Auditing:
A Journal of Practice & Theory
Vol. 5, No. 2
Spring 1986

Auditor Judgment Concerning Establishment of Substantive Tests Based on Internal Control Reliability

B. N. Srinidhi
and
M. A. Vasarhelyi

ABSTRACT

The evaluation of Internal Control Systems has been considered mainly as a single-stage process in the current literature. This paper attempts to reconcile some apparently discrepant research findings [Gaunmitz et al., 1982; Joyce, 1976] by dividing the process into three different stages: 1) identification, 2) evaluation, and 3) interpretation. The final stage, where auditors restrict substantive tests based on their perception of *internal control strength*, was analyzed in descriptive terms. A high degree of consensus was found in the substantive test planning decisions when the system reliability numbers were provided. Meanwhile, large divergence was observed when only *component* reliability numbers were provided. Therefore, the experimental results indicate that *auditors will disagree on how to aggregate audit evidence, but once one aggregation rule is established, high consensus will follow.* These findings reconcile the above-mentioned, previously discrepant findings and lend credibility to the need and desirability of using internal control reliability *decision aids.* To shed further light on this judgment process, a descriptive mathematical model is also developed and tested, with encouraging results for further research.

INTERNAL control evaluation has long been recognized as an important aspect of external auditing as evidenced by the second standard of field work [AICPA, ¶ 320, 1985]. The mandated requirement of communication of material weaknesses [AICPA, ¶ 323, 1985] and the passage of

B. N. Srinidhi is Assistant Professor of Accounting, Graduate School of Business Administration, New York University, New York, NY 10006. M. A. Vasarhelyi is Associate Professor of Accounting, Graduate School of Business, Columbia University, New York, NY 10027.

the Foreign Corrupt Practices Act [1977] have further enhanced the importance of and the need for a systematic procedure of internal control evaluation. Therefore, behavioral and audit researchers have shown increasing interest in internal control evaluation, as well as its incorporation into an overall audit plan.

Most of the descriptive laboratory studies on internal control evaluation judgments have required the subjects to rate the strength of various hypothetical systems on some scale. For example, Ashton [1974], Ashton and Brown [1980], and Hamilton and Wright [1977] used a six-point scale to evaluate internal control strength. Joyce [1976], however, focused on the time estimates of substantive testing procedures. Mock and Turner [1981] required the subjects to specify the sample sizes for substantive tests. While the former tests showed a high degree of consensus and agreement among auditors, the latter tests did not.

Gaunmitz et al. [1982] tried to reconcile the differences in the findings of Ashton [1974] and Joyce [1976]. They required the subjects to specify both the internal control strengths and audit hour estimates, finding that if auditors make an explicit judgment of the internal control strength and *then* provide audit hour estimates, good agreement is reached in both internal control evaluation and audit hour judgments.

Their conclusions imply that internal control evaluation and an interpretation of its strength in terms of the audit effort are two different stages in the decision process of auditors. By making the auditors state internal control strength, the consensus in audit hour estimates was improved. However, despite the fact that the estimates were correlated, a large variance in hour estimates was found.

Therefore, auditor decisions may be *improved* in terms of agreement by dividing them into different stages and providing linkages between the stages to guide the decision-making process. The possibility of improving decisions motivated a detailed study [Srinidhi, 1984] into the explication and quantification of internal control decisions.

This paper identifies three stages in evaluating internal controls which can be interlinked using *reliability* concepts [Cushing, 1974; Stratton, 1981]. Next, an experiment comparing the consensus in audit planning judgments under two conditions is described and its results discussed. These two conditions relate to: (1) when auditors make judgments given all relevant information about the internal control system, and (2) when the decision stages are *separated* and auditors judge singly on substantive test restrictions given the internal control strength. This design enables the assessment of the value of requiring explicit decisions at each of the specified stages.

The next section introduces the reader to reliability terminology and the identification of decision stages. The third section describes the methodology including the experiment, research design, and data analysis. The fourth and fifth sections, respectively, present the experimental results and conclusions from the study.

Support was provided by the Peat, Marwick, Mitchell Foundation as part of the Research Opportunities in Auditing Program. The views expressed here are our own and do not necessarily reflect the views of the Peat, Marwick, Mitchell Foundation. This support is gratefully acknowledged. The suggestions of Professors D. Miller, G. Shillinglaw, D. H. Bao, and J. Bell, comments from seminars at New York University and University of Florida, as well as the suggestions of two anonymous reviewers, were of great value for this paper.

STAGES OF DECISION-MAKING

Concepts from Reliability Theory

A brief description of terminology from Reliability Engineering is necessary prior to stage identification. In this study as well as in Srinidhi [1984], the accounting system is viewed as a series of different *procedures* (e.g., preparation of purchase order, preparation of journal vouchers) and *controls* (e.g., matching involved with vendor invoice approval). The individual procedures and controls are referred to as *components*. Each component can be characterized by *component reliability*; the likelihood that the procedure (control) is correctly performed (applied). The system, comprising a major *activity* of the firm such as purchasing, is characterized by a *system reliability* score between zero and one, which represents the likelihood that the system output (General Ledger entry of a purchase transaction, in this case) is correct. The system reliability score is a function of the component reliabilities. The function that relates system reliability to component reliability is called the *structure function*.

For example, if the system output is correct only when two components are correct, the two component reliabilities are multiplied to get the system reliability. If the system output is correct when either of the two components is correct, then the system reliability is the complement of the product of the complements of the component reliabilities. In the former case, the structure function has the form:

$$SR = CR1 \cdot CR2$$

and in the latter case it has the form:

$$SR = 1 - (1-CR1)(1-CR2)$$

where SR = system reliability and CRi is the reliability of component i.[1]

Stages in the Measurement of Internal Control Strength

Three distinct stages of measuring the strength of an internal control system are identified in this study:[2]

1. *Identification Stage:* Identification of the components and structure of the system of internal controls. Evaluation of each component (including compliance testing where deemed necessary) and recording the judgments in terms of component reliabilities.
2. *Evaluation Stage:* Aggregation of component reliabilities into system reliability (using the structure function).
3. *Interpretation Stage:* Judgment on how the audit plan should be modified based on the system reliability.

The extant literature has concentrated primarily on the evaluation stage of auditor judgment. Some studies, however [Gaunmitz et al., 1982], have treated the evaluation and interpretation stages together.

This study focuses on the interpretation stage, in which auditors make substantive test planning decisions *given* the strength of internal controls. System reliability is used to measure internal control strength.

Srinidhi & Vasarhelyi [1985] discuss in detail the aggregation stage as well as the usage of a *decision aid* to assess system reliability. An aggregation model was developed to combine the component reliabilities using the structure function to yield system reliability. The natural question is:

> Does the knowledge of the system reliability improve audit planning judg-

[1] For a detailed treatment on how to apply reliability concepts to accounting systems, see Srinidhi [1984].

[2] These three steps are analogous to Simon's [1960] three stages of human decision making: intelligence, design, and choice.

ments (restricting the scope of substantive tests)?

This question is addressed through the examination of auditor judgments in terms of consensus and stability. It is found that both of these measures are high at the interpretation stage, adding to the understanding of the composite (three-stage) judgment process. For further examination, a descriptive nonlinear regression model is fitted to auditor judgment. The results of this effort show both promise and the need for further auditor judgment modeling.

METHOD

A laboratory experiment was conducted to elicit auditors' responses to an instrument that provided details about an internal control system with different system reliabilities. This section describes the experimental task, the research design, and the method used in data analysis.

The Experiment

Seventy-seven auditors from 20 offices of a major CPA firm took part in the experiment. All auditors were trained in the evaluation of internal control systems and had, on the average, about three years of field experience.

The experimental task consisted of evaluating the purchase transaction cycle of a hypothetical firm. The auditors were given a brief narrative of the organization and a description of the procedures and related controls in the purchase transaction cycle. The description was accompanied by a flowchart and other documentation with which the auditors were familiar.

The component reliabilities of four major procedures and controls were provided to the subjects in the first section of the instrument.[3] Each procedure and control was presented at two levels of reliability. Each mix of the component reliabilities was one treatment. This forms a 2^4 factorial design where the four components are the factors and the two component reliability levels are the two treatment levels. This constitutes 16 combinations or treatments. They were presented to each auditor in a randomized sequence to ensure that *sequence* effects did not influence the results. Four of these treatments were then replicated. Care was taken to ensure that the four replicates were presented to that auditor earlier in the sequence to minimize the *memory* effect.[4]

The auditors judged the degree of substantive testing[5] using a seven-point scale. In the first section the auditors went through both the evaluation and interpretation stages without explicitly deciding on the system reliability.[6]

In the second section of the instrument, auditors were provided with a system reliability measure in the same range as the system reliabilities of section 1. Based on the background information and system reliability, the auditors were required to estimate the extent of substantive testing required.[7] Here auditors were provided with

---

[3] In the instrument, five procedure and control reliabilities were presented to the auditors. One procedure was presented at the same level of reliability in all combinations. Therefore, it does not constitute a treatment.

[4] A pilot test was conducted on seven practicing auditors and six academics using the same instrument. One of the questions concerned noticing replicates and remembering earlier answers. No subject could recall whether he/she had answered any replicates.

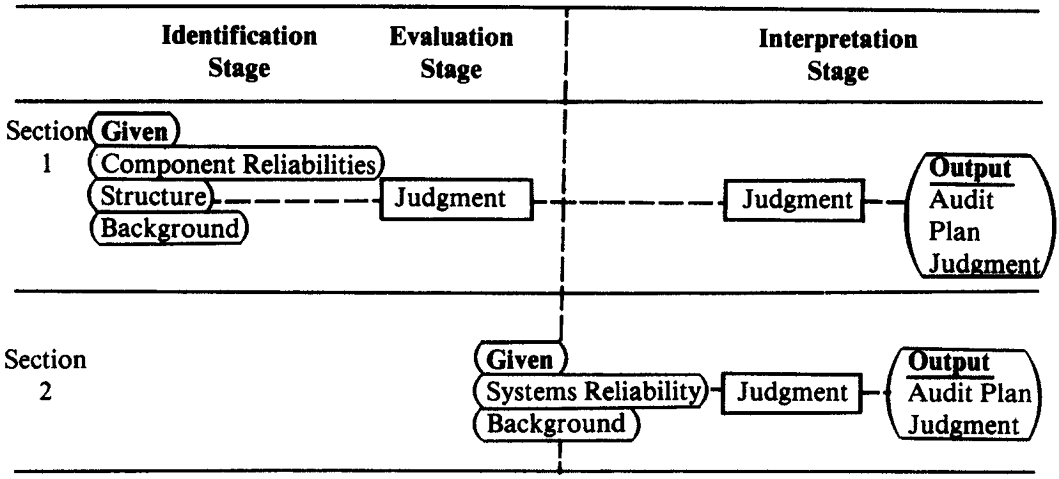
[5] Care was taken to see that subjects did not make system reliability judgments *before* deciding on an audit plan. They had to give the audit plan judgment first. Afterwards, they also indicated their estimate of the probability of error.

[6] The subjects also decided on the system reliability in section 1 *after* they gave their restriction judgment. The purpose of these judgments is to enable the researcher to check whether in this "reverse" judgment, auditors retain similar mappings between system reliability and test restriction as in section 2.

[7] The subjects completed section 1 before section 2. A prior pilot test did not show significant differences in the judgments of auditors that completed section 2 prior to section 1.

FIGURE 1

Decision Making in the Two Sections of the Instrument



the output of the evaluation stage and had only to perform the interpretation stage.[8]

In the last section of the instrument, information was collected on the age, sex, education, and experience of the auditors.

By comparing the audit plan judgments in the second section with those from the first section, the improvements resulting from *knowledge of* the system reliability can be established.

Figure 1 displays the conceptual division of the judgments into three stages and the decision steps involved in the two sections of the instrument.

In the first section of the instrument, the component reliability pattern, the structure, and the background of the firm constitute the cues which affect the audit plan judgment. The structure and the background are kept constant and the component reliability pattern is changed across the 16 original combinations and four replicates. The component reliability patterns constitute the independent variable and the degree of substantive test restriction constitutes the dependent variable.

In the second section of the instrument, the system reliability constitutes the independent variable and the degree of substantive test restriction the dependent variable.

In comparing the decision process of section 1 with that of section 2, consensus between auditors is used as one of the measures of the quality of judgment. Section 2, where the system reliability is provided, is expected to show higher consensus.

If judgment consensus is high in the second section, then the variation of "mean judgment" with system reliability will not be very different in shape from the variation of individual auditor judgments. This provides an opportunity to develop a one-to-one functional mapping between the mean auditor judgment on test restriction and the system reliability. The shape parameter of such a descriptive function will be a characteristic of the group of auditors making the decision. A nonlinear regression method will be used to determine

---

[8] There were no replicate judgments in section 2.

the functional form and the parameters of such a function.

Data Analysis

In situations where a normative criterion is not available, agreement among decision makers has been used as a measure of expertise. There are two dimensions of such agreement. First, *consensus* level indicates the agreement between different decision makers in the same situation. For example, the degree to which different auditors agree on substantive test restriction for a given reliability in section 2 is a measure of consensus in that decision stage. The second dimension, *stability*, indicates the degree of agreement between the two judgments made by the same decision maker on the same situation at two points in time. The time period should be such that the decision maker does not "remember" the first situation or the judgment when the second decision is made. In section 1, the agreement between auditor judgment on the original treatments and replications is a measure of stability.

If there is no expertise available and many different judgments are possible, one would expect each judgment to be made with equal probability [Einhorn, 1974]. Expertise involves knowledge about the entity on which judgment is being made. If that knowledge is shared, the probability of one judgment should become higher than that of the others. Such expertise is reflected in the "peakiness" of the probability distribution over possible judgments which is manifested empirically by a high stability and consensus among "experts." If each possible judgment is observed to be made with equal probability (no stability or consensus), one can infer a lack of expertise. The converse — that a high degree of consensus and stability indicates expertise — does not follow. However, in the absence of a better measure, consensus and stability are used as the indicators of expertise.

The level of consensus between any two auditors typically is measured by the correlation between the responses of the two auditors. The overall measure of consensus used is the mean inter-auditor correlation [Ashton, 1974; Joyce, 1976]. The use of product moment correlation assumes the data to be an interval scale and results in a greater weight being placed on highly opinionated judges who give extreme judgments. To ensure that the extreme judgments do not unduly bias the results, both product moment and rank correlations are presented. The coefficient of concordance is used as the nonparametric measure of the mean consensus level.

Test-retest correlation is used to evaluate the stability of auditor judgments. In the last section of the instrument, the memory effect was retested and was found to be insignificant. Given that response to replications is not influenced by memory effect, a high test-retest correlation would indicate intra-auditor stability.

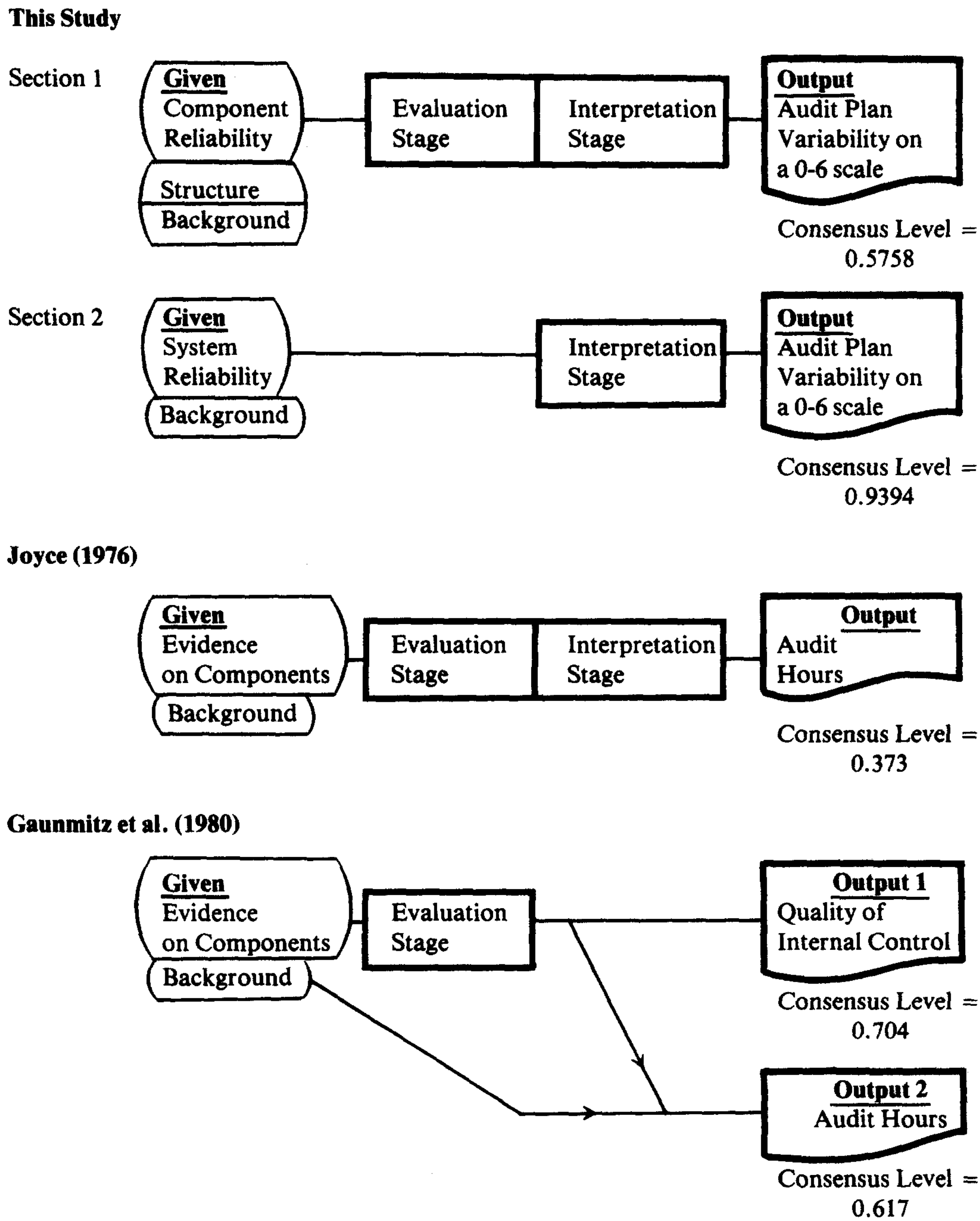
The auditor judgment on the degree of restriction of substantive testing is the dependent variable in both the sections. It is measured on a seven-point scale.[9]

Figure 2 compares the results of this and earlier studies [Gaunmitz et al., 1982; Joyce, 1976]. This comparison is meaning-

[9] The rating scale used:

0 - No reliance and timing restricted to the year-end.
1 - Low reliance but timing restricted to the year-end.
2 - Low reliance with timing restricted to within one month of year-end.
3 - Moderate reliance with timing restricted to within one month of year-end.
4 - Moderate reliance with timing restricted to within two months of year-end.
5 - High reliance with timing restricted to within two months of year-end.
6 - High reliance with timing allowed to be more than two months from year-end.

FIGURE 2
General Comparison of the Results of this Study with Selected Previous Studies [Joyce, 1976; Gaunmitz et al., 1982]



ful only at a general level because of the following differences:

— This study provides specific inputs with reliability numbers as the measure. Outputs are also measured differently. The output from the "evaluation stage" is a system reliability number between zero and one; in other studies this measure was a qualitative rating scale. Joyce [1976] and Gaunmitz et al. [1982] used audit hours as the audit plan variable, while this study uses a scale that incorporates both timing and extent of substantive tests.

— This study conceptually divides decision making into stages with defined inputs and outputs. Joyce does not present such a division, while Gaunmitz et al. imply such a division but do not clarify the decision's sequential nature nor its inputs and outputs.

RESULTS

Sample Characteristics

The sample is homogeneous in age, education, and training, and consists of 77 professional auditors. Most of the auditors were about 25 years of age with two to three years experience in auditing. All of the subjects had undergone the basic training given in the firm up to the senior level. Almost all of the subjects had an accounting education, either at the undergraduate or at the graduate level. A majority had experience in documenting and evaluating the purchase transaction cycle internal control system.

The subject selection procedure restricted the generalizability of the study's findings, as the findings apply to a *single firm*. On the other hand, this allows for a much more homogeneous subject sample and a common understanding of the specific meaning of internal control evaluation ratings. Most firms have their own internal control evaluation procedures, but these vary substantially among firms [Cushing & Loebbecke, 1983]. Sample choice from different firms may introduce substantial ambiguity into the subjects' task or require thorough training of subjects in the specific internal control representation used in the instrument.

Consensus in Auditor Judgments

Table 1 gives the interauditor judgment correlations on substantive test restrictions (denoted as "ratings") when only the component reliabilities are provided (column 1) and when the system reliability numbers are provided (column 2). The mean interauditor correlation in the former case is 0.5758 and in the latter case is 0.9394. The important finding, however, is what happens when the two stages are separated and only the interpretation stage is analyzed. Given the system reliability, the mean consensus is 0.9394. The difference between the Kendall coefficients of concordance is significant at the 0.00001 level.

These findings need to be interpreted in light of earlier findings. Joyce [1976] found that the consensus level in audit hour estimation when the two stages are combined is very low (i.e., 0.373) with a range of -0.687 to 0.937. In this study, the comparable consensus level is that in the first section with a mean of 0.5758 and range of 0.260 to 0.718. The relative narrowing of the range may be attributable to the more homogeneous subject population. Allowing for the differences mentioned earlier, the results seem to be consistent across these two studies.

Gaunmitz et al. [1982] allowed an explicit *judgment* of internal control quality and found a consensus level of 0.617 in audit hour estimates. They also report a wide variance in the audit hour estimates between different auditors. In this

TABLE 1
Consensus Data — Number of Auditors = 77

	Ratings Based on Compotent Reliabilities	Ratings Given the System Reliability
1. Mean Pearson Correlation Coefficient	0.5758	0.9394
2. Standard error of the mean	0.0093	0.0032
3. Kendall's Coefficient of Concordance	0.5566	0.9465
4. Range of the Pearson Correlation Coefficient	.260 - .718	.827 - .963

study, however, providing system reliability estimates increased the consensus level to 0.9394 with a range of 0.8265 to 0.9634. There are two main reasons for this difference. First, in the Gaunmitz et al. study, the two stages are not separated and no overall reliability measure was provided, i.e., the variance in the "quality" judgment was carried over to the audit hour estimate. Second, the audit hour estimates may be based on different assumptions on timing by the auditors.

The mean test-replication correlation of auditor judgments of substantive test restrictions was 0.759 (Spearman rank correlation = 0.751). Though the measurement of stability was slightly different in Ashton [1974], the results obtained here are in the same range as in that study. Ashton reports stability from a low of .43 to a high of .96 with an average of .81. Ashton and Brown [1980] report a stability measure from a low of .62 to a high of .99 with an average of .91.

Functional Relationship

The "rating function" of an auditor is the functional relationship the auditor used in the experiment between system reliability numbers and substantive test restriction judgments, denoted by the term "system ratings." The "mean rating function" is the functional relationship between the mean reliabilities and mean ratings.

The existence of a very high level of agreement on substantive test restriction judgments, given system reliability, makes a functional relationship between system reliability and mean auditor judgment meaningful. A function that can describe this relationship is the heuristic used by the group of auditors in making the judgment. The parameters of the function then can be viewed as the characteristics of the group.

Table 2 gives the mean system ratings that the auditors associated with mean judged system reliabilities based on the first part of the instrument. The corresponding plot is given in Figure 3.

The auditors adhered to certain predictable patterns. It was expected that the auditors would not recommend any restriction of substantive tests below a minimum reliability level. They chose a cut-off reliability of 0.50. Predictably, the restriction decisions were much more sensitive at the up-

TABLE 2
Substantive Test Restriction Ratings When System Reliabilites are Given

Reliability	Rating
Less than 0.50	0
0.60	0.104
0.70	0.351
0.75	0.623
0.80	1.065
0.85	1.727
0.90	2.506
0.95	3.571
0.98	4.558
0.99	5.623
1.00	6.000

FIGURE 3
Mean Rating Function

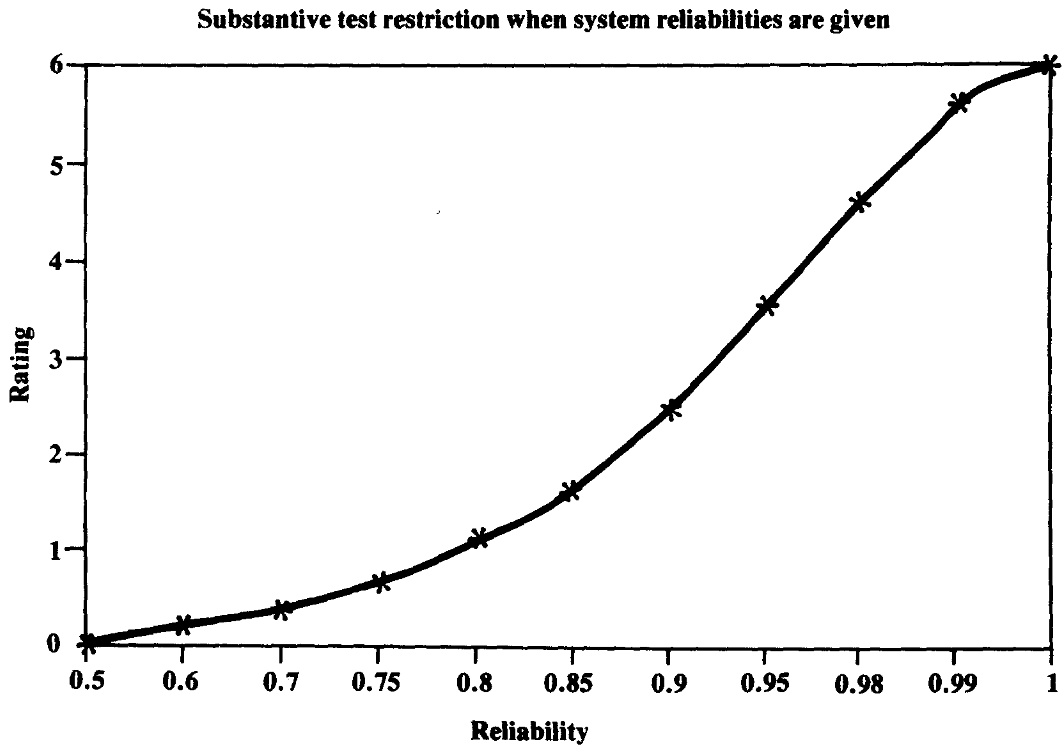


TABLE 3
Results of Nonlinear Regressions on Rating Functions

Variables: x: system reliability (judged or given)
y1: mean ratings given by auditors when reliability is also judged
y2: mean ratings given by auditors when reliability is provided

Model: $y = a(e^{kx} - 1)$
$a = 0.00073$ $k = 9.01$

Source	Degrees of Freedom	Sum of Squares	Mean Square
Model	2	146.62	73.31
Residual	9	0.36	0.04

per end of the reliability scale, i.e., between 0.98 and 1.00. A convex function could describe the responses of the auditors.

The above results are likely in an audit setting. For low values of reliability, the question of restricting substantive testing does not arise. It is only at the upper end of the reliability scale that auditors can reduce substantive testing. Clearly, the restriction of tests will be more pronounced as the reliability approaches one. This means that any functional form which "makes sense" in an audit setting should have increasing marginal test restriction as reliability increases towards one. In the absence of a more specific known relationship, it can be expected that the increase in marginal test restriction is proportional to the increased reliability. A function that has this property is an exponential of the form:

$$Y = a(e^{kx} - 1),$$

where x is the system reliability, Y is the mean rating, a is the scale parameter, and k is characteristic of the transformation.[10] The results of nonlinear regressions on mean rating functions are given in Table 3. In both cases, the model accounts for most of the variance in the data, as can be seen by comparing the mean square for the model (73.3) and for the residual (0.04). The parameter k which is characteristic of the transformation is 9.01. This number represents the sensitivity of the group of auditors to the variations in reliability. A higher value of k indicates that the audit group is more sensitive to the variation in reliability than for a lower k. This provides a way of comparing different groups of auditors on their sensitivities to system reliability or the same group of auditors under two different audit situations.[11]

[10] In the absence of a theoretical normative relationship, descriptive relationships are often hypothesized. For example, Stringer [1975] hypothesized a descriptive relationship between the substantive test reliability S, combined reliability R, and internal control reliance C in the formula $(1-R) = (1-S)(1-C)$. An improvement of this model was suggested by Warren [1975].

[11] The subjects also used the rating judgment to infer back to the system reliabilities in section 1 (see footnote 7). Using a similar nonlinear regression on those data, k was found to be 7.62. This means that when other data such as component reliabilities are available, the auditors are less sensitive to the system reliability number in their test restriction decisions.

CONCLUSIONS

The purpose of this paper was to identify and analyze auditor interpretation of internal control system strength in terms of planned substantive audit test restrictions. In particular, this study attempted to separate the interpretation stage by clearly defining inputs and outputs for each stage. In the earlier studies, auditors seemed to agree on internal control assessment but diverge in hour estimates. This study focused on whether this divergence in audit plans could be observed when system reliabilities were provided.

This study expanded on prior studies by clearly distinguishing among the stages, allowing "broken-down decisions" to be made in each stage and using a different (and perhaps improved) audit planning variable where both the *extent and timing* of substantive tests were included.

The method employed was one of laboratory experimentation in which practicing auditors were asked to respond to two situations, one in which they had to judge substantive test restriction based on component reliabilities and another in which they had to judge the substantive test restriction when system reliability was provided for them.

The auditors exhibited a great degree of consensus in their substantive planning decisions when the system reliability numbers were provided, though large divergence was observed when only component reliability numbers were provided. The main conclusion is that auditors tend to disagree on how to aggregate the evidence from the system in order to arrive at a system reliability measure. Once a system reliability measure is agreed upon, however, there is less disagreement on the degree of substantive test restriction. This indicates the need for decision aids that normatively provide a consistent aggregation rule.

Conceptually, the aggregation of evidence into a system reliability measure can be modeled as an objective statistical process. Its interpretation in terms of audit planning, however, is a process in which the training and experience of an auditor are important. The findings of this study are consistent with Winkler and Murphy's [1968] assertion that professionals such as auditors can be expected to be "substantive" experts but not "normative" experts. "Substantive" relates to the core of professional knowledge, while "normative" refers to statistical or mathematical information processing.

An attempt was then made using a nonlinear regression method to build a descriptive model of auditor judgment. An exponential function performed very well in describing the auditor judgments and provided some insights into their decisions. Admittedly, mathematical modeling of auditor judgments in this area is in an exploratory stage and caution has to be exercised in reaching any conclusions. It is very important, however, to continue researching this area because it provides the basic information on building expert systems to perform these tasks.

REFERENCES

American Institute of Certified Public Accountants, *Statements on Auditing Standards* (New York: AICPA, 1985).

Ashton, R., "An Experimental Study of Internal Control Judgments," *Journal of Accounting Research* (59, 1974), pp. 143-157.

——— and P. R. Brown, "Descriptive Modeling of Auditors' Internal Control Judgments: Replication and Extension," *Journal of Accounting Research* (Spring 1980).

Cushing, B. E., "A Mathematical Approach to the Analysis and Design of Internal Control Systems," *The Accounting Review* (49, 1974), pp. 24-41.

——— and James K. Loebbecke, " Comparison of Audit Methodologies of Large Accounting Firms," Unpublished working paper, University of Utah, July 1983.

Einhorn, R., "Expert Judgment: Some Necessary Conditions and an Example," *Journal of Applied Psychology* (October 1974), pp. 562-571.

Gaunmitz, B. R., T. R. Nunamaker, J. J. Surdick, and M. F. Thomas, "Auditor Consensus in Internal Control Evaluation and Audit Program Planning," *Journal of Accounting Research* (Part II, Fall 1982), pp. 745-755.

Hamilton, R. E. and W. F. Wright, "The Evaluation of Internal Controls Over Payroll," Unpublished manuscript, Stanford University, 1977.

Joyce, E., "Expert Judgment in Audit Planning," *Supplement to the Journal of Accounting Research* (14, 1976), pp. 29-60.

Mock, T. J. and J. L. Turner, "Internal Accounting Control Evaluation and Auditor Judgment," Audit Research Monograph (New York: AICPA, 1981).

Simon, H. A., *The New Science of Management Decision* (Harper and Bros., 1960).

Srinidhi, B. N., *Probability Modeling of Internal Control Systems in Audit Decision Making*, Ph.D. Dissertation, Columbia University Graduate School of Business, 1984.

——— and M. A. Vasarhelyi, "Reliability Theory & Internal Control Evaluation," Unpublished paper to be presented at the Audit Symposium, University of Southern California, 1985.

Stratton, W. O., "Accounting Systems: The Reliability Approach to Internal Control Evaluation," *Decision Sciences* (12, 1981), pp. 51-67.

Stringer, K., "A Statistical Technique for Analytical Review," *Supplement to the Journal of Accounting Research* (15, 1975), pp. 1-9.

Warren, C. S., "A Statistical Technique for Analytical Review — A Discussion," *Supplement to the Journal of Accounting Research* (13, 1975), pp. 11-14.

Winkler, R. L. and P. Murphy, "Good Probability Assessors," *Journal of Applied Meteorology* (October 1968), pp. 751-758.