Appendix B – Survey Design

The survey was designed to validate claims that system-level security metrics are better indicators of overall system security than component-level security metrics.

The survey has four distinct sections: demographic, contact information, security metrics experience, and opinions of the efficacy of various security assessment and implementation techniques.

Demographic information included background on information security metrics experiences, industry affiliation, and education level. Contact information was optional and requested only if it was necessary to clarify answers.

Information security metrics experience was elicited via open ended questions designed to separate experts in security metrics from average security professionals. These questions included defining what is meant by security metrics and commenting on standard information security metrics publications.

The survey was composed of questions in five categories.

- **ISACA Demographic Questions:** Demographic baseline questions are the same as those asked the Information Systems Audit and Control Association, the international organization certification authority for information systems auditors and security managers (ISACA, www.isaca.org).
- **Non-ISACA Demographic Questions:** Demographic baseline questions that are not the same as one that are asked by ISACA. Note ISACA uses these types of demographic categories too, but the choice of answered to these ISACA questions are different than those listed below. In most cases, they have been enhanced to

provide more detail. In some cases, categories have been condensed or ranges expanded to reflect relevance to security architecture questions. For example, ISACA distinguished between organizations that have 150-300 employees, but its largest organization size range is greater than 13,000. The same question below has an answer for 150-500 employees and include an organization size of 75,000 or more. This reflects an expectation that organizational size influences security metrics choices at less granular levels in smaller organizations, but in more granular levels in very large organizations.

- **Security Metrics Baseline Questions:** These questions were designed to assess the respondent's familiarity with the field of metrics and measurement as applied to security.
- **Security Questions:** These questions mapped to the requirements for evidence gathering to support the inductive reasoning about system security attributes. They included questions corresponding to each of the three dimensions of security identified in the *Research Hypothesis.* System-level security metrics were defined as those which supported the research hypothesis that system security can only be measured using system-level attributes of support for mission and purpose, validated input, and incident detection and response. The survey included metrics that corresponded to these three dimensions of security, and it also included every other type of metric identified in the literature of security professional practice. It purposely rephrased similar questions from different perspectives (e.g. security measurement, security utility, security management). Security metrics were

presented in the form of attributes of secure systems as well as methods of security measurement. This questions set was meant to be a large set of options from which consistent security opinions could be extracted.

- **Contact Questions:** These requested a survey recipient to identify themselves for the purpose of ongoing communication, should it be required to evaluate survey results.

Note that Survey Recipients saw only three categories: *Demographics, Security,* and *Contact*. *ISACA, non-ISACA,* and *Security Metrics Baseline* questions were merged into one category called *Demographics*.

The survey was vetted by a team of consulting subject matter experts, two of these were Chief Information Security Officers, one was a retired Chief Information Security Officer, and two were highly respected security architects, one from a financial firm and one from a defense industrial base firm. This review team identified language issues that may interfere with accurate responses and also suggested some additions to the categories of metrics to be included. Their suggestions were incorporated into the survey prior to it being released to participants.

As the survey was intended for completion by human participants, it fell within the domain of scrutiny by survey the Stevens Institute of Technology Institutional Review Board. Topics of interest to the review board and corresponding details were:

- Characteristics of the subject population.

  The population is comprised of security subject matter experts. Security professionals are qualifies as experts by being invited to invitation-only security metrics workshops

run by well-qualified program committees. A return rate of about 25% is anticipated. The population is comprised of both male and female adults of average health and of unknown ethnical diversity. There is no intention to target special classes of subjects or those that may be vulnerable.

- The source of research material.

  All data collected is self-reported data collected through an online survey. Identifiable data is data about which industry the security professional works in, and security profession demographic information (e.g. level of education, number of years experience), as well as an option to provide an e-mail for volunteering to explain survey answers in more detail and/or to obtain survey results. The survey system uses a unique code in the survey link in order to allow the survey taker to quit the survey midway through completion and then return and complete later, from the same computer.

- Plans for recruitment of subjects.

  Those invited to take the survey were qualified in two ways. One set of experts was the group of people invited to an invite-only workshop of security metrics experts. The other was drawn from a database of contacts from our set of consulting experts, and so were prequalified based on personal experience. Each recommended expert will be vetted via biography-checking as well as asked to provide demographic information on security expertise in the survey itself. Consent is given by simply choosing to participate.

- Potential risks and procedures for protecting against or minimizing any potential risks.

  Little to no stress is anticipated, as participation is voluntary, and any stress experienced will most likely be due to inexperience with the survey tool (as is typical in online activity) itself rather than any content contained therein. The data collected is not personal in nature and the respondents may choose at any time to discontinue the survey. The survey population is security professionals who are already motivated to share their opinions about security metrics in forums such as workshops. Professional curiosity motivates security professionals to participate in these surveys because they expect to learn from the results.

Based on the above responses to Review Board concerns, approval for the survey was granted.

Appendix C – Survey Analysis Detail

Survey Method

The purpose of the survey was to elicit expert opinions on the properties and measures that are productively used to attribute security to a system. Five top tier security experts, a group comprised two security architects and three CISOs with strong technical background, were provided with a draft survey and asked to identify any ambiguities in it, or other potential difficulties a security expert may have in responding to it. This review team identified language issues that may interfere with accurate responses and also suggested some additions to the categories of metrics to be included. Their suggestions were incorporated into the survey prior to it being released to participants. These CISOs are also known for their participation in industry committees and other professional activities, and they were also requested to provide contact information for security subject matter experts that they felt would be qualified to opine on the survey content. As security experts are not easy to come by, the sample can only be considered a sample of convenience. Additional respondents were solicited from an invitation-only workshop on security metrics and a highly specialized technical security website blog. It is not known how many security experts may have viewed the survey participation request on the blog, so the percent response in this category is not meaningful. Table 1 summarizes the survey sample.

Table 1: Survey Response				
Survey Participant Source:	Total Solicited	Reminders Sent	Total Response	Percent Response
Security SME CISO Contacts	146	8	62	42.47%
Security SME Workshop	58	7	27	46.55%
SubTotal:	204		89	43.63%
Security SME Blog	20	0	20	100%
Total:	224		109	48.66%

The original survey questions asked respondents to assign ranks and weights to metrics. For example, to assign percentage weights to a list of security attributes, and to ensure that the sum of the weights totaled 100%. All of the expert survey reviewers commented that security experts are busy, and tend to get distracted by changes in the threat environments for systems for which they are responsible. For this reason, they advised that the survey questions would have to be more streamlined and easy to answer quickly. This led to changes in questions that asked for rankings and weightings of security attributes in favor of a simple Likert-scale approach to registering opinions about security attributes. An important design criteria for the survey was that it had to take the minimum amount of time required to deliver opinions on the entire field of study that currently constitutes security metrics.

The change in approach was not viewed as a total setback due to known issues with similar studies which solicited rankings and weights. In a similar study with respect to multi-attribute utility measurement in the domain of nuclear power plant planning, Borcherding et.al, used four weighting methods: the ratio method, the swing weighting method, the tradeoff method and the pricing out method [78]. The comparison of results

showed significant consistency and validity problems in the extent to which the results persist in a carefully designed interactive elicitation process. Speculated reasons for this inconsistency ranged from boredom with the information elicitation process to lack of true expertise on the part of the respondents. The study recommended using carefully designed interactive procedures for elicitation. For this reason, the security survey respondents were requested to provide contact information if they would be willing to participate in interactive follow-up if necessary.

The Boercherding study used an Analytic Hierarchy Process (AHP) approach, wherein one assumes that the problem space can be fully described in a way that priorities, allocations, weights, and preference ratios are judgments that can be represented with meaningful numbers which represent the importance of and dependencies between alternative and competing system attributes [79]. This approach was not used in the security survey because decision analysis in security is not as mature as it is in the domain of nuclear power plant planning. Security outcomes cannot yet be quantified in as clear terms, such as lost lives and environmental damage. The literature review of Chapter 2 makes it evident that there is no starting hierarchy that is agreed upon, and yet there is a wealth of candidate attributes for ranking.

Another approach to structuring this type of problem is described by Thurstone, where participants initially are provided with a blank slate, and iterative ranking exercises reduce the population of the overall attribute list [80]. Unfortunately in this study, the time constraints of potential survey respondents made it improbable that many would participate if they had to start with a blank slate. Moreover, an initial set of properties that

professionals currently use are readily apparent from the literature survey in Chapter 2, and so these were used as a starting point.

Both the Boercherding and the Thurstone studies acknowledge that it is necessary to analyze sensitivity to ambiguous questions, as well as any potential environmental changes in criteria that may result in changes in judgments. Decision theory as applied to security has typically concentrated on one aspect of the security problem, which is investments in a single security technology [81]. Thus the security problem, in contrast to that performed by Boercherding and the Thurstone, does not have a framework waiting to be articulated. Rather, this research is necessary *due* to the fact that system security is not yet well understood enough to place a framework around the problem for others to refine with weights. Yet neither do we begin with a blank slate. This situation is typical in any theory construction for attributes that are not well understood. As observed by Wrenn, "We must subject our constructs to measurement if we are to test our theories, but if we were to insist that theory tests wait until we have a fully axiomatic theoretical model, scientific inquiry would virtually halt" [82]. Hence, in addition to the security attribute criteria gleaned from the hypothesis and literature review, the survey contained other questions of multiple types which were designed to provide background "noise" in order to ensure that bias in attribute select choices was minimized. It also allow respondents to clarify their responses with open ended questions and selections of "other".

To answer prior studies' concerns related to ambiguity and environment, attribute-related questions were ranked using three methods: Thurstone's method [80], the One Number

Method [83], and the Survey Rating System based on proportionate number of respondent selections. These calculations are performed as follows:

- Thurstone's Method

  Post-initial ranking, the positioning of items on the Thurstone scale can be found by averaging the percentiles of the standard normal distribution corresponding to the proportions of the respondents preferring one item over each of the others.

- The One Number Method

  The One Number Method focuses on participant registration of strong opinion, and ignores responses that simply agree with a selection presented. Hence, it is calculated by summing the number of "5s" in a rating response, and subtracting the sum of the "1s", "2s" and "3s" from it, then dividing by the total responses. Where this calculation produced equal values, the number of "4s" was used to disambiguate the responses to allow a basis for selecting the order of the final ranking.

- Survey Rating Method

  Each of the 37 relevant questions on the survey received a rating based on a multiple of the number of respondents who selected a given value multiplied by that value. A straightforward calculation of the rating for a question wherein 5 people selected 1, 10 people selected 2, 15 people selected 3, 20 people selected 4 and the remaining 10 selected 5 was computed as:

  $$(5*1 + 10*2 + 15*3 + 20*4 + 10*5) / 60 = 3.33$$

  These ratings were disambiguated by a second order sort by the number of 5s, then 4s, and so on.

This three-part overall ranking was separated into four groups based on convergence of average rankings. The groups were further analyzed using a *Rank Order Centroid* method [108]. This showed that the differences between the weights in some of the Survey Rating System attributes that ended up in different order were small. The resultant rankings were compared and sent to the CISO-level survey respondents who volunteered to be asked follow-up questions.

Survey Results

Qualifications

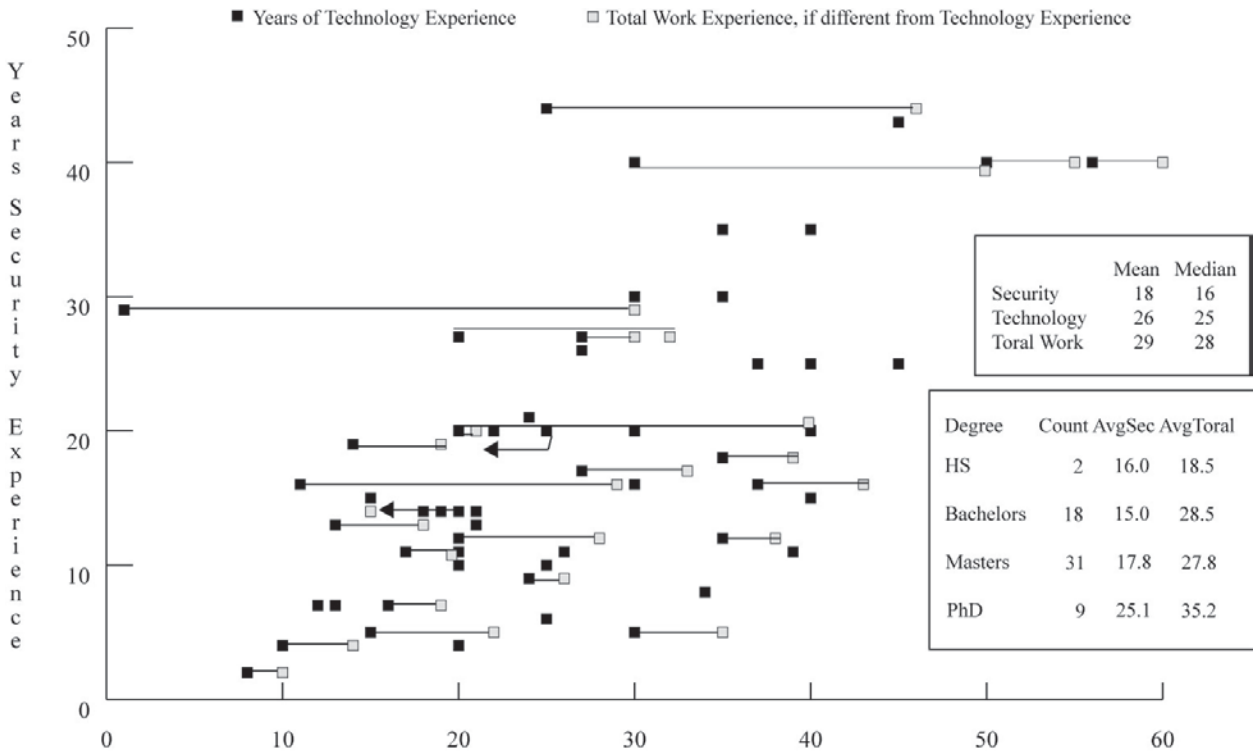
The supposition by the survey reviewers that security experts would get distracted while taking the survey and not finish it was correct. 13 of these had found the survey via the security expert blog site. Therefore, criteria were required to qualify partial respondents for inclusion in results analysis. The criteria were based upon the necessity to include questions on metrics identified in the research hypothesis, as well as a sufficient number of noise questions. This necessity suggested that the criteria include completion of the survey question that contained the widest variety of metrics alternative responses, which was question 21 on the survey. Only 62 respondents of the 109 who started the survey actually completed this question. An additional two respondents were removed from the results because they wrote that they had zero years of security experience. One was a student and another was a network administrator. A few others also missed putting in their years of security experience, but did fill in technology and work experience and were later determined to have been working in security for at least 15 year or more. Some of those who self-selected out of the survey by not reaching question 21 would also have

been removed based on lack of expert qualifications. For example, in response to a question about metrics types, two selected the option, “These terms are unfamiliar to me,” which was purposely included in order to weed out inexpert responses. Three others wrote that their job function was to build, operate or provide project management for security, and these professions are not typically sophisticated in security metrics. However, the ranks of the drop-outs did include at least 10 CISOs, 5 security architects, and a few reknowned security researchers. One researcher complained that the request for assistance with follow-up was improper, given that there was no statement given in advance of taking the survey that identification with the requested. One CISO was from the software industry and did not see a connection between the survey questions and her job function, which was not enterprise but product security. The final count for analysis was 60.

The minimum number of years in the security profession among the 60 was two, but that person had ten years of work experience, and eight years of that experience was in technology. This was also the minimum technology and work experience of the group, no one among the respondents has less than ten years of total work experience. The most experienced in security had 44 years of security experience, 25 years in technology, and 46 years total work experience. The qualifications of the experts are illustrated in Figure 1. Where technology experience and work experience were not the same, they are connected by a line on the graph. Two people reported having a few more years for technology experience than total work experience, and this result is depicted by an arrow pointing to the left in the line which connects them on the graph. The graph shows less

than 60 points because a few respondents had exactly the same number of years experience in all three dimensions. Figure 1 also shows the highest level of education for the individuals. Following the count is the average years in security of individual of that degree level, and the average total work experience at that level.

Figure 1: Survey Respondent Demographics



Two thirds of the participants were active in security professional organizations and over two-thirds had some form of security certification. Seventy-eight percent of the participants were either active or certified. Forty percent of the participants were from the financial industry. This demographic factor was considered large enough to potentially skew the results as financial industry-specific, so a hypothesis was formulated that the distribution of results was the same in this population as compared to the non-financial participants. The Mann-Whitney (Wilcoxon) test was performed on all of the questions

that led to our security attribute ranking, and only one attribute registered a level of significance required to reject that hypothesis. This was the attribute of being able to pass a penetration test. A cross tabulation of industry group with that attribute revealed that financial industry background was strongly correlated with a high rating for penetration testing. This is likely due to the financial industry's relatively higher budgets for hiring outside consultants, and resulting experience that such measures often identify previously unknown vulnerabilities. As this recognition is a sign of experience with a specific tool, rather than being related to financial industry systems, the observation was not sufficiently financial-industry specific to omit either the participants or the question from the sample data. As the Mann-Whitney tests for change in median, an additional test for a more general change of shape, the Kolmogorov-Smirnov test, was performed. The results of both independence tests, and the cross-tabulation results for the question on penetration studies, are included in Appendix E. Given the results, there is no reason to believe that our sample, though skewed toward financial services representation, is not representative of the more general population of security experts sampled.

Rank Results

In the survey, security attributes were rated by experts in six questions, though two of those were confined to systems of a given type, and had lower participation levels. Combining the other four questions provided the general set of opinions on security attributes required for comparison (Questions 21, 24, 25, 26). The three rating methods were compared and disambiguated on this subset of four questions.

The four general questions combined constitute a 44 independent multinomial trial of 5 possible outcomes of the same probabilities. A normal distribution of results would indicate that respondent answers were the equivalent of random selections. This would be the case if the respondents as a whole had ambiguous attitudes toward a given question. By contrast, a positive kurtosis or significant skewness would indicate that the observations are more clustered about an attitude on which respondents agree.

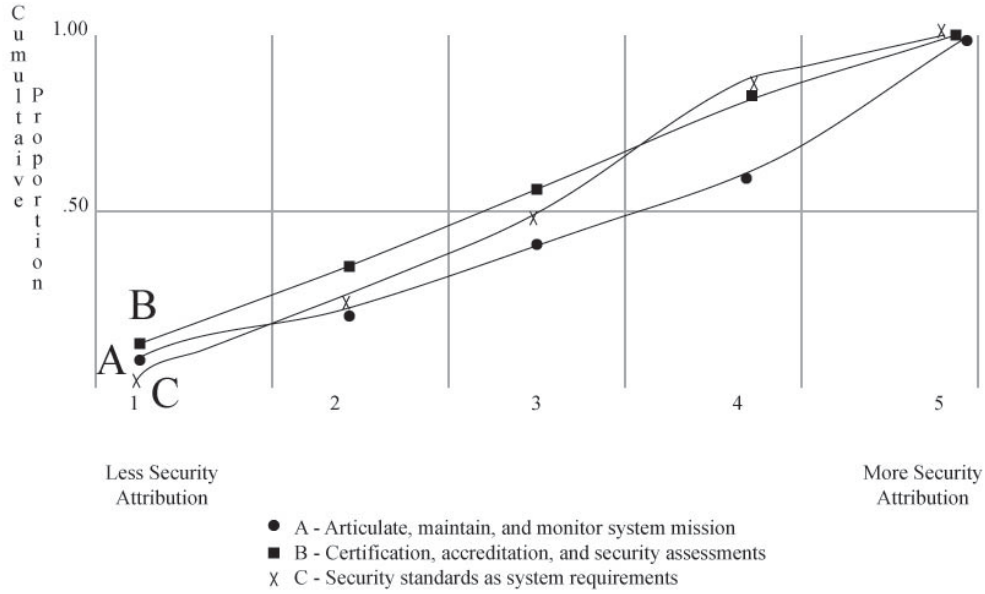
Sorting the statements requires the collected opinions to be compared. As in [80], this was done via a phi-gamma curve as illustrated in Figure 2. The diagram plots the respondent's answers to the first three statements on which they were asked to opine. The steeper the curve associated with a set of opinions concerning the corresponding security attribute, the smaller is the degree of disagreement in the scale by which it was classified by the respondents, so it is a more precise statement. The gentler the slope of the curve, the more ambiguous is the statement. In the example of Figure 2, item C is a less controversial a statement than A or B. Those which are both skewed to the right and have positive kurtosis ranked higher than those with a larger area of the curve in the lower quadrants. Collective responses to any question that approximates a normal distribution or a flat curve are judged too ambiguous to merit inclusion as a security attribute.

Appendix F includes descriptive statistics for all attributes. Those removed due to ambiguity have a skew value below 0.3 and also a central mean (flat) or kurtosis near zero (normal). These were:

- Q21-23-ThreatProb
- Q21-24-DamageProb

- Q24-5-Deliver
- Q24-6-Provenance
- Q25-6-Perform
- Q26-1-IndepComp
- Q26-3-COTS

Figure 2: Example Diagram of Opinions on Security Attribution



The results of the rankings of all three methods are listed in Table 2.:

Table 2: Attribute Rank Order for Survey Responses

Orig Order	Question Label	Thurstone	One Number	Survey Rating
20	Q21-20-IDAuth	1	1	1
27	Q24-4-PassPenTest	2	2	3
11	Q21-11-Incident	3	4	4
36	Q26-4-VaInput	4	3	2

Table 2: Attribute Rank Order for Survey Responses				
Orig Order	**Question Label**	**Thurstone**	**One Number**	**Survey Rating**
1	Q21-1-Mission	5	7	13
8	Q21-8-Awareness	6	5	5
23	Q21-25-ThreatProtProb	7	6	9
14	Q21-14-PhysEnv	8	14	22
15	Q21-15-Personnel	9	12	19
10	Q21-10-Recovery	10	10	7
17	Q21-17-Interfaces	11	11	15
9	Q21-9-SWChange	12	18	10
37	Q26-5-DefOutput	13	8	8
26	Q24-3-PassSecRev	14	20	18
19	Q21-19-AuditTrails	15	15	23
4	Q21-4-Risk	16	9	6
18	Q21-18-Segregate	17	17	17
16	Q21-16-SWIntegrity	18	19	16
7	Q21-7-Acquisition	19	21	24
5	Q21-5-Infrast	20	13	12
6	Q21-6-Features	21	25	21
13	Q21-13-Media	22	26	30
33	Q25-4-Logs	23	16	11
2	Q21-2-Certif	24	32	34

Table 2: Attribute Rank Order for Survey Responses				
Orig Order	**Question Label**	**Thurstone**	**One Number**	**Survey Rating**
22	Q21-22-AssetValue	25	24	29
32	Q25-3-Mgmt	26	28	25
3	Q21-3-Standards	27	30	26
34	Q25-5-BCP	28	23	14
29	Q24-8-FailSafe	29	35	35
28	Q24-7-Interfaces	30	31	33
25	Q24-2-SecAudit	31	29	27
35	Q26-2-Pattern	32	22	20
31	Q25-2-Config	33	27	28
24	Q24-1-RegAudit	34	36	36
12	Q21-12-VendorOver	35	34	32
21	Q21-21-TechCfg	36	33	31
30	Q25-1-Resources	37	37	37

Subsequent Analysis

The result is three ordered lists. Although the rank order of systems properties that merit positive attribution of security are the three types of ranks in Table 2 were different in order, they periodically converged. There were clusters of responses wherein the averaging of responses within ranks show that several sets of values maintained their general order within the more detailed sub-ordering within the clusters. That is, holding

one ordering constant, the sum the ranks for all three of the methods for the corresponding survey question was divided by the rank within the corresponding order. Where these values converged, the ranks within groups were roughly equal. This overall ranking was separated into four groups based on convergence of average rankings holding the Thurstone order constant. The choice of Thurston was based on the scientific validity of that study compared to the other methods. A further test was performed to ensure that the ordering did not overlook the differences in each interval between the ordered ratings. For example, a rank order of 1,2,3 has a different meaning than a rank order of 1, 3.4, 4.6. The groups were further analyzed using a *Rank Order Centroid* method [108]. This showed that the differences between the weights in some of the Survey Rating System attributes that ended up in different order were small. Table 3 shows the four clusters of attributes that resulted from this analysis.

Table 3: Clusters of Ranked Attributes	
1	User identification and authentication
	Withstand targeted penetration attacks by skilled attack teams
	Incident detection and response
	System interfaces accept only valid input
	Articulate, maintain, and monitor system mission
	Security awareness
	Evaluate the extent to which systems are protected from known threats
	Physical and environmental protection
	Personnel screening and supervision

Table 3: Clusters of Ranked Attributes	
2	System recovery planning
	Security features required to maintain integrity over system interfaces
	System and software change control
	System output conforms to well-defined specifications
	Pass internal security review
	Maintain audit trails on use of system functions
	System-level risk assessment
3	Segregate users into groups or roles for access control
	Software integrity preservation
	Due diligence in system and services acquisition
	Infrastructure risk assessment
	Security features that correspond to system functions
	Control over removable media
	Logs that verify that process designed to secure system is followed
	Certification, accreditation, and security assessments
	Quantify the value of assets at risk in system operation
4	Progress in a management plan to secure system
	Use security standards as system requirements
	Successful execution of business continuity procedures
	Fail in denial of service mode
	Maintain integrity of interfaces through system development lifecycle
	Pass security audit

Table 3: Clusters of Ranked Attributes
System follows a commonly used architecture pattern
Percentage of systems or components that have passed security configuration tests
Pass regulatory audit
Oversight of vendor maintenance
Maintain values of standard security variables in system technical configuration
Number of resources consumed in system security-related tasks

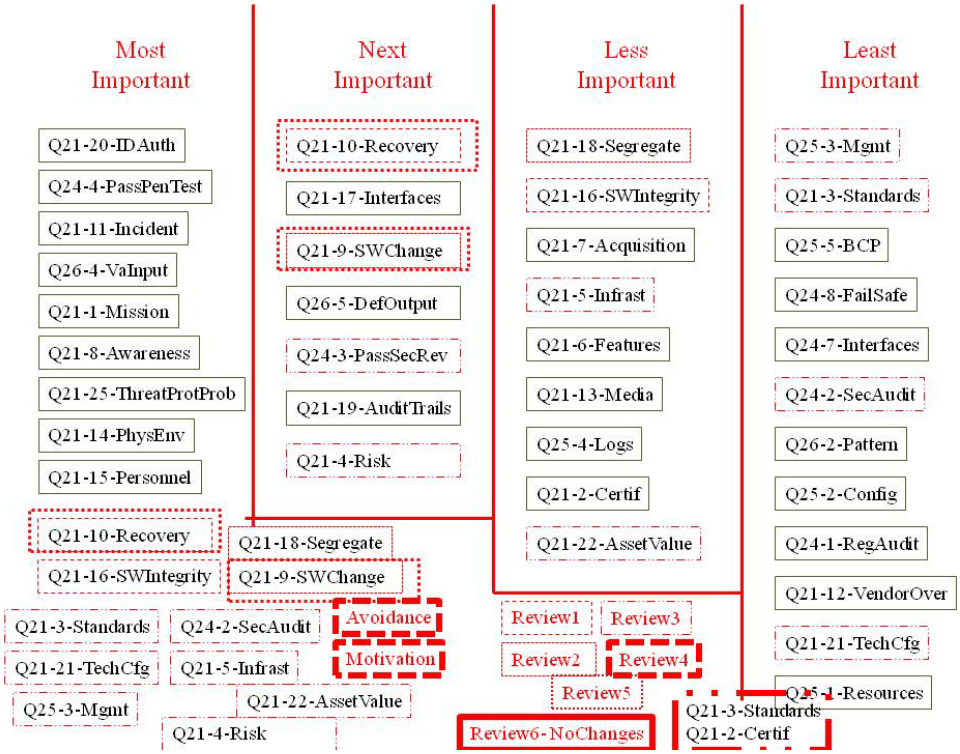
Final Analysis

Of the 29 people who provided an email address for follow-up questions, 19 were either CISOs or consultants with CISO experience. The CISO-level follow-up participants were instructed to review the rankings in the lists attached and make any corrections or comments they thought may be necessary to ensure that this study emphasizes the most important attributes of system security in proper order. Six of these individuals provided detailed feedback. Of those who completed the request for corrections and comments, all but one suggested minor changes in the four categories of groupings, and these are displayed in Figure 3. Only one of the participants suggested that a component level measure (technical configuration) be elevated to "Most Important" status. This person had considerable experience in the component certification and accreditation process, and even had authored a book on the subject [109]. This confirmed some of the other reviewers' comments that bias will of course affect professional judgment. None of the

suggested changes affected the conclusion that the three systems level attributes identified in the research hypothesis are among the most important.

Specific subject matter expert follow-up comments included general disappointment that any security attribute would be considered "not important" as component security could of course be a weak link in a chain or armor. They also commented that responses to the survey were subjective, and complained about the "noise" level of the questions, both of which were, as noted in Section 3.1, intentional.

Figure 3: Rank Shifts Suggested by Survey Follow-Up Respondents



A few respondents that did not provide detailed feedback instead commented either that it was an onerous exercise, or superfluous given the natural bias of participants and inherent limitations of surveys as tools to compare dissimilar concepts.

The full set of survey results in Appendix C includes all comments from all participants. Notable comments supporting this tiered approach to security requirements are:

- *The environment requires easy to understand system documentation from inception to production with security being an identifiable component at all levels. As much detail as is needed to fully describe security related elements/functions is required and development phases are reviewed and accepted or rejected based on completeness and ease of understanding.*
- *System security verification requires an assessment of how the integrated security components combine to defend against, discover or respond to attacks.*
- *Security is an epiphenomenon, a second-order effect of a business process as implemented in a cultural context. As such it is difficult to define repeatable, comparable, quantifiable objective measures of security.*
- *The best security metrics are those that have business correlation, and can be collected analyzed and communicated to support decisions (I assume your context implies that this capability exists, but in truth most organizations struggle reaching a minimal level of maturity) Your question brought to mind a similar question: "What is the best language?" My response to that has always been similar to the one above. Those with the ability to communicate in multiple languages have strong opinions in this area, but if you ask a language professor they will often slap their foreheads and wish that their students knew how to communicate in any language. (My 2 cents from the soapbox)*

Comments also echoed some remarks from Chapter 1 that emphasized the need for reliable security metrics. For example:

- *My use of metrics is not particularly mature or consistent. I rely a great deal on the judgment and consensus of SMEs.*
- *Security Metrics should be used as transitory - they are not true representation of performance or status, but more a convenient means to define targets, benchmarks, status for the temporary time they remain relevant and are not gamed*
- *Like all metrics they become gamed unless reality is more difficult to achieve than the metric itself.*
- *The phrase security metrics means: a migrane.*

The STAC framework provided by this dissertation attempts to answer these concerns and provide some utility to these security experts.