial \theta}\right)+\frac{\partial}{\partial z}\left(k_z\frac{\partial E}{\partial z}\right)=\rho C\frac{\partial E}{\partial t} \qquad (6)$$

3. Control of the Temperature Distribution

Lyapunov theorem is used in derivation of controlling the temperature distribution. The Following steps detail this procedure:

1. A positive Lyapunov candidate should be defined over the domain.
2. The time derivative of Lyapunov theorem will be obtained and appropriate lemmas will be used to study the time derivative of the Lyapunov candidate at the boundaries of the system.
3. Appropriate boundary control heat fluxes will be chosen to make the time derivative of the Lyapunov candidate negative definite.

Figure 4 graphs the numerical representation for positive Lyapunov candidate.

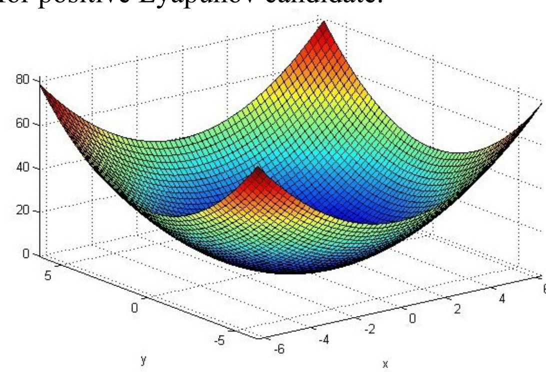


Figure 4. A positive definite candidate for Lyapunov functional

A positive definite Lyapunov candidate can be represented by

$$V = \frac{1}{2}\int_0^l \int_0^{2\pi} \left(\frac{k_\theta}{r^2}\left(\frac{\partial E}{\partial \theta}\right)^2 + k_z\left(\frac{\partial E}{\partial z}\right)^2 \right) rd\theta dz \tag{7}$$

The time derivative of the Lyapunove functional of equation (7) is

$$\dot{V} = \int_0^l \int_0^{2\pi} \left(\frac{k_\theta}{r^2}\left(\frac{\partial^2 E}{\partial t \partial \theta}\right)\left(\frac{\partial E}{\partial \theta}\right) + k_z\left(\frac{\partial^2 E}{\partial t \partial z}\right)\left(\frac{\partial E}{\partial z}\right) \right) rd\theta dz \tag{8}$$

Integration by parts of equation (8) yields

$$\dot{V} = \int_0^l \frac{\partial E}{\partial t}\left(k_\theta \frac{1}{r}\left(\frac{\partial E}{\partial \theta}\right)\right) dz + \int_0^{2\pi} \frac{\partial E}{\partial t}\left(k_z\left(\frac{\partial E}{\partial z}\right)\right) rd\theta - \int_0^l \int_0^{2\pi} \frac{\partial E}{\partial t}\left(\frac{1}{r^2}\frac{\partial}{\partial \theta}\left(k_\theta \frac{\partial E}{\partial \theta}\right) + \frac{\partial}{\partial z}\left(k_z \frac{\partial E}{\partial z}\right)\right) rd\theta dz \tag{9}$$

By considering equation (6), equation (9) is rearranged as

$$\dot{V} = \int_0^l \frac{\partial E}{\partial t}\left(k_\theta \frac{1}{r}\left(\frac{\partial E}{\partial \theta}\right)\right) dz + \int_0^{2\pi} \frac{\partial E}{\partial t}\left(k_z\left(\frac{\partial E}{\partial z}\right)\right) rd\theta - \int_0^l \int_0^{2\pi} \rho C \frac{\partial E}{\partial t}\left(\frac{\partial E}{\partial t}\right)^2 rd\theta dz = \int_0^l \frac{\partial E}{\partial t}\left(\left(\frac{k_\theta}{r}\frac{\partial T}{\partial \theta} - \frac{k_\theta}{r}\frac{\partial T_d}{\partial \theta}\right)\right) dz + \int_0^{2\pi} \frac{\partial E}{\partial t}\left(\left(k_z \frac{\partial T}{\partial z} - k_z \frac{\partial T_d}{\partial z}\right)\right) rd\theta - \int_0^l \int_0^{2\pi} \rho C \frac{\partial E}{\partial t}\left(\frac{\partial E}{\partial t}\right)^2 rd\theta dz \tag{10}$$

Since the net heat flux that passes through the cross section of the shell is equal to zero, the first term on the right-hand side of the equation (10) is zero, and rearranging results in

$$\dot{V} = \int_0^{2\pi} \frac{\partial E}{\partial t}\left(\left(k_z \frac{\partial T}{\partial z} - k_z \frac{\partial T_d}{\partial z}\right)\right) rd\theta - \int_0^l \int_0^{2\pi} \rho C \left(\frac{\partial E}{\partial t}\right)^2 rd\theta dz \tag{11}$$

To make $\dot{V}$ negative-definite, the first two terms on the right-hand side of equation (10) must be vanished. If heat fluxes in the form of equations (12) and (13) are applied, the first term at the right-hand side of equation (11) also vanishes and $\dot{V}$ becomes a negative-definite functional, equation (14).

At $z = 0$ $\qquad U_{z0} = k_z \left.\frac{\partial T_d}{\partial z}\right|_{z=0} f_{z0}$ (12)

At $z = l$ $\qquad U_{zl} = k_z \left.\frac{\partial T_d}{\partial z}\right|_{z=l} f_{zl}$ (13)

where

$$\begin{cases} f_{z0} = 1 & |E(1,\theta,0)| > 10^{-3} \\ f_{z0} = 0 \text{ or } \frac{\partial E}{\partial t} = 0 & |E(1,\theta,0)| < 10^{-3} \end{cases}$$

$$\begin{cases} f_{zl} = 1 & |E(1,\theta,l)| > 10^{-3} \\ f_{zl} = 0 \text{ or } \frac{\partial E}{\partial t} = 0 & |E(1,\theta,l)| < 10^{-3} \end{cases}$$

$$\dot{V} = -\int_0^l \int_0^{2\pi} \rho C \left(\frac{\partial E}{\partial t}\right)^2 rd\theta dz = -\int_0^l \int_0^{2\pi} \frac{1}{\rho C}\left(\frac{1}{r^2}\frac{\partial}{\partial \theta}\left(k_\theta \frac{\partial E}{\partial \theta}\right) + \frac{\partial}{\partial z}\left(k_z \frac{\partial E}{\partial z}\right)\right)^2 rd\theta dz \tag{14}$$

4. Numerical Results

As proved in the previous section, any desired temperature distribution, which satisfies the steady-state heat diffusion equation can be achieved. For the numerical analysis, we consider a cylindrical shell with radius $r = 1m$, height $h = 2\pi\ m$, and conductivity $k_\theta = k_z = 10\frac{J}{m.^{\circ}C}$. The desired temperature distribution is considered to be

$$T_d = e^z \sin\theta \tag{13}$$

Thus, the exact values of the control heat fluxes are

At $z = 0\ m$ $\qquad U_{z0} = 10 \sin\theta$ (14)

At $z = 2\pi\ m$ $\qquad U_{z0} = 10e^{2\pi} \sin\theta$ (15)

A finite element model was developed with 130×130 uniformly distributed nodes at the surface of the shell, and the values of heat fluxes at each node were obtained. The values of the exact, FEA control heat flux, as well as the error at the boundaries $z = 0$ and $z = 2\pi$ are displayed in Tables 1 and 2.

Table 1. Exact, FEA control heat fluxes and the error at $z = 2\pi\ m$

θ (radian) Node Position at $z = 2\pi\ m$	FEM/1000	Exact/1000	Error %
0.193329	5.1301	5.2552	2.380499
0.43499	4.7407	4.8562	2.378403
0.676651	4.0757	4.1751	2.380781
0.918312	3.1739	3.2513	2.380586
1.159973	2.0877	2.1386	2.380062
1.401634	0.8801	0.9015	2.373821
1.643295	-0.3787	-0.3879	2.371745
1.884956	-1.6154	-1.6548	2.380952
2.126617	-2.7582	-2.8255	2.381879
2.368277	-3.7408	-3.832	2.379958
2.609938	-4.5059	-4.6158	2.380952
2.851599	-5.0092	-5.1313	2.379514
3.09326	-5.2214	-5.3487	2.380018
3.334921	-5.1301	-5.2552	2.380499
3.576582	-4.7407	-4.8562	2.378403
3.818243	-4.0757	-4.1751	2.380781
4.059904	-3.1739	-3.2513	2.380586
4.301565	-2.0877	-2.1386	2.380062
4.543226	-0.8801	-0.9015	2.373821
4.784887	0.3787	0.3879	2.371745
5.026548	1.6154	1.6548	2.380952
5.268209	2.7582	2.8255	2.381879
5.50987	3.7408	3.832	2.379958
5.751531	4.5059	4.6158	2.380952
5.993192	5.0092	5.1313	2.379514
θ (radian) Node Position at $z = 2\pi\ m$	FEM/1000	Exact/1000	Error %
6.234853	5.2214	5.3487	2.380018

Table 2. Exact, FEA control heat fluxes and the error at $z = 0\ m$

θ (radian): Node Position at $z = 0\ m$	FEM	Exact	Error %
0.193329	-9.5805	-9.8137	2.37627
0.43499	-8.8532	-9.0687	2.376305
0.676651	-7.6115	-7.7967	2.375364
0.918312	-5.9273	-6.0716	2.376639
1.159973	-3.8987	-3.9936	2.376302
1.401634	-1.6436	-1.6836	2.375861
1.643295	0.7071	0.7243	2.374707
1.884956	3.0167	3.0902	2.378487
2.126617	5.151	5.2764	2.37662
2.368277	6.9859	7.156	2.377026
2.609938	8.4149	8.6197	2.375953
2.851599	9.3548	9.5825	2.376207
3.09326	9.751	9.9883	2.37578
3.334921	9.5805	9.8137	2.37627
3.576582	8.8532	9.0687	2.376305
3.818243	7.6115	7.7967	2.375364
4.059904	5.9273	6.0716	2.376639
4.301565	3.8987	3.9936	2.376302
4.543226	1.6436	1.6836	2.375861
4.784887	-0.7071	-0.7243	2.374707
5.026548	-3.0167	-3.0902	2.378487
5.268209	-5.151	-5.2764	2.37662
5.50987	-6.9859	-7.156	2.377026
5.751531	-8.4149	-8.6197	2.375953
5.993192	-9.3548	-9.5825	2.376207
6.234853	-9.751	-9.9883	2.37578

Figure 5 and 6 are plots of the above numerical results.

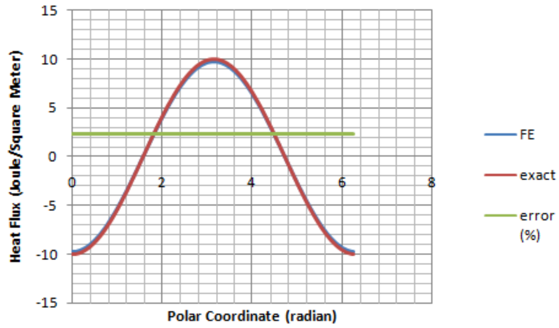


Figure 5 Exact and FE control heat fluxes and the error at $z = 0\,m$

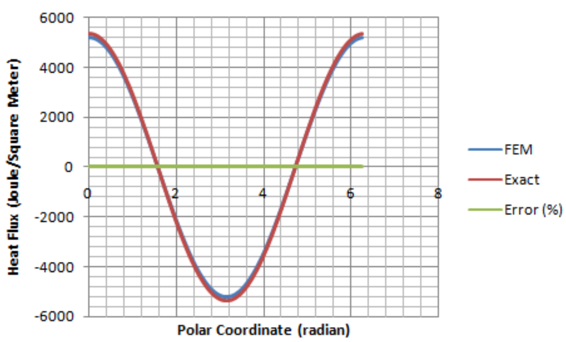


Figure 6 Exact and FE control heat fluxes and the error at $z = 2\pi\,m$

It is worth noting that the most significant source of error between the analytical and FEA values of the heat flux at the boundary is due to the variation in temperature distribution, which is exponential.

5. Conclusion

An elegant analytical inverse method for control of temperature distribution in thin cylindrical shells was developed. The method is significant in the light of inaccessible internal domain such as in cases of control of temperature distribution in a hot body. Lyapunov theorem was used to control the transient temperature distribution. A Lyapunov functional was also defined and the control heat fluxes at the boundaries were assigned in such a way that they make the time derivative of the Lyapunov functional negative definite. Furthermore, the numerical control heat fluxes at the boundaries $z = 0\,m$ and $z = 2\pi\,m$ were obtained by the finite element method. The developed method is shown to be robust, efficient and accurate. The numerical results validated the proposed method, and the maximum error was about 2%. (The numerical accuracy can be increased by increasing the number of nodes in finite element model).

6. References

[1] Y. Miyamoto, W. A. Kaysser, B. H. Rabin, A. Kawasaki, and R. G. Ford, Functionally Graded Materials: Design, Processing and Applications, Kluwer, Dordrecht, The Netherlands, 1999.

[2] L. Marin, "Numerical Solution of the Cauchy Problem for Steady-State Heat Transfer in Two-Dimensional Functionally Graded Materials", International Journal of Solids Structures, vol. 42, pp. 4338–4351, 2005.

[3] B. Chen, L. Tong, Y. Gu, H. Zhang, and O. Ochoa, "Transient Heat Transfer Analysis of Functionally Graded Materials Using Adaptive Precise Time Integration and Graded Finite Elements", Numerical Heat Trans. for B, vol. 45, issue 2, pp. 181–200, 2004.

[4] B. L. Wang and Z. H. Tian, "Application of Finite Element-Finite Difference Method to the Determination of Transient Temperature Field in Functionally Graded Materials", Finite Elements in Analysis and Design, vol. 41, pp. 335–349, 2005.

[5] J. Sladek, V. Sladek, and C. H. Zhang, "Transient Heat Conduction Analysis in Functionally Graded Materials by the Meshless Local Boundary Integral Equation Method", Computational Materials Science, vol. 28, pp. 494–504, 2003.

[6] A. Sutradhar and G. H. Paulino, "The Simple Boundary Element Methods for Transient Heat Conduction Analysis in Functionally Graded Materials", Computer Methods in Applied Mechanics and Engineering, vol. 193, pp. 4511–4539, 2004.

[7] L. F. Qian and R. C. Batra, "Three-Dimensional Transient Heat Conduction in a Functionally Graded Thick Plate with a Higher-Order Plate Theory and a Meshless Local Petrov-Galerkin Method", Computational Mechanics, vol. 35, pp. 214–226, 2005.

[8] M. R. Golbahar Haghighi, M. Eghtesad, D. S. Necsulescu and P. Malekzade, "Temperature Control of Functionally Graded Plates Using a Feedforward-Feedback Controller Based on The Inverse Solution and Proportional-Derivative Controller," Energy Conversion and Management, Volume 51, Issue 1, January 2010, Pages 140-146.

[9] H. Rastgoftar, M. Eghtesad and A. Khayatian, "Boundary Control of Temperature Distribution in a Rectangular Functionally Graded Material Plate", Journal of Heat Transfer, Volume 133, Issue 2, 021304 (6 pages) 2011.

Assessment of Excavator Operator Performance Using an Integrated Human Performance Model

Khaliah Hughes, *Steven Jiang, Zongliang Jiang, Eui Park, Daniel Mountjoy
Department of Industrial and Systems Engineering, North Carolina A&T State University, Greensboro, NC 27411, USA
Corresponding author: xjiang@ncat.edu

Abstract

Studies on human performance often focus on either cognitive or physical aspects, failing to take into account the interaction that produces such performance. In this research, an integrative framework was developed to create a comprehensive human performance model and an empirical study was conducted to investigate the effects of various factors on hydraulic excavator operator performance. Results of the study indicated that the integrated model accounts for the interaction between cognitive and physical components of performance.

1. Introduction

Human performance modeling has been widely utilized in engineering research to create representations of complex systems. Such models range from simple written equations or mathematical statements to complex computer simulation models that provide significant value when predicting performance [1].

Over the years, human performance models have advanced, leading to better predictions of system and human behavior [2]. Progress in the area has been realized through improved software, accurate data, and better knowledge of human capabilities. Much of this advancement, however, is attributable to techniques that more accurately model human performance within acceptable limits, enabling researchers to predict future outcomes, identify system deficiencies, assess the effectiveness of operations, and gain insight on the underlying processes involved in human behavior [3]. One such technique that has been popularized to improve human performance models is the theoretical approach whose models often yield in better logic that parallels human reasoning and interaction [2].

Though many approaches emphasize the human operator, majority of these studies concentrate on cognitive or physical processes separately, when both interact to create human performance. Models that neglect or improperly model either case can have a significant impact on performance predictions, leading to gaps in models and overestimations of both human and system efficiency. Thus, both cognitive and physical aspects must be modeled to accurately predict human performance in emergent research models. Such research is even more critical in the fluid power domain where excavator systems have evolved from hydraulic to electronic control mechanisms, increasing complexity on the human operator.

This research presents a new technique to study hydraulic excavator operator performance. The goal of this research is to utilize a theoretical framework that integrates cognitive and physical aspects, forming a comprehensive human performance model [4]. With this objective, the framework will act as a blueprint in model development by establishing the proper methods to assess, integrate, and implement better models of human performance.

2. Framework Development

In the development of the framework, procedures were followed to support the integration of both cognitive and physical factors to create better models of human performance. Framework structure consisted of performance assessment, model integration, and implementation, acting as a blueprint for integrated models. The following sections describe that development in further detail.

2.1. Performance assessment

Human performance is multi-dimensional, being subject to the effects of many factors that are often considered separately in traditional research models. Since this quality holds the potential to positively or negatively affect performance, it was appropriate to consider at various levels. The appropriate levels at which to assess human performance were determined to set the proper boundaries that constrain performance, convey interaction, and the individual differences that shape behavior.

2.1.2. Definition of performance states: Performance states (Figure 1) were defined to form an environmental wherein human performance can occur. Initial performance consisted of the internal state, representing intangible processes existing within the mind or body. The subsequent external performance state represented tangible processes existing independently from with respect to the body. Both states have a linear relationship that is interrupted by the transformation state, bridging internal and external processes, facilitating action, and creating a full performance representation. Within each state, task, human, system, and environmental information was specified through an approach indicating the required elements to be altered when modeling.

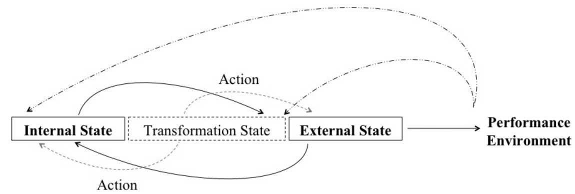


Figure 1. Definition and relationship of performance states [1].

2.1.3. Selection of modeling tools: Simulation provided an efficient method to model and to analyze human performance with regard to the interaction between the human, system, and environment. To avoid loss of accuracy, assumptions regarding behavior, and validate models, cognitive and physical simulation tools were selected to support the framework's structure. Both tools expanded beyond the capabilities of past simulation models and acted as a means to better consider both components of performance. Each tool was later integrated in the framework to create more valid simulation models.

2.1.4. Extraction and linking variables: Based on the modeling capabilities of each tool, cognitive and physical performance variables were extracted to depict interaction and form relationships between independent models. Relationships were formed through linking variables in cognitive and physical human performance models (Figure 2). By linking performance variables, models can be integrated; thus, compensating for a lack of dimension (i.e. cognitive variables compensate for performance lacking in the physical model, and physical variables compensate for performance lacking in the cognitive model). Such compensatory relationships better predict performance by reflecting that which occurs in the real world.

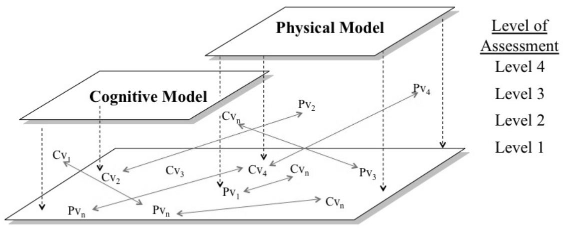


Figure 2. Linking of performance variables and layered metrics [1].

2.2. Performance integration

The integration of models was facilitated through the prior methods to enable an accurate representation of human performance. By using the acquired knowledge, techniques were used to create models that account for both cognitive and physical facets of human performance in an integrated performance model.

2.2.1. Tools: Two tools were selected to simulate cognitive and physical performance. Inputs required for the cognitive tool were identified and compared with the required inputs of the physical modeling tool. Differences in tool features were bridged based on shared modeling capabilities. Tools were then combined to model the system in a virtual environment along with the physical and cognitive tasks of the human operator, returning output on performance.

2.2.2. Variables: Since a bi-directional relationship exists between physical and cognitive components, variables were integrated to produce human performance models that simulate responses and predict interaction. Intra-variable relationships were used to convey the relationship between variables within the same model (i.e. cognitive-to-cognitive variables or physical-to-physical variables); whereas, inter-variable relationships were used to convey the relationship between variables of independent models (i.e. cognitive-to-physical variables). Such relationships depict the impact of cognitive and physical factors and the manner in which they impact performance.

2.3. Framework implementation

Implementation of the framework approach was necessary to demonstrate the benefits of accounting for various factors that impact performance. With the framework's defined structure, a comprehensive representation of human performance was achieved.

2.3.1. Model representation: Upon completing each of the components involved in the development of the framework, a comprehensive representation of performance was obtained. The integrated representation served to gain valuable insight on the correlation between cognitive and physical factors of human performance, to acknowledge the interactions that produce operator behaviors, and to better replicate and predict human performance.

2.3.2. Model structure: Structures of the integrated model representation (Figure 3) were used to depict human performance and to facilitate interaction from the flow of logic that parallels human reasoning and physical action. The integrated performance representation consisted of four primary areas: human centered factors from which performance metrics were used as a subset to classify performance variables, a functional relationship to connect cognitive and physical performance, discrete-event simulation tools to facilitate integration, and model output to provide a representation of integrated performance.

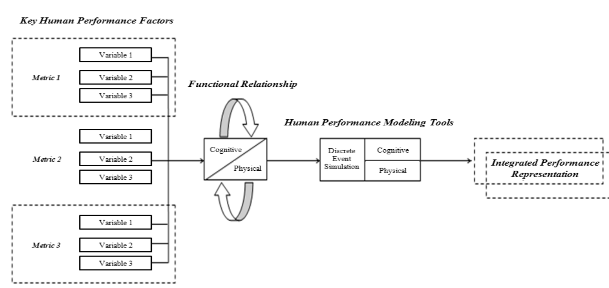


Figure 3. Integrated human performance model representation [1].

3. Empirical Study

For the development of the integrated framework, procedures were followed to ensure that human performance models derived from the framework offer the same degree of validity as real world fluid power applications involving a hydraulic excavator. Procedures consisted of data collection, tool selection, and variable identification to act as components in model development. The following sections describe each of these procedures in greater detail.

3.1. Data collection

Real world data was collected on the work tasks, control operations, and system functions of a hydraulic excavator in its natural work environment. Information was obtained through three primary resources: interviews of expert operators to provide a thorough understanding of work operations and skill requirements; system manufacturers to provide details on implemented design changes, model specifications, and tooling for work applications; and video recordings to gain insight on time variations in dynamic work settings and human-system interaction. Research literature and case studies were also used to supplement general data.

3.1.1. Task analysis: Collected data was used to construct task analyses from 40 video recordings. Such analyses were based on excavator control type (e.g. hydraulic and electronic) and the work environment (e.g. soil and gravel), which have the potential to significantly impact performance. Each analysis outlined the sequence of steps involved in the excavation process, as well as assessed task requirements in terms of the operator's cognitive and physical processes to achieve goals. Based on the observations from the analysis, key tasks were identified.

3.1.2. Time studies: To complement the task analysis, time studies were performed to determine operator efficiency in reaching work goals, identify inefficient work methods, and quantify changes in performance. Timing data was captured from each of the categorized video recordings of excavator operators. Critical tasks were timed, recorded, and organized by the sequence of the task analysis. From this data simple statistics were calculated, and distributions were fit to tasks for accurate modeling.

3.2. Modeling tools and software

For this study, two simulation tools, cognitive and physical, were chosen based on the requirements of the framework to quickly, safely, and economically model performance [5]. In this study, Micro Saint was used to model cognitive performance and Jack was used to model physical performance. A description and purpose of each tool used for the study is provided below.

3.2.1. Micro saint: Micro Saint modeling software was used to model the cognitive components of performance such as human cognition and decision-making. In the software, models were denoted by a task network diagram of nodes and arrows, depicting the sequence of human activities. This tool was used to evaluate performance in areas such as process definition, design, workload, safety, and productivity of hydraulic excavation processes.

3.2.2. Jack: For the physical performance representation, Jack software was used to model physical human capabilities and limitations with regard to the system and the environment. The tool offered visualization models of a digital human, known as Jack, who carries out scheduled tasks and procedures within a virtual environment. Applications with the hydraulic excavator included the analysis of the anthropometry, biomechanics, and task behaviors.

3.3. Model development

As previously stated, the framework was used as the blueprint for model development. Based on the framework's requirements, two tools were chosen. The software was used to build an integrated model that simulates both cognitive and physical human performance.

3.3.1. Cognitive model: In Micro Saint simulation software, findings from the task analysis were used to create task network models, extending each task analysis into a predictive model. Each task network model was organized by the task sequences executed by hydraulic excavator operators to accomplish the desired goal. Task network model hierarchies included the movement task, operator task, and the system task that served as the basis for the task network model and sub-model of the human-machine system. Data recorded from the time studies and statistical distributions were embedded within the task networks to accurately model operator tasks and excavation process. Timing information for each task, in terms of mean time, standard deviation, and the appropriate distribution, allowed the model to simulate process variance and to provide a high level of validity for workload estimates or modeling results [6].

3.3.2. Physical model: In Jack simulation software, manufacturer specifications and human anthropometric data were used to model the excavator system and physical functioning of the human operator. Like Micro Saint simulation models, Jack models were built based on excavator control type and environmental terrain. Development of the models involved building a representation of the system and defining procedures to be performed by the human operator based on manufacturer specifications and the tasks analysis. A digital human was used to simulate excavation procedures in a virtual environment generated in the software. Tasks followed by the digital human were governed by the sequence of activities outlined in the task analyses and Micro Saint network models. Timing and frequency data recorded from the time studies was used to accurately simulate operator tasks and excavation process as well as to validate model results. Models replicated the operator's physical actions, capabilities, and limitations in response to interface designs, creating predictive models of the human-machine system.

Within each software, simulation models were programmed and coded according to the requirements and components defined in the theoretical framework. Variables were extracted and modeled, conveying the functional relationship between cognitive and physical performance and linking cognitive models to physical models. Output obtained from these models was used in an empirical study.

3.4. Data analysis

An empirical study on human performance was conducted using the integrative framework to assess the impact of cognitive and physical factors. The following sections describe the empirical study in further detail.

3.4.1. Stimulus material: The empirical study on the human performance models was conducted using the integrated model derived from the cognitive and physical models.

3.4.2. Simulation tasks: Key tasks were selected for modeling in simulation software to provide insight on the interaction between the operator and excavator system. Tasks were based on their relevance to the framework, significance in the excavation processes, and value in assessing performance. Simulation tasks primarily concentrated on the active work process (Figure 4) to convey the work procedures involved in the excavation of materials. Such tasks are critical to performance because they hold the potential to vary extensively based on the environmental conditions, system design, or human abilities.

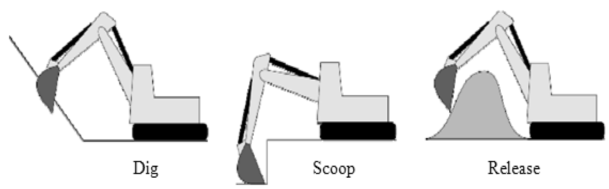


Figure 4. Tasks simulated in human performance models [7].

3.4.3. Equipment: Simulation models were run in Micro Saint and Jack simulation software. The software was installed on a laboratory PC with a Microsoft XP operating system.

3.4.4. Experimental design: A 2x2 factorial design was used in this study to analyze human performance in the non-integrated models. The two independent variables consisted of excavator control type at two levels (e.g. hydraulic and electronic) and environmental terrain at two levels (e.g. soil and gravel). The dependent variables were task completion time and workload.

3.4.4. Procedure: Micro Saint and Jack performance models were completed in compliance with the integrated framework. Initial models were created using Micro Saint, defining the cognitive functioning of the human operator and the sequence of tasks involved in hydraulic excavation processes. Jack software was then used to expand beyond the predictive capabilities of the cognitive models through physical models, which were linked to convey a functional relationship between internal and external human functioning. Models were randomly executed 100 trials and simulation output was simultaneously documented and recorded based on framework requirements.

3.4.5. Data collection: As the simulation models were executed, performance was monitored, documented, and recorded. Output for the models was displayed visually in the form of numerical and graphical data. Descriptive statistics and empirical results were obtained from simulation output. Model output comprised the dependent variables (e.g. completion time and workload).

4. Results

Four Micro Saint and four Jack models were developed to simulate the performance of operators using hydraulic and electronic control excavator systems under two environmental conditions, soil and gravel. Each model was run and performance output was obtained. Models were integrated by interchanging corresponding sets of inputs and outputs, enabling models to comprehensively model human performance.

4.1. Task analysis

For the integrated models, tasks analyses were used to decompose work processes and provide a better understanding of human performance. The following sections provide details from the results of those analyses.

4.1.1. Hydraulic control-soil terrain: Findings of the task analysis for operators of hydraulic control systems in soil terrain revealed that work tasks were performed in three phases. Tasks consisted of monitoring, positioning, and digging processes that were facilitated by cognitive and physical procedures. Physical tasks involved manipulation of controls; whereas, cognitive tasks involved the selection of the proper decision making strategies.

Movement tasks under such conditions consisted of turning the head, extending the leg, reaching the arm, and tilting the hand as seen in Figure 5. The longest task time was observed with reaching the arm which yielded an average time of 2.7 seconds followed by extending the leg with an average time of 1.8 seconds. The shortest tasks consisted of turning the head with an average of 1.2 seconds and tilting the hand with approximately 1.1 seconds. Tasks with longer completion times generally resulted in less variation than those with shorter times (e.g. tilting the hand at 23%).

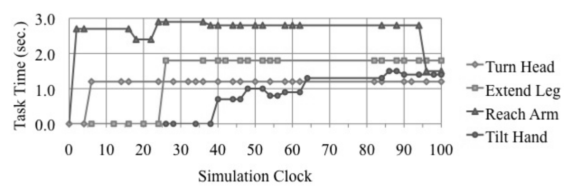


Figure 5. Integrated movement tasks for hydraulic-soil model.

Operational tasks consisted of turning the starter, checking the monitor panel, as well as engaging the lever and joystick controls. Figure 6 illustrates that checking the monitor panel by far yielded the highest time yielding an average time of 6.0 seconds. All other tasks yielded lower values relatively close in nature. Turning the starter switch yielded an average time of 2.0 seconds and pushing the lever yielded an average time of 1.7 seconds. The shortest tasks involved the joystick controller with an average 0.86 seconds for adjusting and 0.84 seconds for pushing or pulling. Again, longer tasks yielded in less variation.

4.1.2. Hydraulic control-gravel terrain: Operators of hydraulic control systems in gravel terrain also engaged in cognitive and physical tasks consisting of monitoring, positioning, and digging as seen in Figure 7.

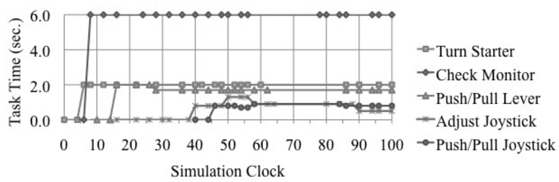


Figure 6. Integrated Operation Tasks for Hydraulic-Soil Model.

The shortest movement tasks for these operators were tilting the hand with an average time of 1.6 seconds, followed by reaching the arm with an average time of 2.5 seconds. As in the prior model, faster tasks generally yielded in the most extreme variation and the least consistency during the simulation with coefficients of variation ranging from 26% to 63%. In contrast, turning the head and reaching tasks took operators longer, averaging 3.5 seconds and 2.6 seconds. These tasks appeared to have the most consistent trends throughout the entire hydraulic excavation process.

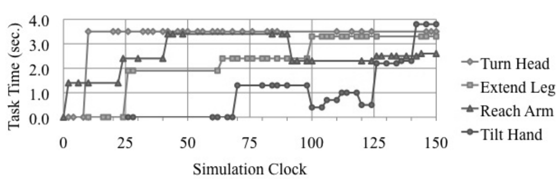


Figure 7. Integrated Movement Tasks for Hydraulic-Gravel Model.

Operational tasks revealed that the longest task performed by the human operator involved checking the system's monitor, taking and average of 10.0 seconds. All other tasks yielded significantly lower values such as pushing or pulling levers, taking an average of 3.9 seconds as well as turning the starter switch taking an average of 2.0 seconds to complete. The shortest times were found to involve the joystick controllers, yielding an average time of 1.4 seconds for adjusting and 1.3 seconds for pushing or pulling. Figure 8 revealed that longer tasks yield less variation (e.g. checking the monitor at 0%) and shorter more variation (e.g. pushing or pulling the lever at 29%).

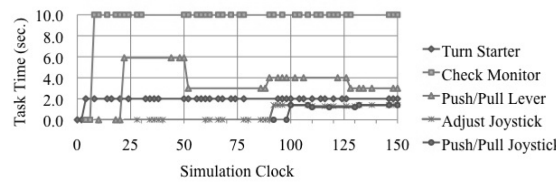


Figure 8. Integrated operation tasks for hydraulic-gravel model.

4.1.3. Electronic control-soil terrain: Work processes for electronic control systems excavating soil involved initialization, active work, and finalization tasks with respect to monitoring, positioning, and digging. Instead of controlling the system through motion of the joysticks, these functions were executed through buttons embedded within the joystick controls.

Movement tasks consisted of reaching the arm, turning the head, and moving the finger. In Figure 9, the shortest time for operators of electronic control excavators was moving the finger with an average time of approximately 0.9 seconds, followed by reaching the arm with 1.5 seconds. Turning the head took the longest time to complete, taking an average of 2.8 seconds with no variation; whereas, moving the finger resulted in high variation at 42%.

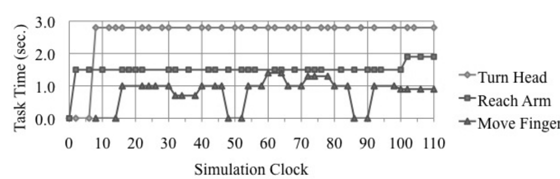


Figure 9. Integrated movement tasks for electronic-soil model.

Operational tasks for hydraulic systems in gravel terrain consisted of turning the starter switch, checking the monitor panel, and pressing the joystick buttons. On average, the longest operational task was associated with checking the system's monitor panel, yielding in a value of approximately 3.5 seconds, followed by turning the starter switch at approximately 2.0 seconds. The most variable operational task was revealed to be that of pressing the joystick button which yielded in an average time of 2.2 seconds. This task yielded in the greatest amount of variation at 90%; whereas, the most consistent task was checking the monitor which yielded in no variation as shown in Figure 10.

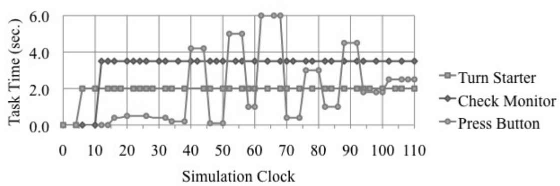


Figure 10. Integrated operation tasks for electronic-soil model.

4.1.4. Electronic control-gravel terrain: Operators of electronic control systems in gravel terrain also executed key tasks such as monitoring, positioning, and digging by utilizing joystick controller buttons. Movement tasks in Figure 11 revealed that the shortest movement was moving the finger with an average time of 1.2 seconds. The longest movement tasks were reaching the arm and turning the head with average task times of 2.0 seconds and 1.5 seconds. Data trends again revealed that longer tasks yield less variation (e.g. reaching the arm at 0%) and shorter tasks yield in more variation (e.g. moving the finger at 40%).

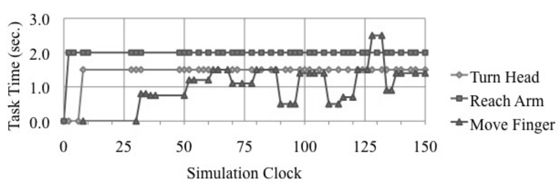


Figure 11. Integrated movement tasks for electronic-gravel model.

The longest task for electronic control systems in gravel terrain as seen in Figure 12 was checking the monitor with a mean time of 10.0 seconds; whereas, the shortest task was turning the starter, yielding a time of approximately 2.0 seconds. Like the other models, the most variation was evident in shorter tasks such as pressing the button (e.g. 82%) with an average time of 3.5 seconds, and the least variation was evident in longer tasks such as checking the monitor or turning the starter at 0%.

4.2. Experimental results

Prior models yielded relevant data regarding the processes associated with various tasks and operations throughout the excavation process. The following sections examine the significance of those effects through an empirical study.

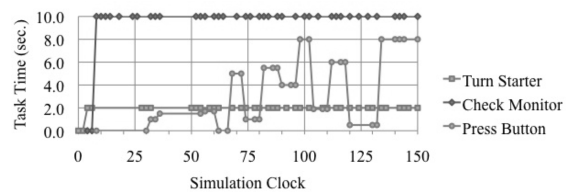


Figure 12. Integrated operation tasks for electronic-gravel model.

4.2.1. Descriptive statistics: In the integrated models, completion time (Table 1) was assessed subject to the effects of physical fatigue. From the results, it was found that operators of electronic control systems in soil terrain yielded in the lowest mean completion time of 165.7 seconds; whereas, operators of hydraulic control systems in gravel terrain yielded in the longest mean completion time of 392.9 seconds. In general, completion time with electronic control was lower in both environments as compared with hydraulic control. Soil also yielded lower completion times than gravel for both systems.

Table 1. Performance model summary for integrated completion time (sec).

Descriptive Statistics	Hydraulic		Electronic	
	Soil	Gravel	Soil	Gravel
Mean	241.6	392.9	165.7	363.6
Standard Deviation	166.6	576.6	125.1	515.9

Table 2 results revealed that the greatest workload for digging tasks occurred with electronic control excavator systems in gravel terrain, yielding an average of approximately 79%; whereas, the least workload occurred with hydraulic control systems in soil terrain, yielding a value of 45%. With respect to control type, operators of hydraulic systems experienced less workload (e.g. 47% for soil and 64% for gravel) than operators of electronic control systems (71% for soil and 79% for gravel).

4.2.2. Inferential statistics

A two-way analysis of variance (Table 3) revealed a significant main effect for the environment (F (1, 396) = 18.9, p < 0.0001) on completion time for excavation processes.

Table 2. Performance model summary for integrated workload (%).

Descriptive Statistics	Hydraulic		Electronic	
	Soil	Gravel	Soil	Gravel
Mean	46.8	63.8	71.4	78.6
Standard Deviation	14.2	16.8	19.5	15.8

Table 3. ANOVA-2x2 factorial design for integrated completion time.

Source	DF	Type I SS	Mean Square	F-Value	Pr > F
Control	1	277,075.9	277,075.9	1.73	0.1897
Environment	1	3,048,341.4	3,048,341.4	18.90	<0.0001
Control*Environment	1	54,400.9	54,400.9	0.34	0.5608
Error	396	63,563,572.3	160,514.1		
Total	399	66,943,390.5			

No significant effect was found regarding the system's control type (F (1, 396) = 1.73, p = 0.1897) or for the interaction between the system's control type and the environment (F (1, 396) = 0.34, p = 0.5608) as seen in Figure 13.

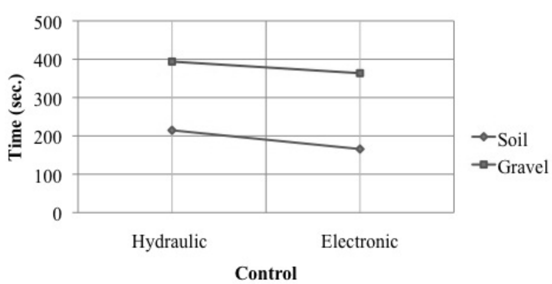


Figure 13. Interaction plot for integrated completion time.

Analysis of workload (Table 4) revealed a significant main effect for control (F (1, 396) = 220, p < 0.0001) and the environment (F (1, 396) = 29.5, p < 0.0001). A significant interaction effect was also found between control type and the environment (F (1, 396) = 29.5, p < 0.0001).

Table 4. ANOVA-2x2 factorial design for integrated workload.

Source	DF	Type I SS	Mean Square	F-Value	Pr > F
Control	1	54,079.5	54,079.5	220.2	<0.0001
Environment	1	7,233.5	7,233.5	29.5	<0.0001
Control*Environment	1	7,233.5	7,233.5	29.5	<0.0001
Error	396	97,278.1	245.7		
Total	399	165,824.6			

Further analysis (Figure 14) revealed a significant effect on workload when sliced by hydraulic control (F (1, 396) = 58.9, p < 0.0001) and no effect for electronic control (F (1, 396) = 0.0, p = 1.0000). When sliced by environment, soil (F (1, 396) = 205.3, p < 0.0001) and gravel (F (1, 396) = 44.3, p < 0.0001) yielded significant effects.

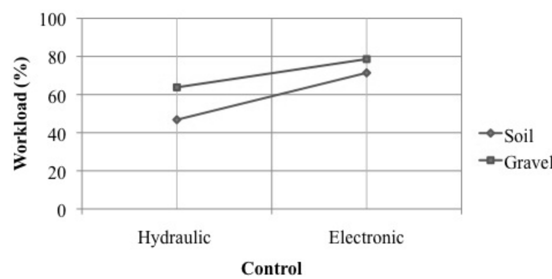


Figure 14. Interaction plot for integrated workload.

5. Discussion

Findings of the empirical study produced relevant implications on the mechanisms of human performance that could not be explained when assessed independently. The following sections describe implication from the results of the empirical study.

5.1. Control

Control was found to have a significant impact on the degree of workload experienced by the human operator. In particular, it was found that cognitive workload was lower in hydraulic control systems as compared to electronic control systems. In prior research, it has been suggested that the mechanics of the system can affect the degree of mental workload experienced by the human operator. It has even been found in some cases to increase workload [8]. Considering this perspective, cognitive workload may be higher with electronic control systems due to the design of its controllers which fail to match the mental models of the human operator. For instance, with hydraulic control systems, the range of motion in the joystick controls match the resultant movements carried out by the system (i.e. downward motion produces a lowing movement of the system's arm). However, in electronic systems, button mappings failed to match the human operator's mental models; thus, violating expectancies.

5.2. Environment

The environmental conditions under which excavation tasks were performed had a significant impact on human performance. For the two test conditions, it was found that operators completed work processes more quickly when excavating soil terrain as opposed to gravel terrain with both hydraulic and electronic control systems. The difference in the overall process completion times and more efficient performance in soil terrain can be explained by the elevated degree of difficulty in performing excavation processes under demanding environmental conditions. Though soil can vary extensively, it generally offers a sufficient degree of porosity and permeability to facilitate excavation. Gravel can also vary tremendously, ranging from fine pebbles to course boulders. However, it results in less efficiency due to its higher mass and weight in conjunction with the capacity of the system. Such difficulties alter the methods implemented by operators to accomplish work goals, resulting in high degree of variation within the models.

Furthermore, the environment also had a significant effect in terms of the workload experienced by the human operator. From the models, it was found that operators experienced more cognitive workload in gravel terrain as compared with soil terrain. Such results are attributable to the additional demands that the operator must take into consideration when undertaking work (e.g. load weights, balance, and bucket capacity); thus, adding to the complexity and difficulty of the work.

These notions lead to three significant ideas that can be formed regarding the impact of such findings from the integrated human performance models. Foremost, both rapid and prolonged work can cause operators to utilize methods reducing accuracy, increasing the risk of error [9]. In reality, excessive cognitive or physical exertions with complex systems can lead to costly mistakes, endangering human workers as well as the surrounding environment. Furthermore, when multiple variables consume attention, the operator can become mentally overloaded, inhibiting performance and jeopardizing safety [10]. Lastly, although operators may perform tasks adequately, excessive workload causes inadequate performance over prolonged periods [11]. Over time as workload increases, performance variations result. These variations increase the difficulty of work tasks and cause operator fatigue, slowing the overall work process

6. Conclusion

As established by this research, the integrated approach provided a strong foundation for the enhancement of methods used to model human performance in complex systems by coupling the interaction among the task, human, system, and the environment. Its structure acted as a guide, linking commonly misrepresented and overlooked elements for better predictions of human performance to understand behavior and its shaping factors. Since it is not possible to develop a framework that is applicable to all systems, the research described herein is limited to the scope of its parameters as well as the capabilities of the chosen software. However, alternatives exist to address such limitations in future research. The integrative framework provides opportunities for application with other complex systems and domains where humans are required to manage cognitive and physical tasks (e.g. transportation, manufacturing, and construction). Software is also now available which has the capability to represent both facets of performance within a single modeling tool. Hence, through continual development, integrative approaches will further provide value by increasing the validity of models to close the gap between existing and emergent research methods that model human performance in complex systems.

7. References

[1] K. Hughes, X. Jiang, Z. Jiang, D. Mountjoy, and E. Park, "A Preliminary Study of an Integrated Human Performance Model," Proceedings of the Industrial Engineering Research Conference, Cancun, Mexico, 2010, pp. 1-6.

[2] G. Campbell and J. Bowers, "The Application of Human Performance Models in the Design and Operation of Complex Systems," Naval Air Warfare Center Training Systems, 2000, pp. 1-4.

[3] C. Chapparo and J. Ranka, "Occupational Performance Model: Definition of Terms," University of Sydney, Sydney, 1996, pp. 58-60.

[4] K. Hughes and X. Jiang, "Development of an Integrative Framework to Model Cognitive and Physical Human Performance in Fluid Power Systems," International Fluid Power Conference and Expo, Las Vegas, Nevada, 2011, pp. 1-13.

[5] T. Baines, R. Asch, L. Hadfield, J. Mason, S. Fletcher, and J. Kay, "Towards a Theoretical Framework for Human Performance Modeling Within Manufacturing Systems Design," Simulation Modeling Practice and Theory, 13(6), 2005, pp. 486-504.

[6] C. Drury and R. Laughery, "Fundametals of Simulation Using Micro Saint," Proceedings of the Winter Simualtion Conference, Boulder, 2004, pp. 446-551.

[7] Y. Sakaida, D. Chugo, K. Kawabata, H. Kaetsu, and H. Asama, "The Analysis of Excavator Operation by Skillful Operator," SICE Annual Conference, Japan, 2006, pp. 543-547.

[8] R. Parasuraman, T. Sheridan, and C. Wickens, "A Model for Types and Levels of Human Interaction with Automation," IEEE Transactions on Systems, Man, and Cybernetics: Part A-Systems and Humans, 30 (3), 2000, pp. 286-297.

[9] R. Hockey, "Compensatory Control in the Regulation of Human Performance Under Stress and High Workload: A Cognitive-Energetical Framework," Biological Psychology, 45, 2005, pp. 73-93.

[10] E. Svensson, M. Angelborg-Thanderz, L. Sjoeberg, and S. Olsson, "Information Complexity: Mental Workload Performance in Combat Aircraft," Ergonomics, 40(3), 1997, pp. 362-380.

[11] N. Mazaeva, C. Ntuen, and G. Lebby, Self-Organizing Map (SOM) Model for Mental Workload Classifications," IFSA World Congress and 20th NAFIPS International Conference, Vancouver, British Columbia, 2002, pp. 1833-1825).

Selection of an Ideal Enterprise Integration Approach using Delphi-ANP Approach

James N. Ngeru, Tridip K. Bardhan and Richard A. Pitts, Jr.
Department of Industrial and Systems Engineering, Morgan State University, Baltimore, MD 21251, USA
jmngeru@gmail.com, tridip.bardhan@morgan.edu, richard.pitts@morgan.edu

Abstract

The need for today's organization to be competitively effective and agile has led to an exceptional interest in the field of enterprise integration (EI). Numerous organizations have already implemented EI while others are still juggling with such initiatives. Despite the growing interest in this field, there is still no clear methodology that guides stakeholders in identifying an EI approach that best meets their organization's business and technical integration needs. Therefore, the objective of this paper is to present formulation and demonstration of such a methodology through a case study. As selection of an EI approach can be viewed as a complex multi-criteria decision making problem, involving several stakeholders with varying preferences and who must gauge a set of business and technical factors to appraise different EI approaches, a multi-criteria decision methodology that blends in both Delphi technique and analytic network process (ANP) method is applied. ANP is favored in this case as it is relatively easy to implement and offers a comprehensive structure that enables extensive analysis of multiple decision elements and their interdependence effects on each other. Delphi technique is absorbed in ANP framework to facilitate pooling of group inputs and bring forth a sense of consensus in the selection process. The formulated methodology is then applied in a state government agency that is involved in implementation of an enterprise-wide integration. The case study's results revealed disparity of preferences towards the evaluation goal and criteria among different stakeholders. However, the use of the proposed methodology helped in assimilating stakeholders varying view-points, and hence a sense of group unanimity.

1. Introduction

Today's organizations are competing and creating new opportunities in global markets with global partners and consumers. The competitive and dynamic nature of the new business playground means new technologies methods and strategies are needed. To survive and thrive in this competitive and dynamic business setting, organizations in both private and public sector must have enough agility to respond to business needs and with effective service delivery [1, 2]. However, this goal is only attainable if there is swift sharing of information among business entities and processes both within and across organization. The last few decades has seen Enterprise Integration (EI) emerging as one revolutionary concept solution that promises to assist organization with achieving this goal. EI aims to link separate business processes giving them increased leverage [3]. Thus it is not surprising that for the past few years, survey after survey of top CIO have ranked EI initiatives, such as Service Oriented Architecture (SOA) adoption, Enterprise Resource Planning (ERP) adoption/upgrades, Data integration, in their list of top ten priorities [4, 5]. To many top executives, EI is now a business necessity. However, implementation of an EI strategy is not an easy task. Past researchers have called for thorough evaluation of organization's requirements before undertaking EI implementation [6].

Thus, this paper aims at presenting formulation and application of a systematic methodology for analysis in the selection of an EI strategy. Selection of an EI strategy is a multifaceted problem involving more than the choosing a vendor or integration software system. It is important to consider the organization's vision, requirements and constraints; the technical merits and limitations of the chosen strategy.

In spite of remarkable interest in overall EI concept, little has been done in development of a framework for EI implementation, especially in EI strategy selection. Most of the EI research has been centered in advancing the technology to enable integration. Janssen argues that organizational issues and stakeholders interests should be addressed in the selection process [7]. Thus, the need to develop a comprehensive framework incorporating a compendium of organizational factors during the selection of an EI strategy is indispensable [8].

The methodology discussed in this paper utilizes multi-criteria analysis principles, specifically integrating ANP method with Delphi technique. The Analytic Network Process (ANP) developed by Saaty [9] was chosen for its ability to incorporate multiple factors with different relative weights and with dependencies characteristics. This gives the decision makers the ability to prioritize the importance of one factor over another. The ANP has been used in complex problems involving technology selection, and thus it is not a far to see that it may be applicable in this situation. To further improve the results on using ANP within this context, Delphi technique is applied. The Delphi technique improves the process by facilitating active participation and consensus among evaluation process partakers representing different organization's functions.

The paper proceeds with an overview on EI, followed by methodology and a case study.

2. An Overview of Enterprise Integration

EI has been defined as *"an alignment of strategies, business processes, information systems, technologies, and data across organizational boundaries to provide competitive advantage"* [10]. Lack of this alignment is source of many organizational problems. According to Chen *et al* [11], integration can be approached in various ways and at various levels, such as, physical, application, business and can also be achieved through enterprise modeling approach or any methodological approach facilitating consistent enterprise-wide decision making. The process of achieving integration involves all managerial and technological factors that enable cross-functional process integration [10]. Silveira *et al* [12] classified EI based on its scope i.e. whether it involves integrating processes and applications within or outside organizational boundaries. Inter-organizational enterprise integration, attempts to integrate business processes between enterprises (B2B), such as Supply Chain Management systems (SCMs), or electronic purchasing processes (e-procurement). Intra-organizational integration is the integration of applications within the organization, attempting to integrate custom applications and packaged systems.

Lammer *et al* [13] acknowledged multiplicity of integration approaches and concepts, differentiated by integration's level and architecture. However, Spackman *et al* [14] and Chandra *et al* [15] argued that beyond the marketing hype; "*there is a great deal of overlap between the integration approaches*". Thus, at highest level of integration, while devising integration strategy for the enterprise, there are limited numbers of overall approaches available [14]. EI approaches can therefore be broadly narrowed down into four basic categories, including [14, 15]: (i) Point-to-point based integration; (ii) ERP integration; (iii) Enterprise Application Integration (EAI) approach; (iv) and Service oriented integration approach.

Thus, during the early phases of EI initiatives planning, determination of a strategic integration approach is among the most important and challenging planning task. The decision outcome of this task carries a great positive or negative impact to most of the subsequent development activities, and consequently overall success and quality of the resulting system. However, this decision making task is generally complicated as it involves different participants, contributing in decision making process and who have diverse goals, criteria and perception about certain alternatives. The activity is further challenging as it comes in early stages of project lifecycle, hence hard to reason about the consequences of the decision made [16].

2.1 EI Adoption drivers

It is well accepted that EI adoption is a strategic investment [17], and should therefore be closely linked to organization's vision, goals and strategies. It is an endeavor involving high expenditure, risk and greatly impacts almost every organization's aspects. Different organizations adopt EI initiatives motivated by completely different integration drivers. However, the initial rationale for EI adoption influences problem definition, methods of achieving goals and other subsequent activities [18]. According to Themistocleous [19], many organizations are motivated to adopt integration initiatives based on organizational, operational, technical and strategic perceived benefits. On the other hand, based on some empirical studies, Lam *et al* [20] and Puschmann *et al* [21] identified several integration drivers and broadly categorized them either as; organizational, project, internal or external drivers. These drivers and benefits should be taken into consideration during EI strategy formulation, and hence during the selection process of an appropriate integration approach [6, 7, 19].

As EI involves an attempt to connect a number of different types of software products, the selection of integration approach should not only be based on functional and business requirements, but should also take into consideration non-functional and other technical requirements. Non-functional requirements describe properties that a system should possess. They are defined by stakeholders and greatly influence adoption of a particular integration approach. Non-functional requirements should not be overlooked, since doing so could lead to a poor system quality, unsatisfied stakeholders and eventually unsuccessful investment [16, 22]. Spackman *et al* [14] proposed a Software Quality Attribute Trading (SQUAT) technique, which takes into consideration known software qualities attributes to evaluate integration approaches and technologies. Silveira *et al* [12] developed EAI evaluation tool by further characterizing and describing international software quality evaluation standard (ISO-9126) for integration tools. Some of these quality attributes includes; adaptability, platform Neutrality, scalability, security, reliability, modifiability, performance, interoperability, maintainability, flexibility and testability

In general, selection and evaluation of integration approaches should take into consideration multiple factors, which include business, organizational, project and technical related factors, as shown in Figure 1.

3. Methodology

Evaluation and selection of integration approach is a complex and challenging process. It involves consideration of several organizational factors, from both technical and business perspective. Moreover, the process consists of several contributors representing different levels of organizational functions and with varying goals and preferences. Thus, the problem requires an elaborate and systematic approach that would allow participants involved in the process to evaluate different alternatives by considering and expressing their preferences on multiple factors. That been the case then, application of multi-criteria decision analysis (MCDA) techniques is warranted.

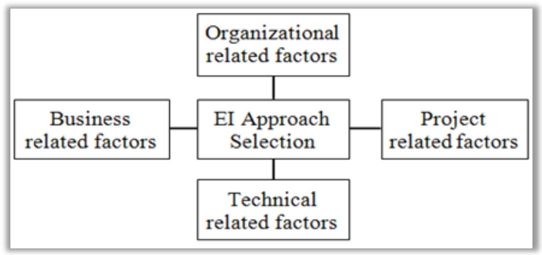


Figure 1: EI approach factors considerations [23]

A plethora of MCDA methods exists, where most common includes; Analytic Hierarchy Process (AHP), Analytic Network Process (ANP), Simple Additive Weight (SAW), Simple Multi-Attribute Rating Technique (SMART), Elimination and Choice Translating Reality (ELECTRE) and Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) [24]. Guitouni [25] argued that the purpose of every MCDA method is to assist in making a good recommendation. However, not all MCDA methods produce a good recommendation for all situations; since every method’s theoretical and axiomatic development is based on some assumptions and hypothesis. Given the nature of EI approach selection problem, ANP is deemed as the most appropriate method to utilize during the evaluation process. ANP is preferred for this case due to its ability to provide an elaborate structure that allows for an intensive analysis on problem elements and the resolution of inter-dependency characteristics among the elements. However, ANP framework does not fully support group decision making process. Since the EI approach selection process involves a group of contributors with varying views and preferences, a sense of consensus among the contributors is an important element in ensuring the process overall success. It is therefore posited that incorporation of Delphi technique in ANP framework, illustrated in Figure 2, will facilitate vigorous involvement of process contributors and bring forth a sense of unanimity among them. More discussion about ANP can be found in [9, 26] and on Delphi technique in [27-29].

The Delphi-ANP methodology presented in this article consists of four Delphi rounds, which are blended in with ANP procedure to yield the following eight steps; (i) Problem formulation and factor identification; (ii) Forming evaluation panel; (iii) Defining evaluation Criteria; (iv) Structuring the problem network model; (v) problem elements appraisal; (vi) judgment weights analysis; (vii) elements ranking approval; (viii) and finally the best alternative selection (See Figure 2 below)

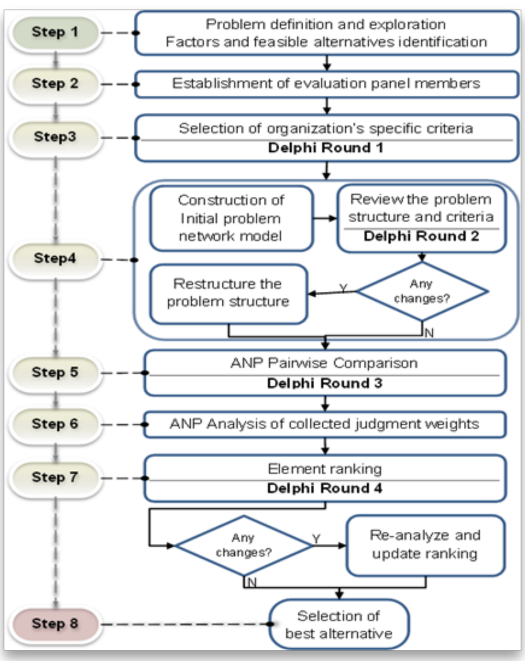


Figure 2: Delphi-ANP evaluation Methodology

To illustrate the ANP-Delphi methodology, this paper will consider an actual case study in a state government agency, referred hereinafter as TGA (real name withheld).

4. Case study: Application of Delphi-ANP methodology

The Delphi-ANP methodology was applied and validated at TGA, a state government agency in east-coast of USA. TGA is composed of several organizational functions and manages several engineering and transportation facilities across the state. However, many of these functions are supported by different systems for their day-to-day operations. Unfortunately, most of these systems exist within technical and functional walls of separation that creates barriers to integration and sharing of data; and thus inefficient business processing. Accordingly, the agency has taken several measures to address its integration challenges as an effort to achieve its strategic goals of effectiveness, efficiency and service improvement. At the time of this study, TGA was embarked in efforts to integrate some of its backbone information systems, such as those used in agency's financial management, supply chain management and maintenance operations. In other similar integration projects, TGA has successfully developed in-house integration solutions by either establishing direct interfaces between applications, replicating data between databases or simply databases-consolidation. However, for the particular initiative considered in this study, any of these traditional approaches may not have been viable option for several reasons. First, the systems intended to be integrated are the backbone of the agency's operations; and must therefore integrate with other systems in a different platform and located within and outside the agency's boundary. Secondly, these systems uses different data format, and thus the issue of data uniformity presented a major integration challenge. Finally, these systems are in constant upgrade to the latest versions, which also requires any direct interface existing between them to be upgraded or redeveloped, which is rather laborious and expensive.

Consequently, the management was seeking for a forward-looking approach that will allow future expansion of the integrated network system, and resonate with organization's strategic goals. The approach must also provide sound, cost effective and long-term solution that is efficient and meets all of the stipulated needs of the agency. As a start, some high level integration approaches were suggested. These suggested approaches consisted; (1) Implement a Service oriented based integration, (2) Adopt a vendor-based EAI adapter, (3) Implement a single enterprise system such as ERP that would either replace these systems or integrate them and (4) built in-house integration solution that establishes direct interface links between the systems.

Deciding on the approach to commit the SGA presented a great challenge to the management, as these approaches were perceived with varying strengths and weakness. The formulated Delphi-ANP methodology was therefore used in assisting the TGA management to decide the appropriate EI approach for their agency.

The Delphi-ANP methodology process involved the following steps;

Step 1: Problem formulation

The step comprised of preparatory and necessary tasks to help understanding the problem domain and set forth the evaluation process. The initial task is to set the objectives and goals of the evaluation process. Then, a review of literature, as well as interviews with domain experts was conducted in order to broadly identify feasible integration approaches and a list of criteria, which will possibly be used in gauging the identified alternatives.

Step 2: Evaluation panel establishment

This step involved setting up a team of participants to contribute in the evaluation process. In a Delphi evaluation process, a group of five to twenty willing experts [30, 31] is generally considered adequate. Thus, a panel of six members consisting of two IT planning manager, two project managers and two IT analysts was established. The members selected were believed to be knowledgeable about the integration subject and organizational requirements.

Step 3: Selection of organization's specific criteria factor

In this step, also marked as Delphi first round, the evaluation panel, formed in the previous step, was presented with an overview explaining the problem at hand, the purpose of the evaluation task, the expectations and procedures of the entire evaluation task process, and a list containing description of feasible alternatives and possible evaluation criteria. The panel was requested to select and justify factors that they would consider important in evaluating the presented alternatives. They were also requested for additional factors that may have been excluded from the list. From this round, the group identified a list of eighteen factors, categorized as strategic related, cost related, technical related and other organizational factors. The final list of the approved criteria is presented Table 1.

Table 1: Evaluation criteria for EI approaches

		Dependent on
Strategic related factors (SC)		
SC1	Approach alignment with business strategy	*TC1, TC6, SC2, SC3, SC4*
SC2	Approach ability to enhance cooperation with others	*TC1, OC2*
SC3	Approach ability to improve business agility	*TC1, TC6*
SC4	Approach ability to allow organization restructuring	
Technical related factors		
TC1	Scalability	
TC2	Maintainability	*TC3, TC4, TC5, TC6*
TC3	Maturity	
TC4	Security	
TC5	Complexity	
TC6	Extensibility	
Cost related factors		
CC1	Implementation cost	*OC5*
CC3	Maintenance cost	*TC2*
CC3	Re-engineering cost	
Other operational related factors		
OC1	Acceptability	
OC2	Approach ability to enhance process interoperability	
OC3	Approach ability to facilitate process management	*TC5, TC6*
OC4	Support/expertise availability	
OC5	Implementation time	

Step 4: Structuring problem into an ANP network model

In this step, the problem was configured to form initial ANP network structure, which consists of all the problem elements (i.e. the goal, criteria and alternatives) and also indicating any relationship existing between the elements.

This step also entailed Delphi second round, where the evaluation panel was presented with proposed ANP network structure as well as Delphi first round results, which contained other members' opinions (anonymous feedback) and any of their discrepancies. The members were given an opportunity to reconsider their judgments'. The results from this round were similar to the result from round one, and thus, consensus on evaluation criteria, presented in Table 1 and final problem structure formed, shown in Figure 3.

Step 5: Appraisal of problem's model elements

First, a questionnaire was developed to assess the relationships between the elements of the network structure, through pair-wise comparison. The questionnaire was instituted in MS-excel application so as to facilitate analysis, like checking consistency on the fly. Step five was also marked as Delphi round three, where the participants were presented with instructions on how to use excel-ANP template and the final problem network structure developed in the previous step. Using Saaty's 1-9 scale [9] shown in Table 2, the participants were requested to gauge (i) relative importance of criterions with respect to goal; (ii) relative preference of alternatives with respect to each criterion and; (iii) any relative influences existing between criterions within and outside the cluster.

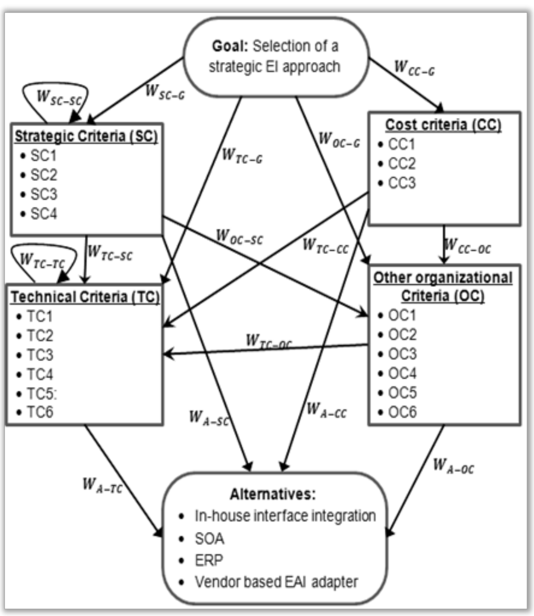


Figure 3: Integration approach selection problem network system

Step 6: Gathered judgment weights analysis

The pairwise comparison was done using MS-excel template, with in-built capability to calculate elements local priorities and pairwise-matrix consistency. Thus, the first task of the analysis step involved obtaining and organizing each panel member's pairwise comparison matrices, and their respective element priority vectors, such as in Figure 4. The pairwise matrices were also assessed for the completeness and consistency.

The element priorities were calculated using Saaty's [9] additive normalization algorithm, which consists of the following three steps:

For each pairwise comparison matrix A_c^m:

$$A_c^m = \begin{bmatrix} a_{11} & a_{12} & \cdots & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & \cdots & a_{nn} \end{bmatrix}$$

Where

A_c^m is nxn pairwise matrix from panel member m, for cluster c and containing relative measures, a_{ij}, assigned to element i when compared to element j (Note that $a_{ji} = \frac{1}{a_{ij}}$ and $i, j = (1,2, \dots, n)$)

i) Sum the values of each column in comparison matrix i.e. $\sum_{i=1}^{n} a_{ij}$

ii) Divide each element in matrix with respective column sum i.e. $\frac{a_{ij}}{\sum_{i=1}^{n} a_{ij}}$ to obtain normalized pairwise comparison matrix.

iii) Sum each row's element and divide row's sum by n elements in the row, $\frac{\sum_{j=1}^{n}\left(\frac{a_{ij}}{\sum_{i=1}^{n} a_{ij}}\right)}{n}$, to obtain vector (Eigen vector), w_c^m, containing estimates of relative priorities of the elements in the cluster c, i.e.

$$w_c^m = \begin{pmatrix} w_{1c}^m \\ \vdots \\ w_{ic}^m \\ \vdots \\ w_{nc}^m \end{pmatrix} \quad (1)$$

Where:

w_{ic}^m is the local priority for element i of cluster c, for panel member m

Table 2: Scale of Preferences between elements [9]

Saaty Scale	
1	Equal importance of both elements
3	Weak importance of one element over another
5	Strong importance of one element over another
7	Very strong importance of one element over another
9	Absolute importance of one element over another
2, 4, 6, 8 - Intermediate values	
Reciprocal Saaty Scale	
1/3	Slightly less importance of one element over another
1/5	Less importance of one element over another
1/7	Far less importance of one element over another
1/9	Absolute less importance of one element over another
1/2,1/4,1/6,1/8 - Intermediate values	

Technical Criteria w.r.t. to Goal	TC1	TC2	TC3	TC4	TC5	TC6	W_{TC-G}	W'_{TC-G}
Scalability (TC1)	1	4	5	1/3	2	2	**0.217**	1.45
Maintainability (TC2)	1/4	1	4	1/4	2	2	**0.131**	0.813
Maturity (TC3)	1/5	1/4	1	1/5	1/2	1/3	**0.045**	0.279
Security (TC4)	3	4	5	1	6	6	**0.438**	2.85
Complexity (TC5)	1/2	1/2	2	1/6	1	1	**0.080**	0.506
Extensibility (TC6)	1/2	1/2	3	1/6	1	1	**0.088**	0.552
λ_{max} = 6.35 CI = 0.07 RI = 1.240							CR = 0.0565	

Figure 4: Pairwise Comparison matrix as built in Microsoft excel

The consistency of each pairwise comparison matrix was checked using Saaty's [9] consistency index (CI) calculated as shown by equations 2 below:

$$CR = \frac{CI}{RI} \qquad (2)$$

Where

RI = consistency random index of randomly generated reciprocal matrix form scale 1-9, (see Table 3), and

$$CI = \frac{\lambda_{max} - n}{n-1} \qquad (3)$$

Where,

$$\lambda_{max} = \left(\frac{1}{n}\right) * \left(\frac{W_{1c}^{m'}}{w_{1c}^{m}} + \cdots + \frac{W_{nc}^{m'}}{w_{nc}^{m}}\right)$$

and

$$W_c^{m'} = A_c^m \, w_c^m = \begin{pmatrix} W_{1c}^{m'} \\ \vdots \\ W_{ic}^{m'} \\ \vdots \\ W_{nc}^{m'} \end{pmatrix}$$

Table 3: Average Random Consistency [9]

Size of Matrix (n)	1	2	3	4	5	6	7	8	9	10
RI	0	0	0.58	0.9	1.12	1.24	1.32	1.41	1.45	1.49

Aggregating members' priorities: - Since the local element priorities varied from one panel member to another, there was a need to aggregate the results for all the participants. Saaty [9] argued that geometric mean can be used to correctly represent a group of experts' consensus. Thus, geometric mean, calculated using equation 4 below, was used to aggregate participants' element priority into a group element priority, shown in Figure 5.

$$W_{ic} = \sqrt[m]{\prod_{m=1}^{m} w_{ic}^{m}}$$

Where: W_{ic} = the group priority of element i in cluster c

Combining technical criteria elements priority vectors with respect to goal							
	Individual member priorities						Group priority
	m1	m2	m3	m4	m5	m6	
TC1	*0.22*	*0.19*	*0.15*	*0.17*	*0.12*	*0.11*	***0.16***
TC2	*0.13*	*0.2*	*0.1*	*0.19*	*0.22*	*0.13*	***0.16***
TC3	*0.05*	*0.06*	*0.07*	*0.04*	*0.09*	*0.1*	***0.07***
TC4	*0.44*	*0.36*	*0.32*	*0.42*	*0.29*	*0.46*	***0.38***
TC5	*0.08*	*0.09*	*0.26*	*0.05*	*0.16*	*0.09*	***0.11***
TC6	*0.08*	*0.1*	*0.1*	*0.13*	*0.12*	*0.11*	***0.11***

Figure 5: Combining of individual element priority using geometric mean

Building super-matrix – To resolve interdependencies characteristics existing between the elements of the problem network system and eventual derivation of final elements' priorities, 'supermatrix' concept proposed by Saaty [26] is applied. A supermatrix is partitioned matrix, where each supermatrix's block is composed of a set of relationships between elements as represented by the problem model. The elements' relative weights obtained in the preceding steps, which indicates relative importance and influence of elements with each other, are entered in their respective block of the supermatrix, to form un-weighted supermatrix. The un-weighted supermatrix's columns are not stochastic i.e. each column of supermatrix does not add to one. Thus, to obtain stochastic supermatrix, each block of un-weighted supermatrix is multiplied by corresponding cluster priority (for this study case, all the clusters had relative-equal weights). To obtain long-term stable set of weights, the stochastic supermatrix is successively raised with large powers, usually 2k+1 (where k is a large arbitrarily number), until the process converges to a limit supermatrix. The final priority weights of the elements are obtained by normalizing the limit supermatrix.

<u>*Step 7: Element Ranking Approval*</u>

In this step, marked as fourth and last Delphi round, a report describing the analysis results was prepared and sent to the evaluation participants for their review. The report contained all the elements ranking based on the final priority weights obtained from normalized limit supermatrix. The report underscored the elements' ranking, by indicating each alternative performance with respect to the criterions and each criterion relative importance with respect to the goal. Any major discrepancies between members' weights was highlighted, and members were given a last opportunity to revise and justify their pairwise relative judgments', especially where notable discrepancy exists. In this last Delphi round, two members revised some of their pairwise comparison, arguing that, while undergoing through the evaluation process, they have become more acquainted on EI related issues, and therefore felt the need to make adjustment to reflect their true judgments. The final element rankings are presented in Table 4, Figure 6, Figure 7 and Figure 8.

<u>*Step 8: Selection of best alternative*</u>

This step marked the end of the evaluation process. The final report, which included members' comments and changes requested from the previous step, was produced. The final report emphasized on the alternative ranking, and more in particular to the first ranked alternative, which is actually the selected alternative.

Rank	EI Approach	From Limit supermatrix	Cluster normalized
4	In-house application integration	0.093	0.215
2	SOA	0.112	0.258
3	ERP	0.101	0.233
1	EAI	0.127	0.293

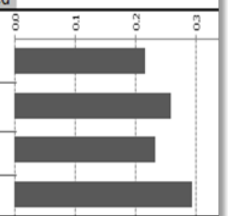


Figure 6: Alternative final score and ranking

5. Results Discussion

The use of Delphi-AHP evaluation approach provided a systematic framework in selection and evaluation of integration approaches. From the case study, EAI was ranked as the most appropriate approach to commit TGA (See Figure 6).

The Delphi-ANP analysis presented members with comprehensive breakdown on how they reached to the final decision, by prioritizing factors with respect to goal, in Figure 7, and by how each alternative relatively support different factors, as shown by the radar chart in Figure 8.

Factors' Ranking		
1	Implementation cost	12.8%
2	Implementation time	10.9%
3	Scalability	7.8%
4	Security	7.6%
5	Business agility	6.8%
6	Process interoperability	6.0%
7	Extensibility	5.7%
8	Overall alignment	5.3%
9	Cooperation	5.3%
10	Maintainability	5.1%
11	process management	4.6%
12	Complexity	4.3%
13	Maintenance cost	4.3%
14	Support/expertsise availability	4.1%
15	org. restructuring	3.3%
16	Maturity	2.3%
17	Re-engineering cost	1.9%
18	Acceptability	1.6%

Figure 7: Criteria ranked by their % contribution to overall decision

The case study results also showed that implementation cost was the major factor influencing the final decision, scoring 12.8%, followed by implementation time with 12%. Scalability and security were considered as the major technical factors with each scoring 7.8% and 7.6% respectively. In strategic perspective, business agility, EI approach overall alignment with organization strategy and cooperation with other business partners are leading factors, each scoring 6.8%, 5.3%and 5.3%, respectively.

Overall, factors deemed to enhance organization interoperability were well favored; although ranked behind the factors related to approach implementation and security. These results are presented in Table 4, Figure 6 and Figure 7. EAI approach scored average in almost every criterion considered. SOA approach was viewed as the most outstanding approach as far as organization's strategy is concerned, but performed relatively poor in terms of cost related factors. ERP scored relatively well in technical related factors, especially on issues related to security, maintainability and complexity. However, it was viewed as an expensive approach.

Table 4: Alternative performance with respect to criteria elements

	in-house int.	SOA	ERP	EAI	
Overall alignment *(SC1)*	0.4%	2.3%	1.4%	1.4%	
Cooperation *(SC2)*	0.3%	2.8%	1.0%	1.4%	
Business agility *(SC3)*	0.4%	2.7%	0.9%	1.6%	
org. restructuring *(SC4)*	0.6%	1.1%	2.8%	1.2%	
Stategic Criteria score	*1.7%*	*8.9%*	*6.0%*	*5.6%*	∑= 22.2%
Implementation cost *(CC1)*	2.5%	0.8%	0.4%	1.8%	
Maintenance cost *(CC2)*	0.6%	1.0%	2.3%	1.7%	
Re-engineering cost *(CC3)*	3.2%	0.6%	0.4%	1.4%	
Cost related criteria score	*6.3%*	*2.3%*	*3.1%*	*5.0%*	∑= 16.7%
Acceptability *(OC1)*	3.2%	0.9%	0.4%	1.1%	
process interoperability *(OC2)*	0.2%	2.6%	0.9%	1.8%	
process management *(OC3)*	0.4%	1.3%	2.2%	1.6%	
Support availability *(OC4)*	2.3%	0.5%	0.6%	2.2%	
Implementation time *(OC5)*	2.3%	0.7%	0.4%	2.1%	
Other/organization criteria Score	*8.4%*	*6.0%*	*4.6%*	*8.8%*	∑= 27.8%
Scalability *(TC1)*	0.4%	3.3%	0.7%	1.2%	
Maintainability *(TC2)*	0.7%	1.2%	2.0%	1.7%	
Maturity *(TC3)*	1.9%	0.9%	0.9%	1.9%	
Security *(TC4)*	1.0%	0.7%	2.8%	1.0%	
Complexity *(TC5)*	0.3%	0.8%	2.2%	2.2%	
Extensibility *(TC6)*	0.3%	2.0%	1.0%	2.3%	
Technical Criteria Score	*4.6%*	*8.8%*	*9.6%*	*10.3%*	∑= 33.3%
					∑= 100.0%

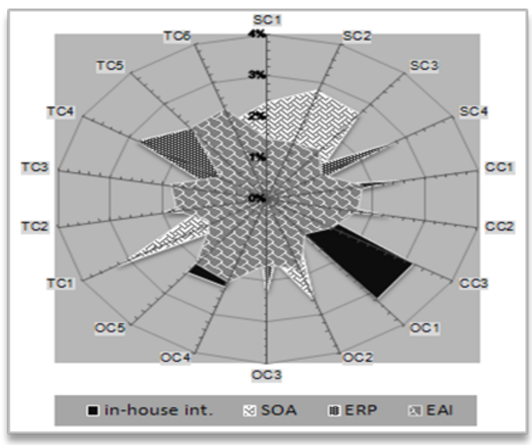


Figure 8: Radar Chart showing alternatives performance with respect to criteria

6. Conclusion and future work

Decisions related to EI implementation and technologies are unstructured and complex, due to consideration of wide range factors, and involvement of different stakeholders. This paper demonstrated a Delphi-ANP evaluation approach, which systematically considers multiple factors and improves the ANP decision process by incorporating Delphi technique. The Delphi technique improved the process by first, allowing selection of requirements considered by the participants. Many factors included in the problem network mean more pair-wise comparison, and this may affect the process ability. The Delphi feedback also helped in improving consensus in the decision making process.

The use of the Delphi-ANP approach allows EI selection to be logically and systematically structured within the context of the organization's strategic vision. It enables decision makers to comprehensively consider the strengths and weakness of each EI strategy against various criteria and sub criteria. Although the process is quite demanding, its adoption allows productive and informative participation of different stakeholders with varying perspective to arrive at a consensus with proper documentation for such a complex decision.

Further validation of this model in different environments and with existing decision making methods is encouraged. Additionally, sensitivity analysis to determine robustness of the decision should be performed.

7. References

[1] R. Myilsamy, "Event Driven Data Integration an Analysis," *Journal of Global Research in Computer Science,* vol. 2, p. 3, 2011

[2] C. Ling and L. Xin, "Achieving Business Agility by Integrating SOA and BPM Technology," in *Information Technology and Applications (IFITA)*, Kunming, China 2009, pp. 334-337.

[3] W. Lam and V. Shankararaman, "An enterprise integration methodology," *IT Professional,* vol. 6, pp. 40-48, March-April 2004.

[4] Gartner-inc, "Gartner EXP Worldwide Survey of More than 1,500 CIOs Shows IT Spending to Be Flat in 2009," ed: Gartner Newsroom, 2009.

[5] Nascio, "State CIO Priorities 2009," ed: National Association of State Chief Information Officers Website, 2009.

[6] M. Themistocleous, "Justifying The Decisions For EAI Implementations: A Validated Proposition Of Influential Factors," *Journal of Enterprise Information Management,* vol. 17, p. 85, 2004.

[7] M. Janssen and A. M. Cresswell, "Enterprise Architecture Integration in E-government," in *38th Hawaii International Conference on System Sciences*, 2005.

[8] R. E. Giachetiti, "A Framework to Review the Information integration of Enterpise," *International Journal of Production Research,* vol. 42, pp. 1147-1166, 2004.

[9] T. L. Saaty, *The Analytic Hierarchy Process*. New York, NY: McGraw-Hill, 1980.

[10] D. W. Frye and T. R. Gulledge, "End-to-End Business Process Scenarios," *Industrial Management \& Data Systems,* vol. 107, pp. 749-761, 2007.

[11] D. Chen*, et al.*, "Architectures for enterprise integration and interoperability: Past, present and future," *Computers in Industry,* vol. 59, pp. 647-659, September 2008.

[12] R. Silveira and J. A. Pastor, "A Model For Enterprise Application Integration Tools Evaluation," in *European and Mediterranean Conference on Information Systems (EMCIS)*, Costa Blanca, Alicante, Spain, 2006.

[13] A. Lämmer*, et al.*, "A Procedure Model for a SOA - Based Integration of Enterprise Systems," *International Journal of Enterprise Information Systems,* vol. 4, pp. 1-12, 2008.

[14] D. Spackman and M. Speaker, *Enterprise Integration Solutions*: Microsoft Press, 2004.

[15] S. Chandra and R. Juarez, "A Practical Approach to Enterprise Integration". *Defense AT&L,* 2009.

[16] T. Al-Naeem*, et al.*, "A Quality-Driven Systematic Approach for Architecting Distributed Software Applications," ed, 2005.

[17] A. R. Venkatachalam, "A Holistic Perspective on Enterprise Integration," *Journal of Information Technology Case and Application Research,* vol. 8, 2006.

[18] C.-C. Wei*, et al.*, "An AHP-Based Approach to ERP System Selection," *Int. J. Production Economics,* vol. 96, pp. 47-62, 2005.

[19] M. Themistocleous and Z. Irani, "Benchmarking the Benefits and Barriers of Application Integration," *Benchmarking: An International Journa,* vol. 8, pp. 317 - 331, 2001.

[20] W. Lam, *Enterprise Architecture and Integration: Methods, Implementation, and Technologies*: Idea Group Inc (IGI), 2007.

[21] T. Puschmann and R. Alt, "Enterprise Application Integration - The Case of the Robert Bosch Group," *Proceedings of the 34th Hawaii International Conference on System Sciences,* vol. 34, 2001.

[22] T. Al-Naeem*, et al.*, "Systematic Approaches for Designing B2B Applications," *International Journal of Electronic Commerce,* vol. 9, pp. 41 - 70, 2004.

[23] J. Ngeru*, et al.*, "Developing an Enterprise Integration Strategy Using an AHP-Delphi Approach," *Journal of Management and Engineering Integration* vol. 2, pp. 92-101, 2009.

[24] E. Triantaphyllou, *Multi-Criteria Decision Making Methods: A comparative Study*, 1 ed.: Springer, 2000.

[25] A. Guitouni, "Decision-aid to improve organisational performance," *European Journal of Operational Research,* vol. 162, pp. 581-583 2004.

[26] T. L. Saaty, *Decision Making with Dependence and Feedback: The Analytic Network Process*. Pittsburgh: RWS Publications, 1996.

[27] B. B. Brown, "Delphi Process: A Methodology Used for the Elicitation of opinions of Experts," ed, 1968.

[28] F. T. Hartman*, et al.*, "The Delphi Method for Graduate Research," *Journal of Information Technology Education,* vol. 6, pp. 1-21, 2007.

[29] C. Okoli and S. D. Pawlowski, "The Delphi method as a research tool: an example, design considerations and applications," *Information & Management,* vol. 42, pp. 15-29, 2004.

[30] G. Rowe and G. Wright, "Expert Opinions in Forecasting. Role of the Delphi Technique," in *Principles of forecasting: a handbook for researchers and practitioners*, J. S. Armstrong, Ed., ed Boston: Kluwer Academic Publishers, 2001.

[31] J. S. Armstrong, *Principles of forecasting: a handbook for researchers and practitioners*: Springer, 2001.

The Use of Analytical Hierarchy Process (AHP) to Design a Healthcare System

*J. S. Sutterfield, Charles Evans, Dominique S. Crawford and Brandon L. Clark
School of Business and Industry, Florida A&M University, Tallahassee, FL 32307, USA
*Corresponding author: pisces4914@earthlink.net

ABSTRACT

Although not often thought of as a production system, healthcare is nonetheless one type of production system, in particular a service production system. In order to design any system, hardware or service, it is first necessary to select the features that are to comprise the system. One such approach that might be used is Quality Function Deployment (QFD). Another is Analytical Hierarchy Process (AHP). In this paper, we demonstrate how AHP might be used to determine the features of a healthcare service system. In practice, the results would then be used as the basis for conceptual design of a healthcare production service system.

1. INTRODUCTION

A production system is an aggregation and arrangement of machines, processes, etc., in general production resources, so as to maximize the value added to some product through a transformation process (Russell and Taylor, 2003). Early in the 20th century the *Production Operations Management* function began to come under the emerging discipline of scientific management, as enunciated by such pioneers as Frederick Taylor, Frank and Marian Gilbreth (of *Cheaper by the Dozen* fame), Henry Gantt and others (Martinich, 1997). It soon began evident that the use of scientific principles could be employed in the design and construction of production systems to substantially increase profits.

One of the most important insights to come out of the quality movement was that services can be viewed and analyzed as production systems. A robust service system can provide significant added value to the customer (Chase, Jacobs and Aquilano, 2006). Just as it is possible to analyze product oriented production systems with scientific management techniques, it is equally possible to treat service oriented production systems with such analytical techniques.

If a service system is viewed as a process, or sequence of processes, that adds intangible value, it immediately becomes evident that a healthcare system qualifies as a service system. For example, someone undergoing successful cataract surgery emerges with value added in the form of restored vision and an improved quality of life. Or again, someone undergoing a successful appendectomy is able to return to a normal life with restored health. However, before a healthcare system can deliver its intended services it is necessary to design that system. Further, the design of the healthcare system must be founded in the design of a philosophical framework within which to design the delivery system. Thus, the basic philosophical framework must be established first from multiple variables, originating in philosophical principles that sometimes are in competition with other.

Numerous decision situations involve multiple variables and objectives that are in competition with each other. Situations in which such variables can be quantified are rather commonplace. Methods for solving such problems have been used for decades. One simply devises some measure for the variables and the objectives involved, after which the relevant computational methodology is selected. Then, after the necessary data is obtained, an answer can be calculated. Frequently, however, decision problems involve variables and objectives than do not have a direct measured, or at best can be only partly measured. In such problems the solution approach is not always so obvious. This is particularly true when the variables and objectives involve personal preferences. The approach frequently taken with such problems is to simply prioritize the decision considerations and to try to choose a solution that maximizes the desired decision quantities at minimum cost. Although this may not be too difficult with a limited number of decision quantities, it can become very difficult when the number of such quantities is large. In addition, the problem becomes vastly more complex when some of the decision quantities are in mutual conflict. Thus, making rational decisions under such circumstances may become extraordinarily difficult.

The type of decision problems discussed above are attacked with a class of techniques known as multi-attribute decision analysis. Unfortunately, most of these require that the attributes be measurable. When the attributes are of more qualitative in nature, the multi-attribute problem becomes much more difficult to handle. One technique available for solving problems of a qualitative nature is Analytical Hierarchy Process (AHP). With AHP it is possible to give a qualitative type problem a quasi-quantitative structure, and to arrive at decisions by testing as to whether the selections have been treated with rational consistency.

In this paper, we first administer an instrument to determine which features of are preferred in a healthcare system. Next, AHP is used to analyze the responses to determine which might be desirable to incorporate into a healthcare service production system. This is done by first converting the instrument responses into a partially complete paired comparison matrix. Next, the logical principle of inverse implication is used to complete the paired comparison matrix. Matrix operations are then used to arrive at a principal vector. A consistency ratio is eventually computed to determine how rigorously rational consistency has been maintained in the analysis. Upon determining the consistency of the respondent selections, the features to be provided by the system are selected. In practice, the next phase would involve implementing these features into the actual healthcare service production system.

Although QFD might just as well have been chosen for this initial design phase, it would have resulted in an entirely different process. Hence, it was considered preferable for our purposes to employ AHP.

2. LITERATURE REVIEW

AHP was originally conceived by Thomas L. Saaty as a structured method for solving problems involving decision variables or decision attributes, at least some of which, are qualitative, and cannot be directly measured (Saaty, 1980). Its ability to address a wide range of problems has met with extremely wide acceptance. Since its inception, it has been applied to a wide range of executive decisions, including those involving conflicts in stakeholder requirements (Saaty, 1982). The real power of AHP consists in its use of fairly elementary mathematics to structure complex problems in which decisions involve numerous decision makers, and multiple decision variables. Another facet of the power of the AHP approach consists in its ability to impose a quasi-quantitative character on decision problems in which the decision variables are not necessarily quantitative. The power and versatility of AHP are demonstrated by the wide range of problems to which the approach has been applied. It was used early for such problems as the justification of flexible manufacturing systems (Canada and Sullivan, 1989), and continues to be used in such applications (Chan and Abhary, 1996; Chandra and Kodali, 1998; Albayrakoglu, 1996). It has been used in such widely different applications as business crisis management (Lee and Harrald, 1999) and pavement maintenance (Ramadhan, et al, 1999). Other interesting applications of AHP include the evaluation of personnel during the hiring process (Taylor, et al, 1998), determination of investor suitability in structuring capital investment partnerships (Bolster, et al, 1995), apportioning public sector funds where numerous projects usually compete for limited resources (Barbarosoglu, and Pinhas, 1995), and determination of real estate underwriting factors in the underwriting industry (Norris and Nelson, 1992). Thus, AHP is a very powerful, versatile and generalized approach for analyzing multi-attribute decision problems in which the decision considerations do not necessarily need to be directly measurable. Although AHP has been used in a wide range of applications, this literature search has not disclosed any in which it has been used for the design of a healthcare system.

Since its' original statement by Saaty (Canada and Sullivan, 1980) the approach in using AHP in multi-attribute decision applications has been very well defined. Some few refinements and extensions have been added, but the original work remains unmodified. However, one such extension is that devised by C. Muralidharan, et al, (2003) in which Saaty's test for consistency is supplemented by constructing a confidence interval for the responses. This work discusses the consistency ratio (C. R.) devised by Saaty to make certain that attribute ratings were internally coherent. Saaty recommended that the C. R. be no greater than 0.1. The work then points out that although all correct decisions are consistent, not all consistent decisions are correct. It then proceeds to employ the student's "t" distribution to determine a confidence interval within which responses should lie in order to be mutually consistent. The "t" distribution is, of course, employed because the number of participants in an AHP analysis is normally fewer than twenty, this being the lower limit for using the normal distribution. For this analysis the number of respondents was well over the minimum required for a Student's "t" correction.

3. AHP METHODOLOGY

As noted above, the procedure for using Analytical Hierarchy Process is well established. It is a tribute to Dr. Saaty that his original work, done nearly a quarter century ago, remains virtually unmodified. Thus, the present work will follow rather closely Dr. Saaty's original approach. This statement involves the following steps:

1) A clear concise description of the decision objective. In practice, this is probably best done by a team of 4 to 6 people who have good knowledge of the objective to be achieved, and who have a stake in arriving at the best possible decision.
2) Identification of those attributes that are to be included in arriving at the desired objective defined in step 1. These attributes are also best identified by a team of 4-6 people and preferably the same team that is used to identify the objective of the analysis.
3) Determination of any sub-attributes upon which an attribute might be based.
4) Identification of a set of alternatives that are thought to achieve, at least partially, the desired objective. We say "at least partially" because probably no alternative will completely provide all desired attributes. However, the alternatives should be selected because of providing some degree of satisfaction to all attributes.
5) Once the attributes are identified, they are entered into a preference matrix, and a preference or importance number is assigned to reflect the preference for/importance of each attribute relative to all others. The strength of preference or importance is indicated by assigning a preference or importance

number according to the rating scale shown in Table 1.

Table 1: AHP Strength of preference rating scale

Intermediate degrees of preference for attribute A over attribute B are reflected by assigning even numbered ratings "8", "6", "4" and "2" for one attribute over another. For example, assigning a preference number of "6" would indicate a preference for attribute "A" over attribute "B" between "Strongly more important or preferred" and "Very strongly more important or preferred." Also, the logical principle of inverse inference is used, in that assigning attribute "A" a preference rating of, say "8", over attribute "B", would indicate that attribute "B" were only "1/8" as important as attribute "A." Either such a preference matrix is developed by each participant in the process, or the participants arrive at agreement as to the preference numbers for the attributes.

6) Mathematical operations are then used to normalize the preference matrix and to obtain for it a principal or characteristic vector.
7) Next, participants collectively rate each of the possible approaches or options for satisfying the desired objective(s). This is done along the same lines as shown above for attributes. The same rating scale is used, except here the ratings reflect how well each possible approach or option satisfies each of the attributes. For example, suppose that Alternatives "1" and "2" were being compared as to how well each satisfies Attribute "A" relative to the other. If Alternative "1" were given a preference rating of, say 5, relative to Alternative "2" it would mean that for Attribute "A", Alternative "1" were thought to satisfy it 5 times as well as Alternative "2". Or conversely, Alternative "2" were believed to satisfy Attribute "A" only "1/5" as well as did Alternative "1". Thus, a set of preference matrices is developed, one matrix for each attribute, that rates each alternative as to how well is satisfies the given attribute relative to all other alternatives under consideration.
8) Again, as with the attribute preference matrices, mathematical operations are used to normalize each preference matrix for alternatives, and to obtain a principal or eigenvector for each. When complete, this provides a set of characteristic vectors comprised of one for each attribute. Each of these vectors reflects just how well a given alternative/option satisfies each attribute relative to other alternatives.
9) Then with an eigenvector from "8" above, we cast these into yet another matrix, in which the number of rows equals the number of alternatives/options, and the number of columns equals the decision attributes. This matrix measures how well each alternative or option satisfies each decision attribute.
10) Next, the matrix from "9" above is multiplied by the characteristic vector from "6" above. The result of this multiplication is a weighted rating for each alternative/option indicating how well it satisfies each attribute. That alternative or option with the greatest weighted score is generally the best choice for satisfying the decision attributes, and thus achieving the desired objective.

11) Last, it is necessary to calculate a consistency ratio (C. R.) to ensure that the preference choices in the problem have been made with logical coherence. The procedure for calculating the C. R. may be found in a number of works on AHP[15] and will not be repeated here. According to Saaty, if choices have been made consistently, the C. R. should not exceed 0.10. If it does, it then becomes necessary to refine the analysis by having the participants revise their preferences and re-computing all of the above. Thus, the entire process is repeated until the C. R. is equal to or less than 0.10.
12) Having completed steps "1" thru "11", that alternative/option with the greatest weighted score is selected as best satisfying the decision alternatives, and thus achieving the desired objective.

4. APPLICATION OF METHODOLOGY

For the current analysis, twenty of the principal features were taken from the new national healthcare legislation, and 10 respondents were asked to identify the top ten features that each would prefer in a healthcare system. Next, the top ten features were cast into an AHP preference matrix, and forty more respondents were asked to rate each of the top ten features relative to the remaining nine. Once this was complete, this set of forty ratings was averaged into a single set of ratings for purposes of analysis. This AHP preference of averaged ratings is shown in Table 2.

Table 2: AHP averaged preference ratings

A number of observations can be made about the values in Table 2. First of all, it will be seen that the diagonal in the table is comprised of "1s." These ratings represent the preference of a feature relative to itself. Since it is logically absurd to speak of preferring a feature relative to itself, it can only be said that a feature is equally preferred to itself. In the first row for feature "A," Public Option," is preferred to feature "B," Healthcare Rationing," by a factor of 4. Further, in the second row for feature "B," it will be found to be preferred to feature "A" by a factor of ¼ or 0.25. Thus, the preference for feature "B" relative to feature "A" is obtained by taking the reciprocal of the preference for feature "A" relative to feature "B." It will be seen the then that the feature rating system makes use of the logical relation of inverse implication.

One final detail must be pointed out concerning the handling of the averaged data. The AHP preference rating scale ranges from 9 thru 1 thru 9, that is to say from rating feature A as strongly preferred to feature B, to Feature B as strongly preferred to feature A. The middle of the AHP preference scale is of course 1. On the other hand the common number scale is centered at 0. Thus, before the respondents preferences could be averaged, it was necessary to deduct 1 from each of them to transform them to the common number scale, compute the averages, and then transform each average back to the AHP preference rating scale by adding 1 back to each average. This operation yielded the values shown in Table 2.

After the original respondent data were compiled to yield the preference matrix shown in Table 2, the data were then normalized in each column by summing the values in the column, and then dividing each value in that column by the sum. Further, each row in the normalized AHP matrix was summed across and divided by 10, the number of features in the row. This operation yielded the principal vector for the respondent data. The results of these operations are shown in Table 3.

Table 3: Normalized AHP preference ratings along with principal vector

It will be observed that the components of the principal vector sum to 1.

It should be pointed out that one variation of the procedure for calculating the principal vector is to compute the geometric mean for each row. Next, these geometric means are summed, and each geometric mean is divided by their sum. As noted by Saaty (1996) this approach to calculating the principal vector provides only an approximation, and is useful when the amount of data is relatively small. Otherwise, the standard approach, as previously mentioned, is to be preferred. Further, it can be shown that in general the arithmetic mean is equal to or greater than the geometric mean. Thus, the arithmetic mean leads to a more demanding test for the consistency ratio (Selby, 1970).

Continuing, an AHP analysis relies upon determining an eigenvector (or characteristic vector) for the preference matrix. This is obtained with a matrix multiplication between the principal vector and each normalized row of the preference matrix. The results of this matrix multiplication are identified as the eigenvector shown in Table 4.

Table 4: Eigen-vector for AHP preference ratings

Next an intermediate vector, identified as **D is calculated**. This leads to the calculation for λ_{max}, the maximum eigen-value for the characteristic matrix. The magnitude of this vector is crucial to determining the consistency of selections, as it is an estimate of the maximum characteristic value that might be expected for a preference matrix with n-1 degrees of freedom. These calculations are shown in Table 5.

Table 5: Maximum eigen-value for AHP preference ratings

The calculation for λ_{max} is obtained from …

$$\lambda_{max} = \frac{\sum_{i}^{10} D}{10}$$

The calculation for the consistency index, denoted C.I., is obtained from …

$$CI = \frac{\lambda_{max} - 10}{10 - 1}$$

It will be seen that if choices were made perfectly consistently that λ_{max} would equal 10, which would make the C.I. zero. Thus, C.I. is a statistical measure of the dispersion resulting from inconsistent choices distributed over the number of degrees of freedom for the preferences. From Table 6 above, it will be seen that this calculation for our application yielded a C.I. of 0.09.

The last step of AHP is that of comparing the consistency index to that which would be expected if the choices had with say 90% consistency. This is called the consistency ratio, (C.R.), and is obtained by dividing the C.I. from above by a standard value for 90% consistency. The standardized values are shown below in table 6.

Table 6: Maximum random indicies for consistency of choice

Then the C.R. for our healthcare production system is obtained as …

$$CR = \frac{\ldots}{\ldots} = 006$$

6. ANALYSIS OF RESULTS

By Saaty's consistency criterion, the C.R. for a given application should not be greater than 0.10 for consistency. Thus, feature choices made by participants for the healthcare system yielded a C.R. well below that necessary for consistency. We are now prepared to select those features to be included and those to be excluded for our healthcare production system. Referring to Table 2, those features are chosen for inclusion having a sum greater than 10. Those having a sum of 10 or less are excluded. The rationale for this is that a feature having a row sum equal to 10 would indicate no preference for that feature, or in other words, a rating of 10 indicates indifference as to a feature. Using this criterion, the features shown in Figure 1 would be included in a healthcare production system:

Features to be included
A: Public Option
C: $500B in Medicare reductions
D: Income tax exclusion for premiums
E: Marriage and family counseling
G: Universal coverage
H: Policy portability
I: Coverage of pre-existing conditions

Figure 1: Features to be included in healthcare system

Further, the criterion stipulated above would also exclude the features shown in Figure 2.

Features to be excluded
B: Healthcare rationing
E: Enforced participation
J: Coverage of illegal aliens

Figure 2: Features to be excluded from healthcare system

The features selected by the sample population were very much as expected. Features H and A were strongly favored over all other options. Further, Features D, G and I also exhibited strong preferences. Feature C, a $500B reduction in Medicare to fund the new healthcare system, demands a bit of special explanation. For this feature, participants were not told what the implications of this feature might be. If this feature were included, it would mean that the parents and grandparents of participants would likely receive a lesser level of healthcare in order that participants themselves might receive healthcare. Presumably, if this had been explained before administration of the test instrument, participants would not have favored this feature.

Feature J, Coverage of Illegal Aliens, was the least preferred of all of the ten features, and decisively rejected by the participants. This result was entirely contrary to the conventional wisdom about ensuring illegal aliens, and would have to be borne in mind in any healthcare system implementation.

7. CONCLUSIONS

The thesis of this paper has been that a healthcare system may be viewed as a service production system, and that Analytical Hierarchy Process may be used in the initial advanced design phase to identify the principal features necessary to complete the design. In an actual application, the features selected in the advanced design phase would be subjected to engineering design in order to conceive of the best hardware configuration for implementing the features selected. However, since the present application lies within the realm of public policy, the next phase would involve designing an administrative system to best implement the features selected using AHP. Once the processes were put in place via a regulatory mechanism, the system could be staffed with personnel and could begin to function. It would then be necessary for healthcare providers to offer any features required by public policy.

It had been thought initially that the number of respondents might be small, so that it might be necessary to test the results using a Student's-t distribution. As matters turned out the number of respondents was approximately 40, or more than enough to obviate the need for a Student's-t correction for a small sample size. For the application of a Student's-t correction in such applications, see Muralidharan, Anantharaman and Deshmukh (2003).

AHP is one of several multi-attribute approaches for making decisions in which the decision involves numerous variables or attributes. QFD is another such approach that might have been used, but would have required an entirely different approach to analysis. AHP was chosen for this paper, rather than QFD, because it better conformed with the purpose of the study.

There is currently an on-going discussion as to whether healthcare might better be delivered by a public system or by the private, free market system. The question of which is preferred for delivering healthcare service, public or private, is not addressed in this paper. Future research may focus upon using AHP to answer this question.

8. ACKNOWLEDGEMENTS

The authors wish to thank the students from the School of Business and Industry at Florida A&M University for participating in this project.

9. REFERENCES

[1] Albayrakoglu, M. Murat (1996). *Justification of New Manufacturing Technology: A Strategic approach Using the Analytical Hierarchy Process*, Production and Inventory Management Journal, Alexandria, First Quarter, Vol. 37, Issue 1, pgs. 71-77

[2] Barbarosoglu, Gulay, and David Pinhas (1995). *Capital Rationing in the Public Sector Using the Analytical Hierarchy Process*, The Engineering Economist, Norcross, Summer, vol. 40, Issue 4, pgs. 315-342

[3] Bolster, Paul J., and Vahan Janjigian, Emery A. Trahan (1995). *Determining Investor Suitability Using the Analytical HierarchyProcess*, Financial Analysts Journal, Charlottesville, July/Aug., Vol. 51, Issue 4, pgs. 63-76

[4] Canada, John R., and William G. Sullivan (1989). *Economic and Multiattribute Evaluation of Advanced Manufacturing Systems,* Prentice-Hall, Englewood cliffs, New Jersey, Ch. 10, pgs. 268 - 279

[5] Chan, Felix T. S., and K. Abhary (1996). *Design and Evaluation of Automated Cellular Manufacturing Systems with Simulation Modeling and the AHP Approach: A Case Study,* Integrated Manufacturing Systems, Bradford, Vol. 7, Issue 6, pg. 39

[6] Chandra, S. and Rambadu Kodali (1998). *Justification of Just-in-Time Manufacturing systems for Indian Industries*, Integrated Manufacturing Systems, Bradford, Vol. 9, Issue 5, pg. 314

[7] Lee, Young-Jai, and John R. Harrald (1999). *Critical Issue for Business Area Impact Analysis in Business Crisis Management: Analytical* Capability, Disaster Prevention and Management, Bradford, Vol. 8, Issue 3, pg. 184

[8] Muralidharan, C., N. Anantharaman and S. G. Deshmukh (2003). *Confidence Interval Approach to consistency Ratio Rule in the Application of Analytical Hierarchy Process*, West Indian Journal of Engineering, Republic of Trinidad and Tobago, Vol. 26, No.1, pgs. 17-28, July

[9] Norris, Daniel M., and Mark Nelson (1992). *Real Estate Loan Underwriting factors in the Insurance Industry*, Real Estate Finance, New York, Fall, Vol. 9, Issue, 3, pg. 79

[10] Ramadhan, Rezquallah H., Hamad I. Al-Abdul Wahab, and Salih O. Duffuaa (1999). *The Use of an Analytical Hierarchy Process in Pavement Maintenance Priority Ranking*, Journal of Quality in Maintenance Engineering, Bradford, Vol. 5, Issue 1, pg. 25

[11] Saaty, Thomas L. (1980). *The Analytical Hierarchy Process*, McGraw-Hill, New York, NY

[12] Saaty, Thomas L. (1982). *Decision Making for Leaders*, Wadsworth Publishing Company, Inc., Belmont, CA

[13] Saaty, Thomas L. (1996). *Decision Making for Leaders: The Analytical Hierarchy Process for Decisions in a Complex World*, The Analytical Hierarchy Process Series, vol, 2, pgs.71-74

[14] Taylor, Frank A. III, Allen F. Ketcham, and Darvin Hoffman (1998). *Personnel Evaluation with AHP*, Management Decision, London, Vol. 36, Issue 10, pg. 679

[15] Selby, Samuel M. (1970). *Standard Mathematical Tables*, The Chemical Rubber Co., Cleveland, OH, pg. 554-555

Preference Number for A is ….. Then Attribute A preferred …	…over (or to) Attribute B as …
9	Absolutely more important or preferred
7	Very strongly more important or preferred
5	Strongly more important or preferred
3	Weakly more important or preferred
1	Equally important or preferred

Table 1: AHP Strength of preference rating scale

	Identified as factor		A	B	C	D	E	F	G	H	I	J	Sums
Public option	A	A	1.00	4.00	3.00	1.00	6.00	2.00	0.33	1.00	0.33	4.00	22.67
Healthcare Rationing	B	B	0.25	1.00	0.50	0.50	2.00	0.50	0.33	0.25	0.25	4.00	9.58
$500B in Medicare reductions	C	C	0.33	2.00	1.00	0.50	3.00	2.00	0.50	0.33	0.50	3.00	13.17
Income tax exclusion for premiums	D	D	1.00	2.00	2.00	1.00	5.00	2.00	1.00	0.50	1.00	4.00	19.50
Enforced participation	E	E	0.17	0.50	0.33	0.20	1.00	0.50	0.33	0.20	0.50	2.00	5.73
Marriage and family counseling	F	F	0.50	2.00	0.50	0.50	2.00	1.00	0.50	0.33	0.50	3.00	10.83
Universal coverage	G	G	3.00	3.00	2.00	1.00	3.00	2.00	1.00	2.00	1.00	3.00	21.00
Policy portability	H	H	1.00	4.00	3.00	2.00	5.00	3.00	0.50	1.00	3.00	5.00	27.50
Coverage of pre-existing conditions	I	I	3.00	4.00	2.00	1.00	2.00	2.00	1.00	0.33	1.00	4.00	20.33
Coverage of illegal aliens	J	J	0.25	0.25	0.33	0.25	0.50	0.33	0.33	0.20	0.25	1.00	3.70
		Sum	10.50	22.75	14.67	7.95	29.50	15.33	5.83	6.15	8.33	33.00	

Table 2: AHP averaged preference ratings

	A	B	C	D	E	F	G	H	I	J	Sums	Principal vector
A	0.10	0.18	0.20	0.13	0.20	0.13	0.06	0.16	0.04	0.12	1.32	0.13
B	0.02	0.04	0.03	0.06	0.07	0.03	0.06	0.04	0.03	0.12	0.51	0.05
C	0.03	0.09	0.07	0.06	0.10	0.13	0.09	0.05	0.06	0.09	0.77	0.08
D	0.10	0.09	0.14	0.13	0.17	0.13	0.17	0.08	0.12	0.12	1.24	0.12
E	0.02	0.02	0.02	0.03	0.03	0.03	0.06	0.03	0.06	0.06	0.36	0.04
F	0.05	0.09	0.03	0.06	0.07	0.07	0.09	0.05	0.06	0.09	0.66	0.07
G	0.29	0.13	0.14	0.13	0.10	0.13	0.17	0.33	0.12	0.09	1.62	0.16
H	0.10	0.18	0.20	0.25	0.17	0.20	0.09	0.16	0.36	0.15	1.85	0.19
I	0.29	0.18	0.14	0.13	0.07	0.13	0.17	0.05	0.12	0.12	1.39	0.14
J	0.02	0.01	0.02	0.03	0.02	0.02	0.06	0.03	0.03	0.03	0.28	0.03
											Sum =	1.00

Table 3: Normalized AHP preference ratings along with principal vector

AHP preference matrix											Principal vector		Eigen vector
1.00	4.00	3.00	1.00	6.00	2.00	0.33	1.00	0.33	4.00		0.13		1.44
0.25	1.00	0.50	0.50	2.00	0.50	0.33	0.25	0.25	4.00		0.05		0.54
0.33	2.00	1.00	0.50	3.00	2.00	0.50	0.33	0.50	3.00		0.08		0.82
1.00	2.00	2.00	1.00	5.00	2.00	1.00	0.50	1.00	4.00		0.12		1.33
0.17	0.50	0.33	0.20	1.00	0.50	0.33	0.20	0.50	2.00	X	0.04	=	0.38
0.50	2.00	0.50	0.50	2.00	1.00	0.50	0.33	0.50	3.00		0.07		0.70
3.00	3.00	2.00	1.00	3.00	2.00	1.00	2.00	1.00	3.00		0.16		1.82
1.00	4.00	3.00	2.00	5.00	3.00	0.50	1.00	3.00	5.00		0.19		2.02
3.00	4.00	2.00	1.00	2.00	2.00	1.00	0.33	1.00	4.00		0.14		1.56
0.25	0.25	0.33	0.25	0.50	0.33	0.33	0.20	0.25	1.00		0.03		0.30

Table 4: Eigen-vector for AHP preference ratings

$$|D| = \left| \frac{1.44}{0.13} \quad \frac{0.54}{0.05} \quad \frac{0.82}{0.08} \quad \frac{1.33}{0.12} \quad \frac{0.38}{0.04} \quad \frac{0.70}{0.07} \quad \frac{1.82}{0.16} \quad \frac{2.02}{0.19} \quad \frac{1.56}{0.14} \quad \frac{0.30}{0.03} \right|$$

$$|D| = |\; 10.93 \quad 10.43 \quad 10.62 \quad 10.73 \quad 10.57 \quad 10.71 \quad 11.25 \quad 10.89 \quad 11.21 \quad 10.66 \;|$$

$$\lambda_{max} = 10.80$$

$$CI = 0.09$$

$$CR = 0.06$$

Table 5: Maximum eigen-value for AHP preference ratings

N	1	2	3	4	5	6	7	8	9	10	11
RI	0.00	0.00	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49	1.51

Table 6: Maximum random indicies for consistency of choice

Applications of Data Mining in Pharmaceutical Industry

John Wang[1], Bin Zhou[2], Jeffrey Hsu[3] and Ruben Xing[1]
[1]*Department of Management and Information Systems, Montclair State University, Montclair, NJ 07043, USA*
[2]*Department of Management, Kean University, Union, NJ 07083, USA*
[3]*Department of Information Systems and Decision Sciences, Fairleigh Dickinson University, Madison, NJ 07940, USA*
prof.johnwang@gmail.com, bzhou@kean.edu, jeff@fdu.edu, Xingr@mail.montclair.edu

Abstract

Data mining benefits every facet of the pharmaceutical industry. In this paper, we focus on three main areas where data mining has proven to be beneficial and effective for pharmaceutical companies. The business components, such as new drug development, sales, service and marketing, as well as inventory control and supply chain, are the main focus points.

1. Introduction

Data mining is simply gathering raw data and constructing that information into something quantifiable and meaningful to the end user. The industry is forever changing and it becomes essential to forecast and analyze all data to fully understand the trends and plan accordingly. Hand defines data mining as "the discovery of interesting, unexpected, or valuable structures in large datasets" [17]. Computers give companies the power to make connections within their data sets that would most likely never be discovered without the aid of complex algorithms.

Data mining is a unique tool that everyone in the business world is seeking to better understand and utilize. When it comes to the pharmaceutical industry, data mining has an edge towards identifying its weakness and developing new ideas. One main process towards developing new problems and solutions is to identify and target the different people in order to resolve and break down this industry [46]. Data mining is used widely in such industries as healthcare, manufacturing, telecommunication, and pharmaceutical. Lately, pharmaceutical companies have increased the use of data mining technology for drug discovery, clinical trials, drug diverse effects, research and development, drug safety, clinical testing, marketing, sales, supply-chain management, and inventory control [10][2]. We will discuss in detail each of these components and how data mining is helping to make each process more effective and powerful.

2. New Drugs from Beginning to the End

With the rapid changes taking place in the pharmaceutical industry, new drug discovery is always a hot topic. To develop any new drugs, it goes through four major steps: Starting from research stage, development stage, clinical trials stage, and post drug production stage. Once it goes through these major milestones, the drug is proposed for potential market segments. Let us understand each of these milestones and how data mining helps to create drugs from beginning until reaching out to customers.

2. 1. Drug Research/Discovery Stage

The pharmaceutical industry is highly competitive, and the research of new drugs is extremely expensive and time consuming. It usually takes 12-15 years to successfully complete the research and development process of new drugs. In addition to having a low success rate of less than 20%, a significant product today cost more than $700 million [8]. Data mining helps to see trends, irregularity, and risk during product research and development stage. Instead of trial and error, data mining can help find drugs that have desirable functionalities. It helps to reduce cost, time, and budget involved in research activities. The research stage in developing new drugs for the market is made easier through data mining. Instead of trial and error experiments, data mining software can help seek out the drug that has the desired activity. There is a vast amount of different side effects and uses of drugs, too overwhelming to sort through without the technological advancement of data mining [9].

When developing drugs, pharmaceutical companies use data mining to analyze and organize the vast amount of outcomes that are occurring during their drug experimenting stages. Data mining eases the process by sorting through the large amount of detailed information and picking out the relevant information. It then finds associations between patients, drugs and outcomes in a faster and more efficient way than when performed manually. Because data mining can record and analyze patterns affecting patients with similar attributes it can also determine what causes diseases in certain patients. Pharmaceutical companies are able to find out how the changes in an individual's DNA sequence affect the risk of developing common diseases [33].

The Geanissance Pharmaceuticals Inc. used data mining tools to analyze how genetic variations affect a person’s response to drugs and how likely someone will get sick. The company also uses data mining tools for their main research components such as cardiovascular disease, central nervous system disorders, and inflammation. BMS (Bristol-Myers) are committed to use the data mining technology. They applied the new approaches in order to help drive drug research process by tracking the information very closely and effectively [3].

2.2. Drug Development Stage

The data mining techniques that are used in developing new drugs are clustering, classification and neural networks. The main goal is to determine compounds with similar activity because compounds with similar activity may behave similarly. This should be performed when we have a known compound or when we do not have a known compound, but have desired activity and would like to find a compound that displays this activity. This can be achieved by clustering the molecules into groups via cluster analysis [39]. Data mining can answer questions such as what are new molecules and factors or what combinations are directly impacting drugs? Data mining provides major chunk of information in the form of structured data such as information pertaining to genes, proteins, diseases, organisms, and chemical substance etc. This can help combine vast amount of information into one place and build a successful predictive model.

Once the data have been clustered for development, the drug goes through the testing in animals and human tissues. The techniques that can be used are classification and neural networks [38]. Every pharmaceutical company has a goal to aid patients. This process will determine what negative and positive effect a drug has and which patients will be benefited from this drug. The clustered data information should be fed into a neural network that will predict a drug’s effectiveness.

Merck-Medco Managed Care is a mail-order business, which sells drugs to the country's largest health care providers like Blue Cross and Blue Shield. Merck explorded data mining to uncover hidden links between illnesses and known drug treatments, and spot trends that help pinpoint which drugs are the most effective for what types of patients. The results are more effective treatments that are also less costly. Merck's data mining project has helped customers save an average of 10-15% on prescription costs [24].

2.3. Drug Clinical Trials Stage

During the clinical trial stage, a company wants to make sure that a drug does not have too many adverse effects because it can be too dangerous and risky. With the help of data mining, one can feed the adverse effects of drugs into a neural network and let the network tell us what these effects are. The collected data or test results are usually submitted to the Food and Drug Administration (FDA) and inspected very closely. The purpose of clinical trials stage is to detect when too many adverse events occur and where is the link between drug and adverse effects. If the doctor prescribes a drug to a patient, he or she can then alert the patient for eventual side effects, or choose another drug if the side effects might be too dangerous for the specific patients. Using data mining techniques, we can find out which adverse events occur more frequently with specific drugs. Data mining protects patients and helps regulate drug testing [38].

Data mining technology, such as signal detection, is used for potential problems with the drugs. There are essentially three aspects to identifying a safety signal within these databases. First, occurrences of the event must be more frequent than would happen by chance. The data mining tools are primarily used *to determine* what the *problem* might be. Second, the event must be medically important or serious, and third, there must be a causal relationship to a drug. Data mining can be a very beneficial tool for companies to detect and study adverse events in already developed drugs and to measure the risk management [23]. Data mining reduces the risk because it contains information supporting the use of the drug and also contains training data about the negative and positive effects of the drug [41].

The FDA issued three safety guidelines in 2005. These guidelines provide the framework for pharmaceutical companies to analyze and report their data throughout the industry. The greatest data mining technologies such as adverse event (AE) provide signals and accurate drug risk assessment in pre-post clinical trial stages [34].

2.4. Post Drug Production Stage

When a drug is developed and sold at market places, the next and most important task is to manage patient and drug related information. Data mining has a tool, which enables companies to manage their patients' information easily and efficiently. User-interface may be designed to accept all kinds of information regarding the beginning up to the end of drug development. Information such as a patient's weight, sex, age, drug dosage, length of usage, foods consumed, and reactions should be reported. Then, based upon the information in the databases and the relevant data entered by the user, a list of warnings or known reactions should be reported. The user profiles can contain large amounts of information that can be easily obtained on a frequent basis. Also, the patient's profile should be recorded along with any adverse reactions reported by the patient, so that future correlations can be reported. Over time, the databases will become much larger, and interaction data for existing medicines will become more complete. The data mining tools are very useful in extracting patterns, and supporting various queries for past, present, and future inquires [14].

The identification and qualification of the information such as interactions among over-the-counter medications, prescription medications, interaction between medications and food, beverages, vitamins, minerals, and determining the adverse effect of particular medication for certain patients can be extremely useful for patients, physicians, pharmacists, health organizations, insurance companies, pharmaceutical manufacturers, and drug testing companies etc.

2.5. Identify and Target Individuals and Demographics

Data mining helps marketers understand the consumer's behaviors; tracking patient habits ultimately helps the companies find areas in the market where there is an unfulfilled need. The variable cost of physically producing the finished drug is relatively low when compared to the high fixed cost of developing the product. The majority of money spent on a new drug is through the extensive 4-stage process. If a company wishes to stay in business it must spend money in research and development. However, regardless of how desirable the new drug may be if it is not aggressively marketed to the appropriate demographic it will fail. Year after year, pharmaceutical companies spend billions of dollars solely on millions of visits of their sales representatives to physicians across the country [36]. For this reason companies yearn to differentiate themselves from the competition and create brand loyalty. It is understood that pharmaceutical companies in the 21st century focus on market potential of a drug rather than health needs of consumers [21]. Typically, a drug being discovered is a better alternative to another drug that already exists. By analyzing data of trends in the industry one can determine if there is a demand for a hybrid alternative to an existing drug. Data mining is a great way to determine if there is a need for a new drug in the market.

Sales figures within the pharmaceutical industry seem to be dropping for quite a few reasons. The innovation of new drugs is starting to slow down, key markets where "blockbuster drugs" are concentrated are experiencing increased competition with similar products, and price controls within the industry today have put a damper on *industrial companies'* margins and profits. Through careful data mining into these factors, some ideas have been discovered that may be able to fix the current problem of diminishing profits. The first is a value-based approach towards pricing which would have regulations that reduce the costs to buyers while providing reasonable incentives to researchers to discover new drugs and therefore help eliminate other companies from marketing similar drugs and simply marketing it better [19].

Law stated the importance of screening that has taken place inside labs that have applied the use of data mining [28]. In line with him, many people throughout the world are infected with chronic HBV (*Hepatitis B virus)*. The main reason of why many people do not know of the infected virus inside them is that it has a non-symptom effect on the body. Many places like Taiwan are now practicing routine checkups upon entering school, the military and other public places. By doing routine check-up and screening, we can help ourselves before it is too late. Unbelievably, these numbers can help determine the need for advance medication or any other solution in the pharmaceutical industry.

2.6. Pharmaceutical Sales

Pharmaceutical companies tend to be organized around the "blockbuster" model, i.e., most of the sales and profits are from a small assortment of broadly active drugs, mass-marketed to a broad patient population by a network of sales representatives. From an industry standpoint, a "blockbuster" is a drug whose annual revenues exceed 1 billion dollars [32]. When companies utilize data mining to analyze industry trends, sales patterns begin to become apparent, and therefore it is possible to accurately forecast which drugs are more likely to become "blockbuster" ones. Pharmaceutical sales are highly competitive and very lucrative. It is estimated that for every six doctors in the United States there is one sales person trying to push the latest and best product [32].

Typically, a company spends less than 1 percent of its sales income on marketing, drug companies spend anywhere between ten and twenty percent [21]. This means that drug companies have tremendously high marketing budgets and therefore are willing to invest in the latest technological tools that will help further advance their strategies. Only two countries in world allow direct to consumer advertising (DTCA), the United States and New Zealand [29]. Some opponents argue that it is wrong to market a drug to a consumer who may not have the education required to fully understand what use the drug has, and if it is a cure for their own illness [15]. With prime time advertising costs so high it is crucial that the company understand exactly who uses the drug and how to effectively launch a campaign that will reach the consumer. The media outlet that effectively reaches the consumer depends heavily on which segment of the market the drug is positioned to. Data mining plays a very important role in properly executing the appropriate campaign by comparing which types of people tend to use different medications. Most marketing is directed at physicians. In 2004, pharmaceutical companies spent 43 billion dollars marketing to physicians, $61,000 per physician in the United States [27].

IMS Health provides business intelligence and strategic consulting services to the pharmaceutical and health care industries. This company utilizes data mining techniques to analyze data from over 29,000 suppliers at 225,000 supplier sites worldwide. IMS monitors over 75% of prescription drug sales in over 100 countries, and 90% of sales within the United States [27]. This type of data gathering allows pharmaceutical companies to better strategize the most efficient ways to reach their demographic. However, gathering such personal information also creates privacy issues, say many opponents of the practice [36]. IMS Health claims that the information obtained is built upon demographics, not on personal selling. This means that IMS simply gathers information based on age, sex, gender, they are not building personal databases of individual users habits. Some people believe this is not the case. The dilemma here is that as long as people use technology, data trails will be created. It is the data miner who collects this raw data and converts the information into something of use.

3. Product Analysis

3.1. Buying Tendencies

There are many factors that contribute to increasing prescription drug expenditures. Such factors include public health and the effects of smoking, consuming alcohol, and obesity. Age referring to ages 65 and older, access to Medicare is also important, as some of the population is uninsured; a few others include new products, unemployment and income. Using data mining techniques one can discover the true effect these factors have on the population's willingness to purchase prescription drugs [12]. Time series models and simple pooling were conducted through studies to observe the determinants of prescription expenditures, however because of "individual effects" some biases had to be fixed. To measure the effect of new drugs from pharmaceutical companies, data are collected from FDA approval percentages and matched against certain independent variables within the population. All of these determinants are compared to the annual rise in prescription expenditures and correlations can then be shown.

3.2. Better Products through Portfolio Management

Portfolio management is used within the pharmaceutical industry to evaluate commercial value and risks of projects that are in the developmental stages. Data mining can help bring the criteria together that will help portfolio management allow you to have the highest value. "Factors brought together include disease percentages, medical needs associated, as well as the developmental projects that are worth continuing. Two major components include evaluation methodology and metrics as well as corporate evaluation process" [6]. Timeline and budget management, sales forecasting, as well as project financial evaluation important roles in the process. The types of risks being undertaken are divided into two groups: strategic risk and operative risk. Process must be cross-functional and include the entire organization in order to properly function and allow the distribution of scarce resources to be allocated to most useful projects [5]. In order to increase the effectiveness and efficiency of the development process, assessments of this process must be made frequently as the field is always changing and the data that are collected becomes outdated quickly.

3.3. Forecasting New Drugs

The problem with a new drug is that it has no history, which in turn makes forecasting difficult. Data mining within the pharmaceutical industry allows us to use three steps to help forecast a new drug: 1. Drug life cycle curve 2.

Growth function 3. Estimation and calibration of growth function [36] [26]. The collection of certain data such as quality, price, and general desirability allow users to see the rate at which a new drug is able to gain use within a particular market. There are many variables that affect use such as age, income, and disease that do not allow the rate of one new drug to correlate to the rate of another. A new drug forecasting approach, the Gompertz curve, is therefore used to express forecast value. This particular method was used to forecast growth of Attention Deficit Hyperactivity Disorder (ADHD). Based on this forecasting it was shown that the Gompertz curve is able to incorporate primary marketing data of an expected new drug within estimated model parameters of any market and calculate estimated drug use [26]. Data mining from a particular market showing the expected acceptance and use of a new product helps to lower uncertainty of result even though no history of drug is incorporated.

Data Mining has become a very important tool in Traditional Chinese Medicine (TCM) research and development. Yang et al. states that "prescription is defined as combinations of herbs with proper proportion" [44]. TCM is a very old segment of the pharmaceutical industry, and they have over one million formulae for prescriptions that have been generated and collected over 3000 years. Prescription discovery is based on a hypothesis that new prescriptions are hidden in data composed of all formulae. The problem for years has been finding a quality method for discovering these new prescriptions. Most prescriptions are created from multiple herbs, so it would take an inconceivable number of man-hours to go through the collective data and determine similarities between different prescriptions for similar ailments. The goal of data mining here is to discover several combinations of herbs in a self-organized way. This can lead to better treatments, and possibly even cures for serious diseases we have been struggling to fight against for years, such as AIDS or forms of cancer.

4. Financial Analysis

Booth observed that companies developing new drugs received more than $150 billion in capital since 1995 [7]. These funds are used not only for the research & development but also to support their operations, drug testing on candidates, hiring of scientists and in some cases, commercializing its products. Where does this money come from? Many private and public equity investors measure their interest based on the valuation of the company. Pharmaceutical companies require millions of dollars and in most case billions over the years in order to fund existing projects. In the private sector, financing has increased from $500 million in 1992 to over $3.5 billion in 2004.

There are various tools and applications of Business Intelligence (BI) including query reporting & analysis tools, data mining tools, and data warehousing tools [25]. By utilizing business intelligence tools, pharmaceutical companies can analyze its recent sales force activity and their results to improve targeting of high-value physicians and determine which marketing activities will have the greatest impact in the next few months.

Companies like Pharmascience Inc. utilize WebFocus BI tools to produce sales reports and provide marketing analysis [18]. Using BI creates one data warehouse with multiple set of data to produce meaningful reports. Depending on the platform, the data can be real-time and can reduce the amount of time it takes to produce reports.

5. Inventory Control & Supply Chain Management

There are several advantages to data mining within the supply chain of pharmaceutical companies. Appropriate data mining tools and methods help companies better forecast, keep flexible production, record sales and operations processes, and make supply chain process more clear. In addition, with some systems, when sales orders are being received, production batches are automatically created [30]. However, this is under the assumption that the supply chain has appropriate data because collecting useful data and organizing it is the hardest part for the entire sector. They need quality and quantity data from inside and outside of companies. Manufactures and distributors develop relationship to share information while some manufacturers' sales representatives try to develop good relationships with physicians or pharmacies for primarily the same reasons.

5.1. Manage Inventories

Pharmaceutical manufacturers are trying to shift their production style from a make-to-stock model to a demand driven approach. It is important to keep better track of warehouse data for managing inventories [43]. Like any other sectors, it is crucial for the pharmaceutical industry to keep an optimal inventory level. What do they have to do in order to obtain the right amount of inventory? First, it is important to forecast their demand curve more accurately. This has a lot to do with quality and volume of information from inside and outside the company. Sharing accurate and timely information between supply chains helps retailers have the right amount of inventory and helps distributors and suppliers monitor the stock level of stores. It contributes to decrease a bullwhip effect [4, 13, 37]. Constantinou and Balakrishnan suggested two examples of data mining tools to get precise demand-signal collection, which are EDI (Electronic Data Interchange) and Electric drug pedigrees [11].

Some companies may also have the financial ability to invest in radio frequency identifier technologies (RFIT). For Epedigree, final drug products are tagged with RFIT that contains a standardized numerical identifier. Information coded on RFIT is also useful for supply chain analysis. Forcinio explains how a RFIT system works by exemplifying track-and-trace pilot of American Source Bergen, a drug distributor [16]. Information of each tagged product is scanned and then stored in a system called EPCIS (Electronic Product Code Information Service). The system obtains all of the tagged items' historical records such as their location in the distribution center, the time that the product shipped, and its destination. This data set is useful for monitoring inventories and product demand. Furthermore, AmerisourceBergen is trying to connect EPCIS to other companies in the supply chain [16].

Memorial Hospital in Chattanooga, TN lost $500,000 from unused or unbilled products such as devices, guide catheters, coronary wires and other high-value items. These items either disappeared or were not charged correctly or they were moved to the wrong department [20]. WaveMark provides real-time inventory solutions in the healthcare supply chain. Data rich RFID labels are always available in an easy to use manner, at point of use, allowing the data to be acted upon in ways that conventional paper and etched labels do not allow [1]. Workers use a special WaveMark desktop reader scanner to scan in any new inventory that has a bar-coded serial number. An RFID tag contains the product's type, size, expiration date and cost. Surgical products are housed in cabinets that have RFID readers that capture each of the tag ID's number. If a nurse removes an item from the cabinet the software changes the product to "missing". The nurse scans the patient's barcode and then scans the product so this links the product to the patient [40].

Johnson & Johnson used RFID tags to track promotional products displayed at RFID-enabled retail locations. As a result, Johnson & Johnson found that more than 79% of its promotional products were placed on the sales floor too early. The data also resulted in stores that complied with the promotional schedule saw increased sales over those that did not adhere to the schedule [35]. By using RFID companies have the opportunity to reduce costs on overstocking as well as increase savings.

DePuy produces implants & orthopedic devices wanted to make its supply chain more efficient with regards to its product Express Care kits. DePuy replaced bar-codes with RFID in order to reduce its return process. The workers would place the kit through an RFID reader tunnel and the software was able to show workers, which components were missing from the kit. The company was able to reduce the return process from 10-30 minutes to less than one minute [22, 35]. This will increase time management as well as increase their revenues by saving the extra 10-30 minutes per kit.

5.2. Reducing Counterfeit Drugs

Many pharmacists have turned to the internet or unknown secondary distributors to purchase certain medications that are in short supply. Not only pharmacists, but consumers have also turned to the internet to purchase drugs rather than getting a prescription from their doctors. Internet and unknown wholesalers offer prices that are often too good to be true and in some cases are exposing customers to misbranded or expired products. Counterfeiting and other forms of intellectual property theft cost American businesses at least $250 billion annually and counterfeit drugs worldwide cost pharmaceutical companies as much as $46 billion annually [31]. Pharmacies should have good inventory methods in place and should monitor purchases for unusual trends. If a patient is not responding to therapy, counterfeit drugs are often overlooked as a possible reason.

Pfizer was not aware that counterfeits of its cholesterol drug Lipitor had infiltrated the United States market until consumers began to complain. More than eighteen million fake Lipitor tablets were recalled following the consumer complaints and the company estimates that counterfeits had reached 600,000 consumers in fifteen states. The company estimates that the total cost of the recall was in the tens of millions of dollars. Pfizer now includes a case study about the counterfeit Lipitor problem in a special section of its website that addresses current issues related to counterfeit drugs. Pfizer developed and posted this case study of its own initiative, indicating that the company sees the Lipitor counterfeit case as an important lesson learned. Today, Pfizer invests time and money into distributor auditing initiatives that aim to prevent counterfeits from causing additional harms to the company's reputation, consumer base, and bottom line [45].

6. Conclusion

Data mining is an important tool to the pharmaceutical industry as well as its suppliers. The data that are aggregated provide many opportunities and trends for companies to produce safe and secure drugs, identify areas that may increase sales, and also create efficiency in its supply chain by reducing errors in reordering supplies. For pharmaceutical companies to be a success, it is crucial for companies to invest time and money in data mining.

The area that requires the most capital is research and development. Data mining gives the companies the ability to research and compile data like never before by using pre-existing data sets coupled with new data to research and design new drugs in the field. While progressing through the four stages of development, pharmaceutical companies have a social responsibility to create and design drugs that are safe for its consumers. Once a company is successful with creating a new drug, the drug has to be marketed to make a return on its investment.

Certainly, pharmaceutical companies have to be careful with how they proceed with data mining. Data mining will help pave the way toward the future in this industry.

7. References

[1] G. Adams, and A. Burke,"Pharmaceutical Manufacturing: RFID– Reducing Errors And Effort", *Filtration &Separatio*, 2007,44(6), pp. 17-19.

[2] M. Aitken, E. Berndt,and D. Cutler, "Prescription Drug Spending In The US" *Healthcare Insight*, 2008, 65(4), pp. 154-163.

[3] S. Altan, J. Bergum, L. Pfahler, E. Senderak, S. Sethuraman, and et al. (2010). "Statistical Considerations In Design Space Development",*Pharmaceutical Technology*, 2010, *34*(7), pp. 62-66.

[4] V.P.Arnum. "Epedigree In The Pharmaceutical Supply Chain",*Pharmaceutical Technology*, 2008, pp.36-39.

[5] F. Balkau, and G.Sonnemann,"Marketing Sustainability Performance Through The ValueChain*", Corporate Governance*, 2010, 10(1), pp. 46-58.

[6] K. Bode-Greuel, andK. Nickisch, "How Can Pharmaceutical And Biotechnology CompaniesMaintain A High Profitability*" Journal of Commercial Biotechnology*, 2008, 14(4), pp. 307–325.

[7] B. Booth."Valuation With Cash Multiples", *Nature Reviews Drug Discovery*, 2005, 4, pp. 533-534.

[8] N.S, Buchan, Rajpal, D.K., Webster, Y., Alatorre, C., Gudivada, R.C., Zheng, C., Sanseau, P., andKoehler, J., "The Role Of Translational Bioinformatics In Drug Discovery", Drug Discovery Today, 2011, In Press.

[9[S.J. Campbell, Gaulton, A., Marshall, J., Bichko, D., Martin, S., Brouwer, C., and Harland, L., "Visualizing The Drug Target Landscape", *Drug Discovery Today*, 2010, 15(1-2), pp. 3-15.

[10] W.Y. Chiang,"To Mine Association Rules Of Customer Values Via A Data Mining Procedure With Improved Model: An Empirical Case Study", *Expert Systems with Applications*, 2011,38(3), 1716-1722.

[11] D. Constantinou, andBalakrishnan, A.,"Beyond The Foggy Supply Chain", *Pharmaceutical Executive*, 2008, 28(1), pp. 34-35.

[12] A. Dattaf, andVandegrift, D.,"Prescription Drug Expenditures In US", *Southern Economic Journal*, 2006, 73(2), pp. 515-529.

[13] S. Ekins, and Williams, A.J., "When Pharmaceutical Companies Publish Large Datasets: An Abundance Of Riches Or Fool's Gold?",*Drug Discovery Today*, 2010, 15(19-20), pp.812-815.

[14] Felix, S., "*Digging For Data – Benefits",*Canada, 2005, 29(9), pp. 47-49.

[15] C. Flood,"The Evidentiary Burden For Overturning Government's Choice Of Regulatory Instrument: The Case Of Direct-To-Consumer Advertising Of Prescription Drugs", *University of Toronto Law Journal,* 2005,60(2), pp. 397-424.

[16] H. Forcinio, "RFID Moves Forward" *Pharmaceutical Technology*, 2007, 31(4), 48-52.

[17] D.J. Hand, "Principles Of Data Mining",*Drug Safety,*2007, 30(7), pp. 621-622.
(Remaining references are available upon request).

A Fully Bayesian Approach to Verifying a Shift in the Process Average

Nasser Safaie and Gamal Weheba
Department of Industrial and Manufacturing Engineering, Wichita State University, Wichita, KS 67260, USA
nxsafaie@wichita.edu, gamal.weheba@wichita.edu

Abstract

A number of sampling techniques can be utilized to quantify the impact of process improvement efforts and verify their effect on its performance. This paper proposes a fully Bayesian approach for determining the maximum economic number of measurements required to verify a shift in the process average. Mathematical models are derived with due considerations of the quadratic loss at specified levels of stable operation. The resulting loss functions are used to establish implementation boundaries from an economic viewpoint. A numerical example is used to illustrate the steps involved and highlight the advantages of utilizing Bayesian methods to support process improvement efforts.

1. Introduction

One of the most important uses of control charts is to distinguish assignable causes from chance causes of process variation. Common causes are inherent in the process all the time. Grant and Leavenworth [1] indicated that control chart applications involve two phases. In phase I, a set of historical data is analyzed to assess stability and identify special causes. If no special causes are present, then the in-control process parameters are estimated. In phase II, the data is sequentially collected over time to monitor performance at the estimated levels. However, there are scenarios where the process is found to be centered away from the desired level, and changes are sought to improve its performance. In this case, the proposed changes must be tested to verify their effect [2]. The duration of this activity is referred to as a transition period, and users of the traditional control charts are limited to two options. One option is to assume complete ignorance of the process level and repeat phase I. The second is to utilize the desired level as the standard value for phase II. The former requires accumulating a new set of data, while the latter may result in an accept/reject decision without providing an updated estimate of the process parameter. The main purpose of this paper is to explore applications of the Bayesian approach as a third option for updating prior knowledge about the process average. The following section represents a review of the relevant literature and highlights some of the merits of the Bayesian techniques. Section 3 represents the procedures followed in deriving the mathematical model for determining sample size. This is followed by an illustrative example and final conclusions.

2. Literature review

The Bayesian and traditional methods deal differently with estimation and test of hypothesis for a parameter of interest. In the traditional methods, parameters are of fixed values and are estimated based on the sample information. Whereas, in the Bayesian methods, parameters are treated as random variables of prior distributions reflecting accumulated knowledge. When new information is obtained, the likelihood is revised so that it may represent all available information [3]. The updated prior, based on the sample, is called posterior distribution $p(\theta|x)$. This approach is based on the Bayes' theorem for densities, which updates the knowledge about the process parameter θ by combining prior belief about its distribution

$p(\theta)$, and the likelihood function resulting from sample data $p(x|\theta)$. This can be expressed as

$$p(\theta|x) \propto p(\theta)p(x|\theta) \quad (1)$$

This approach provides a good bridge between the extremes of total knowledge and total ignorance of process parameters [4].

Sample size determination (SSD) plays a significant role in the design and analysis of any data sampling activity. Directly or indirectly, sample size affects all steps of a study from design to analysis and results. Joseph and Belisle [5] argued that calculating the sample size without considering the prior information and uncertainty may lead to a larger sample size than is necessary. According to Berger and Sellke [6], SSD is essentially Bayesian in nature, since it absolutely requires prior information. Broemeling [7] noted that Bayesian methods are more applicable than traditional methods in decision problems for which the prior and loss information are part of the problem. Roshan [8] noted that the loss and utility functions play a key role in evaluating a decision for Bayesian methods.

Loss functions depict the economic impact of choosing a specific action. Pham-Gia and Turkkan [9] indicated that the Bayesian approach pays attention to the cost corresponding to an optimal decision (d) when the prior information is taken into account. They pointed out that the lower the decision loss, the larger the sample size, and consequently the larger the cost of sampling. The early work by Raiffa and Schlaifer [10] identified a utility function in the form $U(n, x, d, \theta)$ to quantify the sample size n. Following the same modeling approach, Lindley [11] considered a utility function and included a sampling cost of C for each observation as:

$$U(n,x,d,\theta) = U(d,\theta) - Cn \quad (2)$$

Berger [12] determined an upper bound for sequential sampling by expressing a loss function in the form

$$L(n,x,d,\theta) = -U(d,\theta) - Cn \quad (3)$$

The distinguishing characteristic of the sequential analysis method is that the observations are taken one at a time, and a decision is made after each observation either to stop sampling and choose an action or to take another observation.

On the other hand, Press [13] criticized the complexity of Bayesian calculations and the hidden objectivity in selecting the prior distribution. He argued that even though a subjective prior distribution has the value of additional replications, it is not always easy to translate the prior information to a meaningful distribution. Also, Marcellus [14] summarized a number of limitations of Bayesian methods. A practical difficulty is that more information is necessary about the process than that of the traditional methods, even though acquiring this knowledge can yield real benefits. He also noted that assessing the probabilities used in Bayes' theorem require special training and skills. Typically, the Bayesian approach is not recommended when there is no good understanding of the process, since the objectivity of prior information is crucial in SSD and decision-making [15].

These authors believe that most of the above limitations can be avoided by considering the two methods as complementary techniques. The information gained during phase I application of the traditional control charts can be used as a starting point in the utilization of Bayesian methods. The following section represents an attempt to utilize the Bayesian methods in support of the traditional methods used in statistical process control.

3. Mathematical modeling

The process considered is one of producing a single quality characteristic x. The specification limits for x are expressed as $m \pm \Delta$. It is also assumed that a state of statistical control has been established and the quality characteristic was shown to follow a normal distribution with mean $\mu_1 \neq m$ and variance σ^2.

Production is expected to continue over future periods to supply a total of N units. Two options are considered for process management. The first option (a_1) involves taking no action and continuing operation at an estimated per unit loss(L_1), which can be estimated according to

Taguchi [16] as

$$L_1 = \frac{A}{\Delta^2}[\sigma^2 + (\mu_1 - m)^2] \qquad (4)$$

Where, A represents the repair or replacement

cost for one unit of product. The second option (a_2) involves making changes to shift the average closer to m at μ_2. As such, L_2 can be expressed as

$$L_2 = \frac{A}{\Delta^2}[\sigma^2 + (\mu_2 - m)^2] \qquad (5)$$

The second option will also require an initial investment of $\$W$ for tooling and setup. The probability that the changes will successfully result in improving the process level is estimated a priori as p. This information has been summarized in a loss matrix, as shown in Table 1.

Table 1. Loss matrix

	Maintain Current Level (a_1)	Improve to New Level (a_2)
Probability of Success (p)	$N(L_1)$	$N(L_2\text{-}L_1)+W$
Probability of Failure (1-p)	$N(L_1)$	W

The expected value of the loss, namely the loss function, for the two options can be calculated as

$$R(a_1) = N(L_1) \qquad (6)$$
$$R(a_2) = pN(L_2 - L_1) + W \qquad (7)$$

The criterion for economic improvement requires that the average expected value of the loss under the second option be less than that under the first or $R(a_2) < R(a_1)$. When this

criterion is satisfied, a_2 would represent the optimal Bayes rule in which the expected loss is minimized. As such, a sample of size n measurements is required to verify the effect of the proposed changes at the rate of $\$C$ per unit. The average Bayes risk associated with a_2 can be calculated as

$$r(a_2) = E[R(a_2)] + Cn$$

$$= pNE[(L_2 - L_1)] + W + Cn \qquad (8)$$

Now suppose that an unbiased gauge with variance σ_g^2 ($\sigma_g^2 < \sigma^2$) is available to collect the

measurements. The resulting sequence of measurements $x_1, x_2, \ldots, x_n$ can be used to update the prior belief regarding the process average. Since x is normally distributed, the likelihood function of x given μ is expressed by

$$f(x|\mu) = \frac{1}{\sqrt{2\pi\sigma_g^2}} \exp -\frac{1}{2}\left(\frac{x-\mu}{\sigma_g}\right)^2 \qquad (9)$$

As was shown by Box and Tiao [17], the posterior density after n measurements is also a normal function in the form

$$f(\mu|x_1, x_2, \ldots, x_n) =$$

$$\frac{1}{\sqrt{2\pi\rho_n}} \exp -\frac{1}{2}\left(\frac{\mu-\mu_n}{\rho_n}\right)^2 \qquad (10)$$

Where μ_n and ρ_n are the posterior mean and variance given by

$$\mu_n = \frac{n\sigma^2 \bar{X} + \sigma_g^2 \mu_1}{\sigma_g^2 + n\sigma^2} \quad (11)$$

$$\rho_n = \frac{\sigma^2 \sigma_g^2}{\sigma_g^2 + n\sigma^2} \quad (12)$$

Substituting values of L_1 and L_2 into Equation (8), the expected risk $r(a_2)$ can be expressed as

$$r(a_2) = pNk([E(\mu - m)^2 - E(\mu_1 - m)^2] + W + Cn \quad (13)$$

Since $E(\mu - m)^2 = \rho_n$ then,

$$r(a_2) = pNk\left[\frac{\sigma^2 \sigma_g^2}{\sigma_g^2 + n\sigma^2} - (\mu_1 - m)^2\right] + W + Cn \quad (14)$$

To explore the characteristics of this function, the sample size n is treated as a continuous variable as an approximation. The first and second derivatives of Equation (14) reveal that $r(a_2)$ is convex in n. Consequently, setting the first derivative of Equation (14) equal to zero and solving for n results in the maximum economic sample size n^*, expressed as

$$n^* = \sqrt{pN\frac{A}{\Delta^2}}C^{-1/2}\sigma_g - \frac{\sigma_g^2}{\sigma^2} \quad (15)$$

Due to the nature of n^* and the ratio of the gauge variance to the process variance, the last term in Equation (15) can be ignored in most practical scenarios. It is worth noting here that Equation (15) is in agreement with that given by Berger [12], except for the term under the square root. This scaling factor includes important parameters that reflect the economic value of the quality characteristic x, and the level of certainty in prior beliefs. Large values of this factor allow larger samples to be taken economically.

The sampling procedure used depends on the way the process will react to the changes made. If the process is expected to experience a sudden shift, the average of the n* measurements $\bar{X}$ would represent a sufficient statistic for calculating the posterior mean given by Equation (11). To verify the shift in the process average, a credible interval in the form $\mu_2 \pm z_{\alpha/2}\sqrt{\rho_n}$ can be constructed at a given level of α. If such an interval encloses the posterior average μ_n, then it can be concluded that the changes have resulted in the desired shift. In this case, phase II application of the traditional control chart would resume with μ_n established as the new process average. Otherwise, improvement efforts have to continue following the same procedure while testing different changes. A schematic representation of this sampling procedure is shown in Figure1.

On the other hand, if the nature of the process reaction to the changes is unknown, it would be beneficial to update the average sequentially. That is, values of the posterior mean would be updated after each measurement (x_i) using

$$\mu_i = \frac{\sigma^2 x_i + \sigma_g^2 \mu_{i-1}}{\sigma_g^2 + \sigma^2} \quad (16)$$

Where $\mu_0 = \mu_1$. As shown in Figure 2, this procedure should repeat following each measurement until enough statistical evidence of the shift could be obtained or a total of n^* measurements have been collected. Once again, a credible interval in the form $\mu_2 \pm z_{\alpha/2}\sqrt{\rho}$ can be constructed based on a specified value of α. In this case the posterior variance ρ is obtained by setting n=1 in Equation (12). When such an interval encloses the posterior mean μ_i, sampling should stop, and a decision may be reached before collecting the maximum number of measurements n^*. However, the credible interval obtained in this case is based on a larger value of the posterior variance based on n=1. The following section represents a numerical example to help illustrate procedures for calculating the maximum economic sample size n^*.

4. Numerical example

The design specifications for a critical dimension x are given as 8.10 ± 0.05. Based on phase I application of control charts, the process was found to be operating in a state of statistical control. The distribution of x was found to follow the normal function with an average of μ_1 = 8.06, and variance of $\sigma^2 = (0.01)^2$. Production

is expected to continue to produce N = 2,000 units. The cost of replacing a nonconforming unit is given as A = \$100. The process improvement team is considering one of two options. The first (a_1) is to continue operation at the current level with an average per unit loss of L_1 = \$68, as estimated using Equation (4). The second option (a_2), is to make changes to the current process as to cause a shift in the process average closer to the target while maintaining the same process variance σ^2. The setup cost is

estimated as W = \$1,000 and an unbiased gauge with variance of $\sigma_g^2 = (0.002)^2$ is available to take measurements at per unit cost of C = \$10. Based on previous experience, the team anticipates that the changes proposed would cause the average to shift from 8.06 to 8.10 with a probability of p = 0.7. At the new level, and using Equation (5), the estimated loss is reduced to L_2 = \$4 per unit. Utilizing Equations (6) and (7), it can be seen that a_2 represents the optimal Bayes rule in which the expected loss is minimized. Figure 3 shows a plot of Equation (14) representing the Bayes risk at various values of the sample size.

As can be seen, the average risk r (a_2) is minimized at maximum economic sample size n^*= 5 at which an average net gain (negative loss) of \$8,505 is attained. Similar results could have been obtained using Equation (15).

5. Conclusions

The modeling approach used in this paper builds on the ability of the Bayesian methods to update process parameters during transition periods without the need for making questionable assumptions. It is aimed at overcoming the subjectivity in specifying the prior distribution by utilizing knowledge gained during phase I of traditional control charts. The resulting model is simple and accounts for important parameters including the gauge variability, the cost of collecting measurements, and the process setup cost. These authors are currently researching a modeling approach to determine the sample size required to verify a reduction in the process variance.

6. References

[1] E.L Grant, and R.S. Leavenworth, Statistical Process Control (5th ed.), McGraw-Hill, New York, 1980.

[2] H.M. Wadsworth, K.S. Stephens, and A.B. Godfrey, Modern Methods for Quality Control and Improvement (second ed.), Wiley, New York, 2002.

[3] P. Congdon, Bayesian Statistical Modelling (secod ed.), John Wiley and Sons, Chichester, 2007.

[4] T. BIBLIOGRAPHY Bayes, An Essay Towards Solving a Problem in the Doctrine of Chaces. Philosophical Transactions of the Royal Society, 1783, 53, pp. 370-418.

[5] L. Joseph, and P. Belisle, Bayesian Sample size Determination for Normal Means and Differences Between Normal Means. The Statistician, 1997, 46, pp. 209-226.

[6] J. Berger, and T. Sellke, "Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence", Journal of the American Statistical Association, 1987, 397, pp. 112-122.

[7] L.D. Broemeling, Bayesian Methods for Measures of Agreement, Taylor and Fransic Group, Boca Raton, 2009.

[8] J.V. Roshan, "Quality Loss Functions for Nonnegative Variables and Their Applications", Journal of Quality of Technology, 2004, 36(2), pp. 129-138.

[9] T. Pham-Gia, and N. Turkkan, "Bayesian Decision Criteria in the Presence of Choices under Quadratic and Absolute Value Loss Functions", Statistical Papers, 2005, 46(2), pp. 247-266.

[10] H. Raiffa, and R. Schlaifer, Applied Statistical Decision Theory, Harvard University, Graduate School of Business Administration, Boston, 1961.

[11] D.V. Lindley, "The Choice of Sample Size", The Statistician, 1997, 46(2), pp. 129-138.

[12] J. Berger, Statistical Decision Theory and Bayesian Analysis (second ed.), Springer-Verlag, New York, 1985.

[13] S.J. Press, Subjective and Objective Bayesian Statistics, John Wiley and Sons, NJ, 2003.

[14] R.L. Marcellus, "Bayesian Monitoring to Detect a Shift in Process Mean", Quality and Reliability Engineering International, 2008, 24, pp. 303-313.

[15] A. Gelman, J.B. Carlin, H.S. Stern., and D.B. Rubin, Bayesian Data Analysis (second ed.), Chapman and Hall/CRC Press, Boca Raton FL, 2004.

[16] G. Taguchi, Introduction to Quality Engineering, Asian Productivity Organization, Tokyo, 1986.

[17] G.E.P. Box, and G.C. Tiao, Bayesian Inference in Statistical Analysis, Wiley-Hall. New York, 1973.

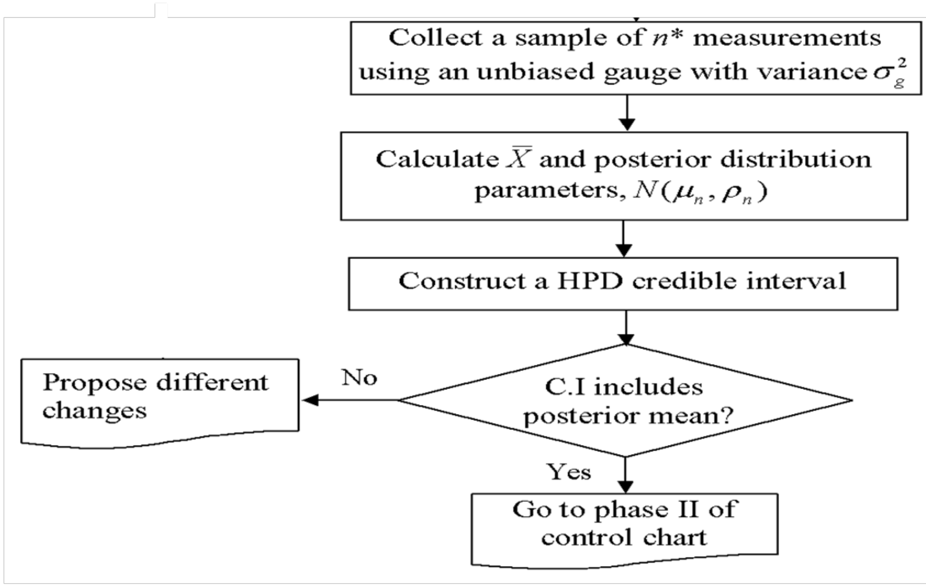


Figure 1. Sampling procedure under sudden shift

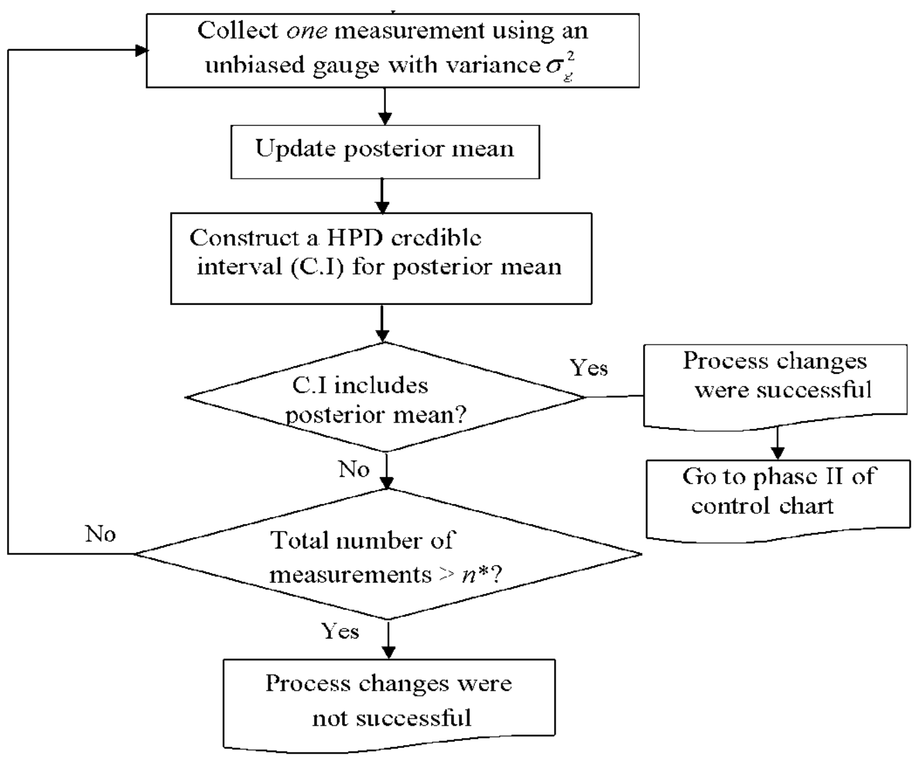


Figure 2. Sequential sampling procedure

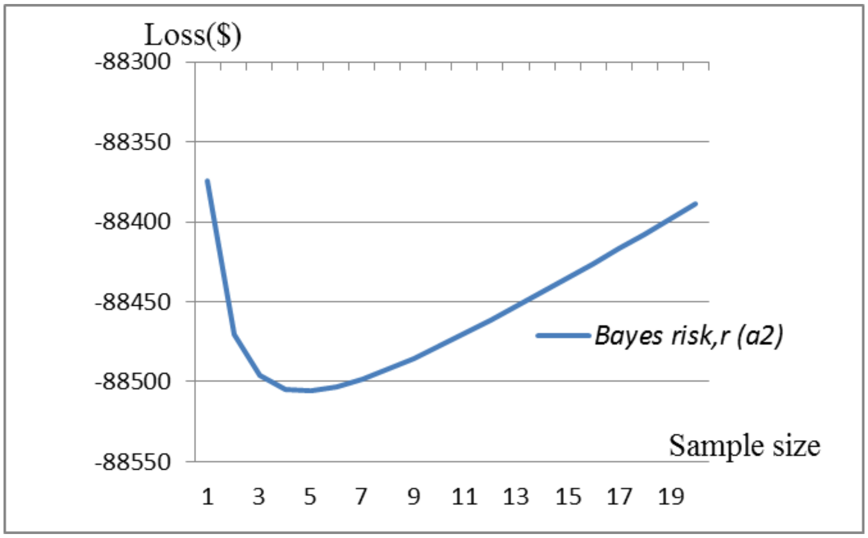


Figure 3. Bayes risk plot

Visual Project Management Tools and Their Application

Wei Zhan and Joseph A. Morgan
Department of Engineering Technology & Industrial Distribution, Texas A&M University
College Station, TX 77843, USA
zhan@entc.tamu.edu, *morgan@entc.tamu.edu*

Abstract

Engineering technology students at Texas A&M University follow a product development process and use many project management tools in their two-semester capstone design projects. This practice allows students to learn how engineering projects are executed in the real world in addition to the technical skills required for their projects. Students used several visual tools including the Project Review Chart (PRC) and the CPI/SPI graph to manage their capstone design projects. These visual tools are proven to be highly effective for management reviews such as weekly or monthly reviews. The creation of these graphs can be done with widely accessible software tools such as Excel and Visio.

1. Introduction

Many companies have their own processes for developing new products. These processes involve the creation of a new product concept, market analysis, establishment of system requirement, system design, testing and validation, manufacturing, and customer support. Unfortunately, most of the educational experience for a typical engineering technology (ET) student focuses only on one of these aspects: system design. The other aspects are mostly neglected. System design requires specific knowledge in a technical field and should be the focus of the educational experience of ET students; however, neglecting the other aspects creates a huge gap between students' educational experience and the real world tasks they will eventually face. There is an urgent need for ET programs to reduce this gap by including all aspects of the product development process in the curriculum.

Beaudoin and Ollis [3] designed a laboratory for freshmen to introduce the concept of product development process. Bartolomei *et al.* [2] successfully used system engineering to design the curriculum for a major at the United States Air Force Academy. Olds *et al.* [21] taught design process to freshmen and sophomores. Zhan *et el.* [30] argued that product design processes should be taught early and repeatedly throughout the curriculum. Recent trends in engineering education have led to an increased use of capstone design projects for teaching product development process [6,7,22,23,26]. A good review in this subject can be found in [5].

Starting in 2001, the Electronics Engineering Technology and Telecommunication Engineering Technology (EET/TET) programs in the Department of Engineering Technology and Industrial Distribution at Texas A&M University now use a two-semester capstone design sequence [18,22] to teach students product development process. During the first course, students look for real-world projects, typically sponsored by industrial partners of EET/TET. The students learn the fundamentals of project management and use these to come up with a project plan which will be executed in the second semester. Typically, three to four students form a start-up company. The format of start-up companies was used to promote invention, innovation, and entrepreneurship in engineering, which was discussed extensively in [27]. The students brainstorm for product ideas, conduct market analysis, define system requirements and a test plan, design hardware and software for their product, and fabricate a prototype. They are also required to create a design document, a test report, and a user manual. The capstone design process closely mimics a typical product development process in industry. This has turned out to be an effective way to reduce the gap between their educational experience and real world tasks. The new capstone design project practice has received positive feedback from the industrial advisory board for EET/TET.

In this article, some of the tools used in the capstone design project courses will be discussed with a focus on visual representation of the concepts. Although most of the concepts can be expressed in words or formulas, a visual representation is often more efficient. Some claim that visuals are six times more efficient than words; a picture is worth a thousand words.

2. Product development process and project management

Product development process and project management are typically taught in Industrial Engineering programs. Although ET students have the option to take these courses as technical electives, the alternative of teaching/learning these concepts in the capstone design project courses seems to be a better choice. Students are more motivated to learn the knowledge if it can be used to solve their problems directly and more efficiently.

Once the product concept and market analysis are completed and the decision is made to move forward with a new product during the first semester of the capstone design project courses, a systematic process needs to be followed to make the product development process more efficient. System engineering [16] is one of such processes. It requires the engineers to establish a system requirement document [9] before the actual design. A test plan is then created so that the requirements can be validated after the product is designed and fabricated. Subsystem level requirements are then derived from the system level requirements followed by a subsystem test plan. The subsystem level requirements are then used to create component level requirements followed by a component level test plan. After the design requirements and test plans are established, the components are selected or designed and then tested against the requirements. After that, the subsystems are fabricated and tested, followed by system integration and testing. The complex process can be easily understood with the visual representation of the system V in Fig. 1.

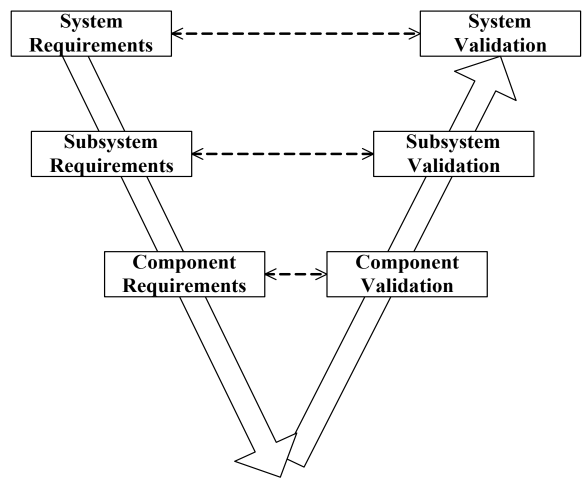


Figure 1. System V

The capstone design projects are medium scale engineering projects which typically require two to three man-years of engineering work. For such projects, one can appreciate the effectiveness of project management methodologies [4,8,13,15, 19]. Since project management tools are widely used in industry for various projects, learning to use these tools during the capstone design project courses better prepares students for their careers in engineering after graduation.

While students design their products following the system engineering process, they use project management tools to manage their project. Just like a real engineering project in industry, they focus on the quality of their products, budget for the projects, and timely completion of the projects. A Work Breakdown Structure [4,8,13,17] is created first, as shown in Fig. 2.

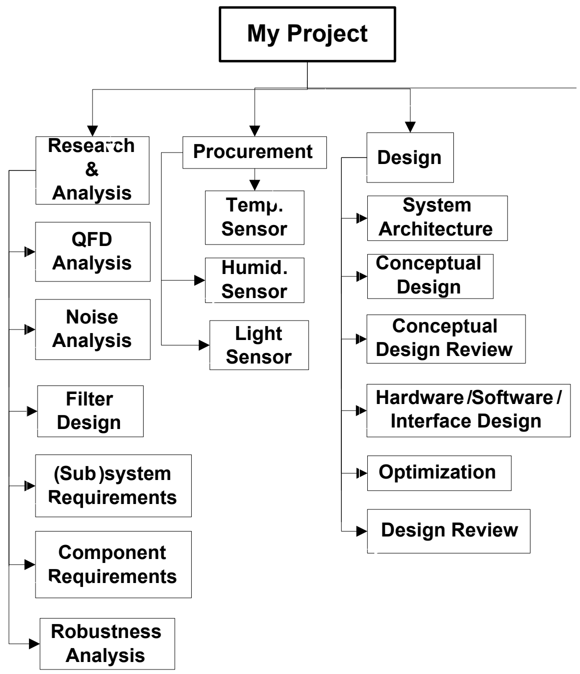


Figure 2. Work Breakdown Structure

With estimated time duration for each activity, the tasks in the Work Breakdown Structure are then put in a logical order to create a Critical Path Method (CPM) model [1,10-12,17,20,24,28,29].

CPM is commonly used for project management in industry. Any project with interdependent activities can apply this method of analysis. There are different formats for CPM models, but they should all contain the following components: 1) a list of all activities required to complete the project; 2) the time required to complete each activity; and 3) the logic dependence among activities. Using the CPM model, the longest path of planned activities from the beginning of the project to the end of the project can be calculated. These paths are defined to be the critical paths. Any delay in the completion of the activities on the critical paths will cause a delay to the overall project completion. CPM was used by students to identify their focus areas and determine when to crash the critical path when the project is estimated to be delayed. Different formats were tried, as illustrated in Figs. 3, 4, and 5.

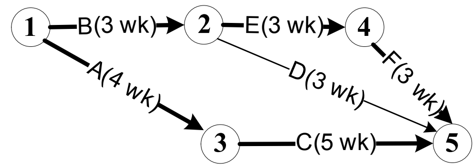


Figure 3. Critical Path Method (Format 1)

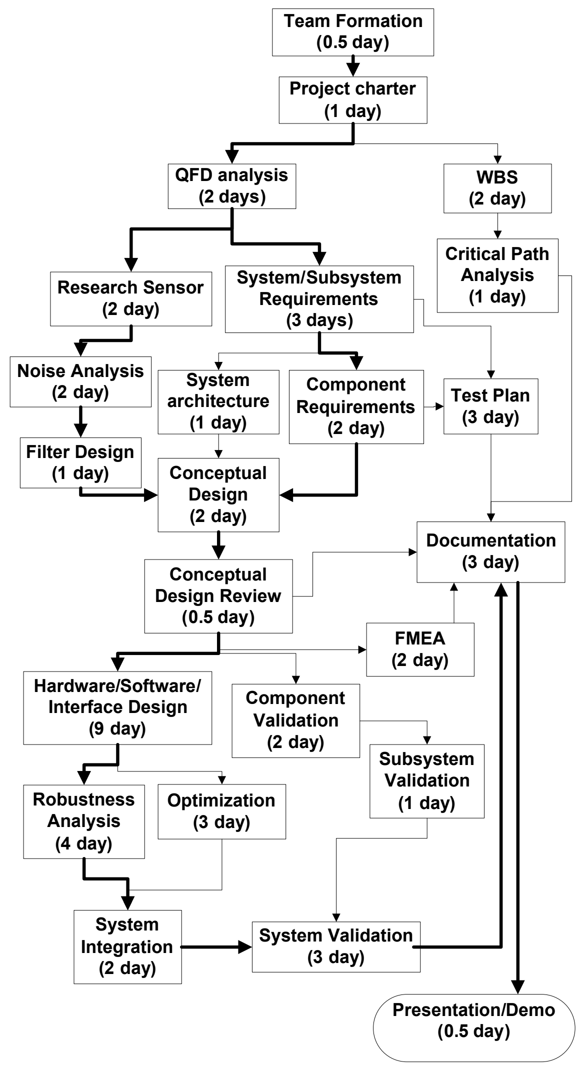
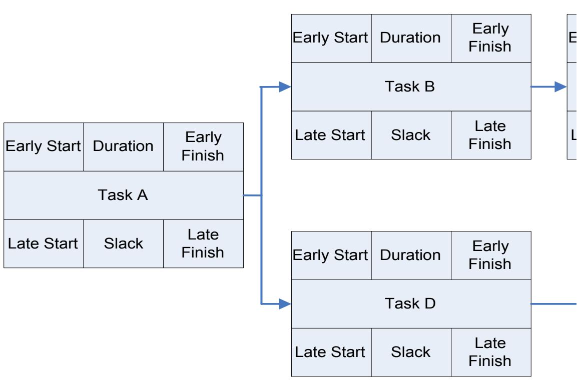


Figure 4. Critical Path Method (Format 2)

Each of these visual representations of the Critical Path Method models has its advantages and disadvantages. Format 1 is easy to create using Microsoft Word, Visio, or Power Point. It has a compact format, and it is thus easier to fit the entire project in one page. It also shows milestones in addition to the activities, their time durations, and critical path. However, if one wants to use more meaningful names for the milestones and activities, the graph can become very crowded and not easy to read. A trade-off is to list the milestones and activities on a separate page with more meaningful labels. Format 2 can be created using Microsoft Word, Visio, or Power Point. It contains the activity names and duration, milestones, and highlighted critical paths (as indicated by the thicker lines in Fig. 4). Format 3 can be easily created using Visio. It has the information on early start, early finish, late start, late finish, and slack for each task. For large scale projects, both Formats 2 and 3 can be lengthy. All three formats require the students to understand how critical paths are identified. There are other commercial software tools available for creating CPM models. These software tools are usually easy to use.

Figure 5. Critical Path Method (Format 3)



A fourth alternative is using Microsoft Project. Once students learned how to use the software, it becomes easy to create a Gant chart; it can be viewed as a network logic diagram where critical paths are displayed. It is easy to make changes and recreate the critical paths after the changes. The disadvantages of using this tool are: 1) Microsoft Project is not as readily available to every student as Microsoft Word or Visio; 2) With everything automatically generated, students are not required to understand how the critical paths are identified. 3) While the software can capture all the details of the project, the document itself can be lengthy and difficult to review. It may be a good tool to use after students have already worked on their own model using Visio or Microsoft Word, provided that Microsoft Project is available to students. If the software is not available to students, it may be worthwhile to present the creation of a Gant chart and CPM model using Microsoft Project.

After learning the underlining theory of project management tools such as the CPM analysis, students were encouraged to create graphical representations using simple software that was readily available to them. This process reinforced the knowledge they learned. Each semester, students had access to a Knowledge Base database that stored what was created in the past semesters. Continued improvements and variations were made by students every semester.

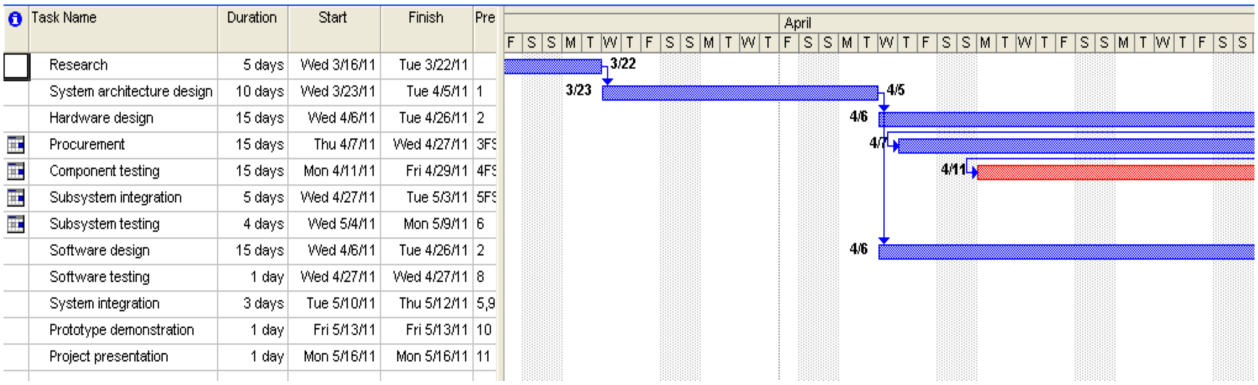


Figure 6. Microsoft Project

3. Visual representations for PRC, CPI, and SPI

Over the years, students learned different project management tools and software packages that can be used to create visual representations. They also came up with their own ideas of generating new visual representations. It was noticed first by faculty members that the length of the Network Logic Diagrams or Critical Path Method models were not convenient for weekly project status review. Other than Microsoft Project, all the CPM formats are difficult to update and do not show the progress from one week to the next. The idea of creating a weekly project status report was discussed. The basic requirements for a Project Review Chart (PRC) were: 1) It must fit in a single 8.5x11 page; 2) It must show the statuses (as a % completion) of major tasks for current and last review meeting; 3) It must identify (e.g. highlighting) the critical paths. With this feedback from the faculty members, students created a PRC, as illustrated in Fig. 7. This has become a convenient project review tool. It is also going through continuous improvement every semester. For instance, students suggested adding the total duration for each task in PRC. PRC can be easily generated in Visio. It is also easy to update the chart on a weekly basis.

As discussed earlier, as far as project management is concerned, the focus is on project budget and in timely completion of the project. To help the student teams with these focuses, the concepts of Cost Performance Indicator (CPI) and Schedule Performance Indicator (SPI) were introduced to the students. The CPI value indicates the efficiency in utilizing the resources allocated to the project. The SPI value indicates the project team's efficiency in utilizing the time allocated to the project. They are defined as

$$CPI = EV/AC \quad (1)$$

$$SPI = EV/PV \quad (2)$$

where EV is the earned value, AC is the actual cost, and PV is the planned value. They can be calculated as follows

PV = The total cost of the work scheduled /Planned as of a reporting date.

AC = The total cost taken to complete the work as of a reporting date.

EV = The total budget * % actual completed.

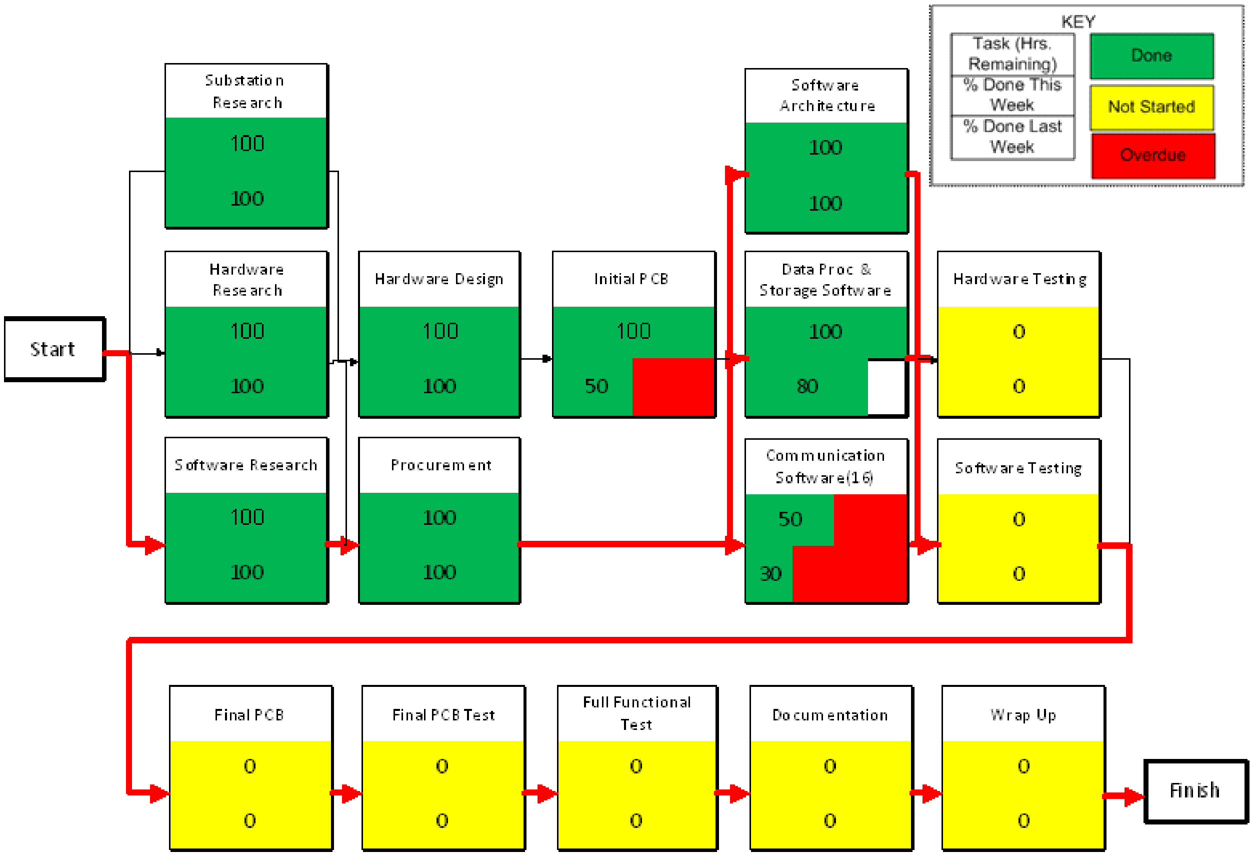


Figure 7. Project Review Chart [14]

If CPI is greater than 1.0, the project is under budget; if CPI is less than 1.0, the project is over budget. The larger the CPI is, the more efficient the project team is in utilizing the resources. If SPI is less than 1.0, the project is behind schedule; if SPI is greater than 1.0, the project is ahead of schedule. The larger the SPI is, the more efficient the project team is in time management. The CPI and SPI values can be calculated over the project duration and plotted as functions of time. Fig. 8 shows an example for a CPI/SPI chart.

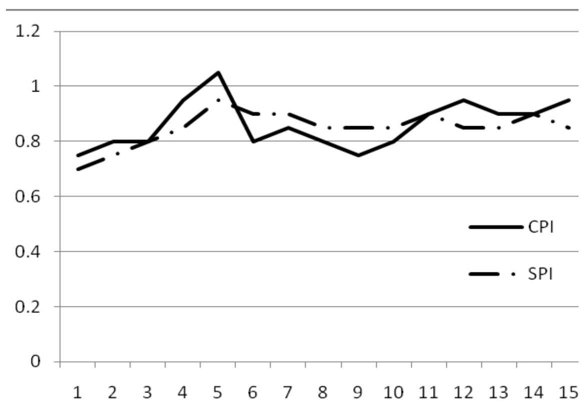


Figure 8. CPI/SPI as function of time

Students took these concepts and thought the CPI/SPI chart was boring. They decided to create an Excel program that takes the inputs from the project manager, does the calculation

[25], and generates a SPI vs. CPI chart, as shown in Fig. 9. In the SPI vs. CPI chart, two target areas are plotted together with the CPI/SPI values for each review period.

With challenges like the creation of PRC and SPI vs. CPI, students were motivated to learn theories, used advanced features in software such as Excel and Visio, and had fun working on their own visual representation of these project management tools. They were also more interested in using tools created by themselves to manage their projects.

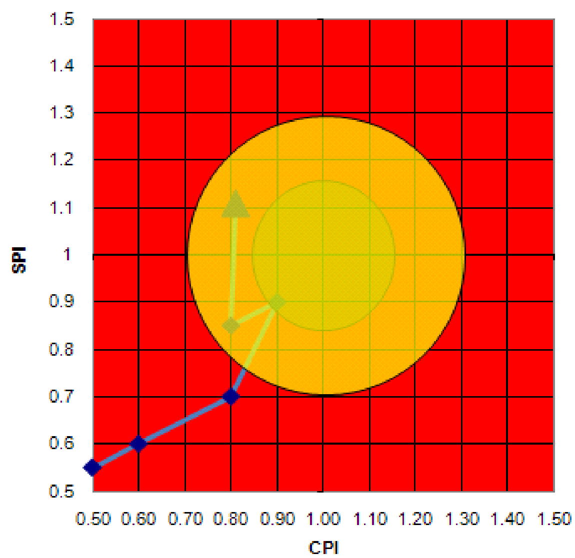


Figure 9. SPI vs CPI [14]

4. Conclusions

Product development processes and project management tools were extensively used in capstone design projects in EET/TET programs in the Department of Engineering Technology & Industrial Distribution at Texas A&M University. In addition to the technical skills, students learned all aspects of the product development process. The gap between educational and real-world experiences is reduced; as a result, they are better prepared for their engineering career after graduation. Several visual representations of project management tools were used. Based on what they learned, the students took up the challenge to create some new project management tools. A continuous improvement process is in place to motivate students to learn, use, and enhance these tools. Future work includes web-based programs for generating these graphs. Initial feedback from students was very positive. Formal evaluation for the effectiveness of using these project management tools will be carried out by surveying the current and former students.

5. References

[1] S. L. Baker, "Critical Path Method (CPM)," University of South Carolina, Health Services Policy and Management Courses, last accessed: March 18, 2011: http://hspm.sph.sc.edu/COURSES/J716/CPM/CPM.html

[2] J. E. Bartolomei, S. L. Turner, and C. A. Fisher, "Using the Systems Engineering Method to Design A System Engineering Major at the United States Air Force Academy," *Proceedings of the ASEE Annual Conference*, 2004-1387.

[3] D. L. Beaudoin and D. F. Ollis, "A Product and Process Engineering Laboratory for Freshmen." *Journal of Engineering Education*, Vol. 84, No. 3, 1995, pp. 279-284.

[4] D. L. Cleland and L. R. Ireland, *Project Management: Strategic Design and Implementation*, 5th ed., McGraw-Hill Professional, 2006.

[5] A. J. Dutson, R. H. Todd, S. P. Magleby, and C. D. Sorensen, "A Review of Literature on Teaching Engineering Design Through Project-Oriented Capstone Courses," *Journal of Engineering Education*, January 1997, pp. 17-28.

[6] G. W. Forman and C. D. Reese, "Capstone Design - On Teaching Non-Engineering Considerations in Engineering Design," *Proceedings, ASEE Annual Conference*, 1984, pp. 808-810.

[7] J. H. Hamelink, "Industrial Oriented Senior Design Projects. A Key for Industrial Experience," *Proceedings, Advances in Capstone Education Conference, Brigham Young University,* 1994, pp. 87-89.

[8] G. Heerkens, *Project Management (The Briefcase Book Series)*, McGraw–Hill, 2001.

[9] G. Herzwurm, S. Schockert, and W. Mellis, *Joint Requirements Engineering*, 2nd ed., Friedrick Vieweg & Son/Morgan Kaufmann Publishers, 2000.

[10] J. Kelley, "Critical Path Planning and Scheduling: Mathematical Basis," *Operations Research*, Vol. 9, No. 3, May-June, 1961.

[11] J. Kelley, and M. Walker, "Critical-Path Planning and Scheduling," *Proceedings of the Eastern Joint Computer Conference,* 1995.

[12] J. Kelley, and M. Walker, *The Origins of CPM: A Personal History*, PMNETwork 3(2), 1989, pp.7-22.

[13] H. Kerzner, *Project Management: A Systems Approach to Planning, Scheduling, and Controlling* (9th ed.), 2005.

[14] J. Kirsey, L. Porter, J. McCracken, and D. Kwan, "T.I.E.S. capstone design final report", Department of Engineering Technology & Industrial Distribution, Texas A&M University, May, 2010.

[15] T. Klastorin, *Project Management: Tools and Trade-offs* (3rd ed.), Wiley, 2003.

[16] A. Kossiakoff, and W. N. Sweet, *Systems Engineering Principles and Practice*, Wiley-Interscience, Dec, 2002.

[17] D. Z. Milosevic, *Project Management ToolBox: Tools and Techniques for the Practicing Project Manager*. Wiley, 2003.

[18] J. A. Morgan, G. Wright, and J. R. Porter, "Managing Senior Design Projects to Maximize Success: The TAT Team," *American Society of Engineering Education Annual Conference*, Portland, Oregon, June 12-15, 2005

[19] M. Newell and M. Grashina, *The Project Management Question and Answer Book*, American Management Association, 2003, p. 98.

[20] J. J. O'Brien and F. L. Plotnick, *CPM in Construction Management, Seventh Edition*, McGraw Hill, 2010.

[21] B. M. Olds, M. J. Pavelich and F. R. Yeatts, "Teaching the Design Process to Freshmen and Sophomores," *Journal of Engineering Education*, July/August, 1990, pp. 554-559.

[22] J. R. Porter, J. Wright, J. A. Morgan, B. Zoghi, R. Fink, and Y. Jung, "Project ELVIS: An example of an innovative capstone process," *Proceedings of The 2006 IJME-INTERTECH Conference*, 2006.

[23] J. R. Porter, B. Zoghi, and J. A. Morgan, "Integrating Project Management into the Capstone Senior Design Course," *American Society of Engineering Education Annual Conference*, Montreal, Canada, June 16-19, 2002.

[24] Project Management Institute, *A Guide To The Project Management Body Of Knowledge* (3rd ed.), Project Management Institute, 2003.

[25] J. Terkelsen, "Earned value user's guide," *Capstone design project report*, Department of Engineering Technology & Industrial Distribution, Texas A&M University, April, 2008.

[26] R. H. Todd, C. D. Sorensen, and S. P. Magleby, "Designing a Senior capstone Course to Satisfy Industrial Customers," *Journal of Engineering Education,* vol. 82, no. 2, 1993, pp. 92-100.

[27] E. L. Wang, J. A. Kleppe, "Teaching Invention, Innovation, and Entrepreneurship in Engineering," *Journal of Engineering Education*, October 2001, pp. 565-570.

[28] M. B. Woolf, *Faster Construction Projects with CPM Scheduling*, McGraw Hill, 2007.

[29] B. Wortman, W. R. Richdson, G. Gee, M. Williams, T. Pearson, F. Bensley, J. Patel, J. DeSimone, and D. R. Carlson, *The Certified Six Sigma Black Belt Primer*, West Terre Haute, IN: Quality Council of Indiana, 2001.

[30] W. Zhan, B. Zoghi, and R. Fink, "The Benefit of Early and Frequent Exposure to Product Design Process," *Journal of Engineering Technology*, Spring 2009, pp. 34-43.