Audit Risk Model:
A Framework for Current Practice and Future Research

Gary L. Holstrum
James L. Kirtland
Deloitte Haskins & Sells

The authors wish to acknowledge the pioneering work in this area by Kenneth W. Stringer. We are also grateful for the helpful comments of participants at the 1982 DH&S AuditSCOPE Research Update Seminar and especially Barry Cushing, Bart Ward, Ted Mock, Ed Joyce, and Bill Felix.
AUDIT RISK MODEL:
A FRAMEWORK FOR CURRENT PRACTICE AND FUTURE RESEARCH

The purpose of this paper is to discuss how an audit risk model, such as the one described in SAS #39, *Audit Sampling*, can provide the appropriate nucleus for both current practice and future research in auditing. Business risk is distinguished from audit risk and each is discussed. A practical audit risk model is proposed that incorporates four risk components: inherent risk, control risk, analytical review risk, and substantive test of details risk. The model integrates various statistical concepts and techniques, such as the Firm's STAR program for analytical review and Audit Sampling Plan for tests of details. The paper discusses the basic structure of the risk model, related statistical applications, and Firm guidelines for assigning audit reliance (risk) to the various risk model components. The paper (1) proposes a practical approach designed to measure and control the risk components in current audit practice, (2) discusses unsolved problems and sources of error in estimating the risk components, and (3) suggests future directions for research. The role of auditor judgment in the quantification of audit risk is analyzed, including the benefits and limitations of objectively and subjectively quantifying the four individual risk components and the resultant overall risk.
BUSINESS RISK

Business risk is conceptually distinct from audit risk. Audit risk is the risk of reporting incorrectly on the financial statements, while business risk relates to the consequences to the auditor or the auditing firm arising from any litigation or criticism concerning the auditor's work or the client's audited financial statements. The auditor is exposed to business risk regardless of whether the financial statements are presented in conformity with generally accepted accounting principles or whether the auditor has complied with generally accepted auditing standards.

There are several aspects of business risk and many potentially adverse consequences to the auditor. Business risk includes (1) the risk of the auditor's work being challenged, (2) the risk of unsuccessfully defending against the challenge, and (3) the potential loss to the auditor. Potential adverse consequences to the auditor include financial losses, damage of reputation, time spent to defend the auditor's position, and loss of clients.

Though conceptually distinct, business risk is associated with audit risk. Business risk increases when audit risk increases and when there has not been compliance with professional standards. As the amount of misstatement of the financial statements increases, the probability of negative consequences to the auditor also increases. In an audit engagement, if factors exist that are normally associated with increased business risk, we may attempt to control our business risk by correspondingly increasing our audit work beyond that normally performed in an audit or by taking unusual
audit precautions. We do not, however, decrease our audit effort below that required by generally accepted auditing standards or by the Firm's commitment to audit quality, even in situations where the expected business risk associated with an audit is low.

AUDIT RISK

Audit risk is the risk that the auditor may report incorrectly on the financial statements. The third standard of field work refers indirectly to audit risk by requiring that the auditor have a reasonable basis for expressing an opinion. Even though sufficient, competent evidential matter may be obtained to afford a reasonable basis for an opinion, the auditor still accepts some level of audit risk because not all of the available evidence has been examined. The auditor's report does not provide certainty, but rather only adds credibility to the client's financial statements. As discussed in SAS No. 31, _Evidential Matter_ (paragraphs 19 through 21), auditors are seldom convinced beyond all doubt with respect to all aspects of the financial statements being examined. Because of the economic limits within which the auditor typically works, his opinion needs to be formed within a reasonable length of time and at reasonable cost. Generally, there should be a rational relationship between the cost and the usefulness of the evidential matter obtained. The auditor must decide however, whether the evidential matter available within the limits of time and cost is, in the auditor's professional judgment, sufficient to justify the expression of an opinion.
The principal audit risk for the auditor is the risk of providing assurance that the financial statements are free from material misstatement when they are not (i.e., the auditor may express an unqualified opinion when the financial statements contain a material misstatement). SAS No. 39, *Audit Sampling*, refers to this portion of audit risk as ultimate risk.

AUDIT RISK MODEL

The risk model that underlies the DH&S audit approach is very similar to, and consistent with, the one discussed in SAS No. 39. This joint risk model considers ultimate risk in relation to the individual risks for each of its separate components and is summarized by the following formula:

\[ \text{UR} = \text{IR} \times \text{CR} \times \text{AR} \times \text{TD}. \]

The factors in this model are defined as follows:

- **UR** is the ultimate risk that the financial statements are materially misstated, after completion of the audit.
- **IR** is the inherent risk of material misstatement (errors or irregularities) occurring in the process of preparing financial statements, without consideration of control procedures.
- **CR** (control risk) is the risk that the system of internal accounting control will fail to detect a material misstatement, given that it has occurred.
- **AR** is the risk that our analytical review procedures will fail to detect a material misstatement, given that it has occurred and has not been detected by the system of internal accounting control.
TD is the risk that our substantive tests of details will fail to detect a material misstatement, given that it has occurred and has not been detected by the system of internal accounting control.

To illustrate how the above formula works, assume an IR of 100%, CR of 37%, AR of 50%, and TD of 28%. The result of applying the formula would be as follows:

\[
UR = \text{100}\% \times 37\% \times 50\% \times 28\% \\
UR = 5\%.
\]

In our sampling plan, we utilize reliability factors (R) that are associated with particular confidence levels and risk levels. The relationships between the reliability factors and corresponding confidence levels and risk levels are shown below.

<table>
<thead>
<tr>
<th>Reliability Factor (R)</th>
<th>Confidence (Reliability) Level</th>
<th>Risk Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.0</td>
<td>95%</td>
<td>5%</td>
</tr>
<tr>
<td>2.3</td>
<td>90</td>
<td>10</td>
</tr>
<tr>
<td>2.0</td>
<td>86</td>
<td>14</td>
</tr>
<tr>
<td>1.5</td>
<td>78</td>
<td>22</td>
</tr>
<tr>
<td>1.3</td>
<td>72</td>
<td>28</td>
</tr>
<tr>
<td>1.0</td>
<td>63</td>
<td>37</td>
</tr>
<tr>
<td>0.7</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>0.5</td>
<td>39</td>
<td>61</td>
</tr>
<tr>
<td>0.0</td>
<td>00</td>
<td>100</td>
</tr>
</tbody>
</table>

The reliability factor represents an exponent related to the risk percentage in the sampling formulas. Therefore, the reliability factor related to a given risk, when added to the reliability factor related to another given risk, will result in the reliability factor related to their joint risk. We find it more convenient to
work with the reliability factors than with the corresponding risk percentages for two reasons. First, since the reliability factors are additive (whereas the risk levels are multiplicative), reliability factors are much easier to process intuitively (through human information processing) in combining and integrating risks from various audit components. This facilitates our understanding and our audit judgments regarding joint risk from all aspects of the audit. Second, the reliability factors are specific parameters in the Poisson distribution, which underlies our Audit Sampling Plan. Consequently, the wide use of reliability factors in our Sampling Plan (for tests of details) facilitates their easy use in other components of the risk model. Their relationship is illustrated as follows (continuing with the above example):

<table>
<thead>
<tr>
<th>Risk Levels</th>
<th>IR</th>
<th>CR</th>
<th>AR</th>
<th>TD</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk Levels</td>
<td>100%</td>
<td>x</td>
<td>37%</td>
<td>x</td>
<td>50%</td>
</tr>
<tr>
<td>Reliability Factors (R)</td>
<td>0</td>
<td>+</td>
<td>1.0</td>
<td>+</td>
<td>0.7</td>
</tr>
</tbody>
</table>

THE ROLE OF AUDIT JUDGMENT IN THE QUANTIFICATION OF AUDIT RISK

The quantification of risk is in keeping with our general view that an auditor can make better and more consistent judgments when risks are quantified. We recognize, however, that the auditor has to assess much information where risk is not susceptible to quantification and that quantification of other aspects of risk is necessarily somewhat subjective. Objectively quantifying all audit risks is a goal that probably is unattainable and that probably
would not be cost-effective. The objective of our audit risk model, therefore, is to optimize the blend of objective quantification of audit risk with subjective judgment in order to perform an audit in the most effective and efficient manner.

We view subjective quantification as a useful way to systematically include some information in the model that otherwise would have to be considered in an even more subjective manner. Subjective quantifications have always been a formal part of our system. Reliance on internal accounting control is a good example. We strive, however, to make these applications as objective as possible by providing objective criteria, because we believe that the more objective the model, the better we can maintain consistent quality in our audits.

At times, in the process of making subjective judgments about the various risks, conservative estimates are used. For example, the assignment of a 100% risk for a given condition does not imply certainty of occurrence but is only a conservative choice to place no reliance on that particular component even though the risk might, in fact, be less than 100%. Also, the assignment of a given level to control risk, or to some other source of reliance, may be based on subjective judgment and may not be as precise as it may appear in the computations.

Although the risk model is conceptually simple, a significant amount of professional judgment is required in applying the model in any audit situation. As to IR and CR, there are presently no totally objective techniques available for quantifying the risks.
The sampling risk portion of TD can be objectively quantified through statistical sampling techniques, and certain aspects of AR can be quantified through a properly designed approach to regression analysis or some similar objective technique. Even though portions of the model are based on subjective judgments that cannot be quantified precisely, we believe that the model helps improve the degree of correspondence between auditing procedures being applied and the amount of risk the auditor is willing to accept.

MEASURING AND CONTROLLING RISK COMPONENTS

Ultimate Risk

SAS No. 39 defines ultimate risk as the allowable risk that "...monetary errors equal to tolerable error might remain undetected in the account balance or class of transactions after the auditor has completed all audit procedures deemed necessary."¹ This is generally planned by the auditor so that the amount of risk that exists at the end of the audit is limited to a level considered appropriate.

The auditor's decision concerning the appropriate level of ultimate risk is arrived at in conjunction with other subjective decisions. Auditing is a process of gathering evidential matter from

¹"Tolerable error" is related to, but not synonymous with, material error. For the financial statements as a whole, tolerable error may be established at an amount equal to financial statement materiality or, because of management expectations concerning audit precision, may be less than a material amount. Tolerable error for individual accounts, locations, or segments of an audit needs to be established so that aggregate error for the entire audit does not exceed an amount considered tolerable for the financial statements as a whole.
many sources, evaluating that evidence, and combining those evaluations to reach an overall conclusion. Much of this process is based on the auditor's professional judgment rather than precise mathematical quantification of results.

Each of the components of ultimate risk can be affected by sampling risk, nonsampling risk, or both. **Sampling risk** is the risk of drawing an erroneous conclusion because the auditor has examined less than all of the items in the population. For example, in confirming receivables or in testing for compliance with pertinent internal accounting control procedures, an auditor may come to a conclusion different from the one that would have been reached had those procedures been applied in the same manner to the entire population. **Nonsampling risk** includes the aspects of the components of ultimate risk that are not attributable to sampling. An auditor may apply an auditing procedure to every transaction or balance and still fail to detect a material misstatement that has occurred or a material weakness that exists in internal accounting control. Nonsampling risk includes the possibilities of selecting auditing procedures that are not appropriate to achieve the specific objective and of failing to recognize deviations or errors included in the documents examined, which would make the procedure ineffective even if all items were examined. Although nonsampling risk normally cannot be measured, it can be controlled by adequate audit planning, appropriate selection of audit procedures, and proper performance and supervision of audit work.
Statistical sampling literature divides sampling risk into alpha risk and beta risk. **Alpha risk** is the risk that the evidence erroneously fails to support the acceptability of the population when it should be accepted. **Beta risk** is the risk that the evidence erroneously supports the acceptability of the population when it should be rejected. These two risks apply to both compliance and substantive tests. SAS No. 39 introduced different terminology for alpha and beta risks. Alpha risk is referred to as the risk of incorrect rejection for substantive tests and the risk of underreliance on internal accounting control for compliance tests. Beta risk is referred to as the risk of incorrect acceptance for substantive tests and the risk of overreliance on internal accounting control for compliance tests. Alpha risk relates to the efficiency, not the effectiveness, of the audit because the auditor's response to a potential alpha risk will normally be to extend the auditing procedures. The beta risk concepts relate to the effectiveness of the audit in detecting material misstatement. Therefore the risk model focuses primarily on the beta risks and not the alpha risks.

Within DH&S, ultimate risk (UR) has been set, as a matter of Firm policy, at 5% for ordinary audit situations. This is based on our professional judgment and is consistent with examples of ultimate risk often cited in the professional literature. Although 5% might appear to be a high risk level when compared to other (non-auditing) applications of statistical sampling, we believe that this level is appropriate for auditing because (1) our experience indicates that the inherent risk of misstatement is normally less than
100%, (2) substantial corroborating evidential matter is usually available that is not comprehended in the model, and (3) special audit risk situations (discussed later in the paper) are identified and addressed separately. Future research and development may provide practical techniques for measuring these factors and incorporating them into the basic risk model.

Inherent Risk

Inherent risk affects ultimate risk and is closely related to control risk because, as the effectiveness of a system of internal accounting control increases, the probability of prompt detection of material misstatement increases and, therefore, the likelihood of those misstatements occurring would be expected to decrease. Although we can conceptually distinguish between inherent risk and control risk and although audit experience indicates that inherent risk is less than 100%, our ability to distinguish them on a practical level is limited. Conceptually, to measure IR independently, we would need to estimate the probability distribution of errors or irregularities arising from the flow of transactions through the accounting system, exclusive of any consideration of internal control procedures. However, in a practical (or empirical) sense, it is not possible to measure or objectively estimate such a probability distribution of errors or irregularities. This is because there is no existing population of transactions that are processed through the accounting system without being subject to the client's existing internal accounting control system, since some of the control procedures (primary controls) are applied during the
processing of transactions. Therefore, we normally do not give explicit recognition to an evaluation of inherent risk in our applications of the audit risk model.

In situations where we perceive an unusually high level of inherent risk, we may judge that a special audit risk exists. The identification of and responses to special audit risks are discussed later in the paper.

Control Risk

A conceptual outline of the steps involved in evaluating control risk is shown in Exhibit A.

EXHIBIT A

THE PROCESS OF EVALUATING CONTROL RISK

1. Understanding the flow of transactions through the client's accounting system.

2. Identifying possibilities for errors and irregularities to occur in the accounting system.

3. Understanding the design of the client's system of internal accounting control, including:
   a. Primary controls,
   b. Secondary controls, and
   c. Internal audit function.

4. Evaluating the reliability of individual prescribed control procedures, given that they are performed by persons with appropriate segregation of duties.

5. Evaluating whether the control system is designed to provide adequate segregation of duties.

6. Estimating the probability of material errors or irregularities occurring and escaping timely detection, based on the design of the control system.

7. Testing compliance with control procedures to be relied upon.

8. Revising the estimate of the probability of material errors or irregularities, based upon compliance testing results.
Control risk may result either from a poorly designed system or from lack of compliance with a properly designed system. As required by the second standard of field work, the auditor's evaluation of the amount of reliance to be placed on the control system ordinarily will be based on both the design of the system and the results of tests of compliance with those controls on which the auditor intends to place reliance. Because there are inherent limitations on the effectiveness of any system of internal accounting control, complete reliance on internal accounting control to the exclusion of other auditing procedures is not contemplated by generally accepted auditing standards (see AU 320.71).

We obtain our understanding of the flow of transactions through the accounting system by making careful inquiry of knowledgeable client personnel and by referring to appropriate documents and records, such as system flow charts and descriptions, procedural manuals or other relevant sources of information. This understanding provides us with a general knowledge of the various classes of transactions; the way in which each is authorized, executed, initially recorded, and subsequently processed; the methods of data processing used; significant records and reports; and the control environment. Our understanding of the control environment provides us with a general knowledge of the methods used by the entity to communicate responsibility and authority; of management's supervision of the system, including the existence of an internal audit function, if any; and of the competence of the personnel.
We may record the information concerning the flow of transactions through the accounting system in the form of answers to a questionnaire, narrative memoranda, flow charts or any other form that best suits our needs in the particular circumstance. Alternatively, the client may prepare the documentation of the flow of transactions for us to review. Regardless of whether we or the client prepares the documentation, we often find it enhances our understanding of the system to trace a few of each kind of transaction through the accounting system. If we decide to extend our study and evaluation of internal accounting control with a view to restricting the extent of our substantive tests, we obtain further information about the primary and secondary control procedures that are prescribed and by whom they are performed.

The auditor's estimate of control risk (CR) is based on the evaluation of the effectiveness of the system of internal accounting control. For purposes of considering the design of systems of internal accounting control, the Firm has identified the basic possibilities for the perpetration and concealment of misstatements and the control procedures that should prevent or detect such misstatements. These possibilities and procedures are reflected in our decision tables and related internal accounting control questionnaires and are referred to as either critical combinations of conditions or single conditions.

The decision tables identify critical combinations of conditions in which a person has uncontrolled access to assets and also has access to records in which the loss of assets could be concealed.
The basic possibilities depicted in the decision tables relate primarily to irregularities, which are intentional and whose underlying causes ordinarily are a lack of integrity and a motivation for personal gain. When a critical combination of conditions is present, the auditor should normally assume a high level of control risk in the absence of some other control.

*Single conditions* identify functions in a system of internal accounting control where errors may occur and not be corrected because some of the necessary functions are absent, but where the individuals involved do not have uncontrolled access to assets. When single conditions are present, the auditor should assess the likelihood of misstatements occurring in amounts that would be material in relation to the financial statements being audited. If single conditions are present but the auditor believes that the likelihood of material error is remote, assuming a low level of control risk is appropriate.

Control risk assigned to the primary internal accounting controls may be reduced because of administrative controls or other management functions that are designed primarily to achieve broader management objectives, but that may also contribute to the achievement of specific control objectives. Such controls, referred to as secondary controls, generally are applied after transactions have been processed and are applied on a basis that is less detailed than primary controls. Accordingly, a secondary control generally cannot be relied upon to the same extent as a primary control and seldom will fully mitigate the absence of a primary control. A secondary
control that fully mitigates a weakness in a primary control could probably have been identified originally as a primary control. Secondary controls that do provide a basis for reliance should be carefully evaluated as to their effectiveness in each particular situation to determine whether and to what extent they do, in fact, mitigate the particular control weakness identified. Controls on which reliance is to be placed should be tested for compliance.

Because of the inherent limitations in internal accounting control, we do not place complete reliance on internal accounting control. As a matter of Firm policy we have assigned a reliability factor of 2.0 (14% risk) to control risk [therefore, substantive procedures are performed using an R=1.0 ] for those situations where we place maximum reliance on primary internal accounting controls. When critical combinations of conditions are present, or when single conditions are present and the likelihood of material misstatements is not remote, we normally place no reliance on internal accounting controls and thereby assign 100% to control risk.

Secondary controls may partially or fully mitigate the lack of primary controls. Generally, secondary controls will not fully mitigate the lack of a primary control but each specific situation in which secondary controls exist should be subjectively assessed by the auditor. When the lack of a primary control is only partially mitigated, we normally assign a maximum reliability factor of 1.0 (control risk of 37%). However, when the particular secondary control fully mitigates the lack of a primary control, a maximum reliance of R=2.0 (risk of 14%) for internal accounting control may be appropriate.
Control risk also may be reduced because of an effective program of internal auditing that functions as a higher level of control. To justify attributing a higher level of control to an internal audit function, however, we evaluate whether such function is effective and suitable for our purpose. We satisfy ourselves that (1) the scope of the internal audit work was relevant to our objectives, (2) the work was properly planned, supervised and reviewed, (3) sufficient competent audit evidence was obtained, and (4) the reported conclusions are consistent with the results of the work performed. When internal auditors are functioning as a higher level of control and we have complied with the the Firm's established requirements for evaluating their work, we believe it is appropriate to reduce the risk level assigned to control risk. The extent of such reduction is a function of the amount of audit work performed by the internal auditors and the basic level of control risk before considering their work.

In addition to the above two situations, control risk will also be modified based on the results of our compliance tests. When those compliance tests relate to inquiry and observation, the auditor makes a judgment as to whether the level of compliance is adequate to justify the planned reliance on the control procedure. When compliance testing involves examination of documentary evidence (a sampling application), the Firm has established a 95% reliability level and a 4.5% upper precision limit for maximum reliance on internal accounting control. The relatively high reliability level for compliance testing was established because compliance tests are
the primary source of evidence concerning whether the controls with documentary evidence are functioning properly. Normally, not all compliance deviations will result in substantive errors. Therefore, we established the upper precision limit for compliance deviations based on our judgment as to the level of compliance deviations that might exist before a material amount of misstatement could exist in the financial statements.

In evaluating compliance tests, the achieved upper limit on deviations may exceed the 4.5% limit, either because the initial sample size was too small or because more deviations were found than were anticipated in designing the test. In such cases, we determine an appropriate adjusted level for control risk and the corresponding adjusted reliability factor needed for substantive testing.

**Substantive Test Risk**

Substantive test risk is the risk that the auditor's substantive procedures will fail to detect a material misstatement. Section 320.73 of SAS No. 1 states that "the auditor's reliance on substantive tests may be derived from tests of details, from analytical review procedures, or from any combination of both that he considers appropriate in the circumstances." However, we believe that, because these procedures complement each other, neither one should be omitted. The maximum allowable level of substantive test risk (for the combination of tests of details and analytical review) can be calculated from the model, given the values assigned for the other risks.
Analytical Review Risk

The process of evaluating analytical review risk is illustrated in Exhibit B.

Exhibit B
THE PROCESS OF EVALUATING ANALYTICAL REVIEW RISK

1. Identifying relevant independent variable.

2. Selecting the general nature of the model of the relationship among variables.

3. Determining the specific values of model parameters (for the base period).

4. Processing the input variables through the model to estimate the expected value of the relevant financial statement variable (for the projection period).

5. Evaluating whether fluctuations from the expected (predicted) value are "unusual" and need to be investigated. This involves estimating the probability of failing to investigate nonrandom material fluctuations that, in the aggregate, are material.

6. Deciding whether investigated fluctuations are attributable to identified specific economic events.

Analytical review utilizes deductive reasoning (i.e., conclusions about the details are inferred from the evidence obtained concerning the total population). The overall reasonableness of the population is tested by identifying observations that are significantly different from the expected amounts and substantiating their validity by investigation. Analytical review risk can be reduced by (1) developing a more precise expectation of the current results, (2) identifying fluctuations on a more extensive basis, or (3) performing more extensive corroboration of explanations of identified fluctuations.
Whether or not analytical review is used as a source of substantive test reliance in our audit risk model, we perform analytical review procedures in all general audits (1) as a preliminary orientation procedure in audit planning to assist us in obtaining an understanding of the client's business and in tentatively determining the nature, timing, and extent of our auditing procedures, and (2) as a review of the overall reasonableness of the financial information on which we are reporting. However, we do not place substantive test reliance on the analytical review procedures envisioned in such a preliminary orientation or final review and, therefore, do not explicitly incorporate them in our audit risk model.

For analytical review procedures to be considered effective and, therefore, relied on as a substantive test in our audit risk model, we believe that the following criteria ordinarily should be met:

1. The base used for comparisons with current financial information should be independent, plausible, and relevant.

2. The identification of fluctuations for investigation should be related to the materiality limits specified for the engagement and to the direction of test for the particular item.

3. Quantified explanations for identified fluctuations should account for a reasonable amount of the fluctuation and should be corroborated by sufficient competent evidential matter.

The maximum amount of substantive test reliance that may be assigned to analytical review procedures should depend on the degree of objectivity of the techniques used to measure or estimate analytical review risk. One objective technique that we have developed
for performing a substantive analytical review is the DH&S STAR Program. STAR, an acronym for Statistical Techniques for Analytical Review, is a computer program for applying such techniques. In using STAR for analytical review, our objectives and judgments about materiality and relative risks are expressed through the same parameters that we use for that purpose in applying our Audit Sampling Plan for tests of details. This feature of STAR aids us in combining the results from the two types of substantive tests and in evaluating them on a more objectively quantified basis.

The primary statistical technique used in the STAR Program is regression analysis. The Program incorporates a combination of a general regression model and an interface with our audit risk model. The general regression model develops projections designed to provide a reasonable base from which fluctuations should be measured. The interface with our audit risk model is designed to determine the amount of fluctuation that would be considered unusual and therefore would need to be investigated.

The regression analysis determines the mathematical relationship, if any, between one dependent variable of audit interest and one or more independent variables, based on a set of values of such variables. In an application of the STAR Program, for example, the dependent variable might be cost of sales, the independent variable might be sales, and the data profile might include the recorded amounts for each of those variables for each of the 36 months preceding the current audit year (base period) and the 12 months in that year (projection period).
Based on its calculation of the standard error of the regression estimate and on the statistical assurance desired by the auditor, the STAR Program determines the portion, if any, of the recorded amounts in the projection period that should be considered as unusual fluctuations and prints such portion as an excess to be investigated.

An application of the STAR Program involves three phases: (1) specifying the audit model, (2) running the Program and reviewing the printout, and (3) investigating unusual fluctuations. Specifying the audit model consists of making the required auditing decisions concerning the variables, the data profile, the direction of tests, the precision limit, and the reliability level.

The STAR Program determines which independent variables have a statistically significant correlation with the dependent variable, computes a regression function, and develops regression estimates of the dependent variable in the projection period. The Program provides a printout for the projection period that shows (for the dependent variable) recorded amounts, regression estimates, residuals, and unusual fluctuations.

Fluctuations identified as "excesses to be investigated" represent the difference, in the specified direction of test, between the recorded amount and the amount considered acceptable without investigation, the latter being based generally on the standard error of the regression estimates and the specified precision and reliability. Such excess could result from (1) identifiable unusual transactions or events, (2) unidentifiable random fluctuations, or (3) accounting errors or irregularities.
The excesses are investigated for possible misstatement of the recorded amount by either (1) obtaining a satisfactory analytical explanation of the excess or (2) testing additional details of the pertinent transactions or balances. If the amount of the excess is substantial in relation to the recorded amount, the analytical approach normally is more efficient and, therefore, preferable. However, if the excess is relatively small, a satisfactory analytical explanation is less likely to be available and additional testing of details may be the preferable approach.

As a matter of Firm policy, we do not place all of our substantive audit reliance upon analytical review procedures; therefore, we perform some substantive tests of details regardless of the reliance assigned to analytical review procedures in our audit risk model. Consequently, in our use of the STAR Program, when we either (1) identify no excesses to be investigated or (2) obtain satisfactory analytical explanations of all identified excesses, we still perform a certain level of substantive tests of details. When no satisfactory analytical explanation has been obtained for an excess to be investigated, we extend our substantive tests of details of the dependent variable to compensate for the decreased reliance on analytical review procedures. Therefore, for each excess to be investigated, the STAR Program (1) computes the increase in analytical review risk that would occur if the excess is not satisfactorily explained; (2) determines the required increase in reliance on tests of details that would be necessary, according to the audit risk model, to compensate for the decreased reliance on analytical re-
view; and (3) prints out information that is used directly in the Firm's Audit Sampling Plan to select the additional sample for the required extension of our substantive tests of details. Because the excess may have been caused by errors in any recording function for any accounting record affecting the dependent variable, each pertinent function and record is comprehended in the tests of additional details. Such functions include the preparation of initial records, summarization of records, and posting of the final records that affect the dependent variable. For a further discussion of our use of the STAR program, see Stringer (1975); and for a conceptual explanation of the integration of regression analysis and audit sampling, see Kinney (1979).

Tests of Details Risk

The process of evaluating tests of details risk is shown in Exhibit C.

EXHIBIT C

THE PROCESS OF EVALUATING TESTS OF DETAILS RISK

1. Selecting and evaluating the validity of procedures for substantive tests of details (i.e., evaluating nonsampling error).

2. Designing and selecting sample of details for testing.

3. Applying audit procedures to sample items.

4. Estimating the probability that a material misstatement would be detected by the sample.

Tests of details utilize inductive reasoning (i.e., conclusions about the total population are inferred from the details examined). Individual items are selected from a population, auditing procedures are performed, and conclusions are drawn concerning the population, based on the results of the procedures applied to the items selected.
We control the sampling risk portion of tests of details risk through use of the Firm's Audit Sampling Plan. For a monetary sample under the Audit Sampling Plan, judgments about materiality and relative risk are expressed through the monetary precision limit (MP) and the reliability factor (R), respectively, that we use to design and evaluate a sample. These parameters (MP and R) also provide a means for us to express our judgment about two other matters that we consider in designing a sample to achieve audit efficiency and to respond to client expectations. The first of these is the level of errors, if any, that we may reasonably expect in the sample on the basis of our past experience or knowledge of current conditions relating to the records to be tested. The second is that the client may expect the extent of our tests to be related to a level of possible errors and irregularities that is lower than we would consider material in relation to the financial statements. In applying audit sampling, either manually or by computer, a simple computation using the parameters indicated above provides the basis for determining the sample size and selecting the sample items.

Audit Sampling Plan

The Audit Sampling Plan was developed between 1958 and 1962 by Kenneth W. Stringer, recently retired partner in Deloitte Haskins & Sells, with consultation from the now deceased Professor Frederick F. Stephan, an eminent professor of statistics at Princeton University. A description of the Audit Sampling Plan may be found in Stringer (1963) or the DH&S programmed instruction course, "Audit Sampling" (1978). Discussions of the problems associated with classical sampling plans based upon normal distribution assumptions,
the development of the DH&S Audit Sampling Plan, and the current state-of-the-art of statistical sampling in auditing are available in Stringer (1979). The Audit Sampling Plan was designed to solve some of the problems associated with the use of classical statistical methods based on normal distribution theory. These classical methods were not well-suited to audit sampling because they provided little or no guidance for typical situations of concern to auditors -- sampling to detect the presence of accounting misstatements (errors or irregularities) where the rate of occurrence may be low. Stringer (1979) described some of the major features of the Audit Sampling Plan and its implementation as follows:

The most significant feature of this plan is that it provides carefully defined and mathematically rigorous reliability levels and precision limits with respect to monetary errors from samples that reveal zero or any other number of such errors. This capability was achieved by the development of a new sample evaluation methodology that does not depend on normal distribution theory.

The plan as initially implemented also provided for numerical (attribute) samples, multiple stratification for monetary samples, a variety of selection techniques and other features designed to facilitate its application in auditing. Although important for their respective purposes, these features were clearly secondary to the new evaluation methodology. Subsequently, the selection techniques incorporated in this plan were expanded to provide for selection proportional to size (except for "top-stratum" items that are always selected) by use of a technique based on "cumulative monetary amounts" in the population. This technique became known by its acronym and is often referred to as "CMA sampling." However, such references focus on a secondary feature rather than on the most significant feature of the plan as noted above. The Audit Sampling Plan is used extensively on substantially all audits in the author's firm.

In 1967 this plan (with a more sophisticated version of the CMA selection technique) was made available to other accounting firms, business organizations, governmental agencies, and universities as part of the firm's Auditape System, the first generalized audit software system developed by a CPA firm. (p.117)
The Audit Sampling Plan is based on the Poisson distribution, which gives the probability that a sample of a specified size will contain at least a certain number of errors, given a specified population error rate. When we adapt the Poisson distribution to monetary sampling in auditing, we can determine the probability of finding at least one error in the audit sample if the audit population contains errors equal to MP.

**Integrating Risk Components**

The following table illustrates the integration of risk components in our audit risk model for three different assumed situations. The assigned individual risks are expressed in percentages and the reliability factor that corresponds to each is also given. In each case, inherent risk (IR) has been conservatively assumed to be 100%, and the ultimate risk has been set at 5%.

<table>
<thead>
<tr>
<th>Type of Risk</th>
<th>Assigned Risk</th>
<th>Related R Factor</th>
<th>Assigned Risk</th>
<th>Related R Factor</th>
<th>Assigned Risk</th>
<th>Related R Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>IR</td>
<td>100%</td>
<td>0.0</td>
<td>100%</td>
<td>0.0</td>
<td>100%</td>
<td>0.0</td>
</tr>
<tr>
<td>CR</td>
<td>14%</td>
<td>2.0</td>
<td>100%</td>
<td>0.0</td>
<td>37%</td>
<td>1.0</td>
</tr>
<tr>
<td>TD</td>
<td>61%</td>
<td>0.5</td>
<td>100%*</td>
<td>0.0*</td>
<td>37%</td>
<td>1.0</td>
</tr>
<tr>
<td>AR</td>
<td>61%</td>
<td>0.5</td>
<td>5%</td>
<td>3.0</td>
<td>37%</td>
<td>1.0</td>
</tr>
<tr>
<td>UR</td>
<td>5%</td>
<td>3.0</td>
<td>5%</td>
<td>3.0</td>
<td>5%</td>
<td>3.0</td>
</tr>
</tbody>
</table>

*Regardless of the reliance assigned to AR in the risk model, a certain level of substantive tests of details is required.

In Example A, we assume that control risk has been specified as 14% (R = 2.0) based on our study and evaluation of internal accounting control (including compliance tests), that there are no other secondary controls or internal audit function on which reliance will be placed, and that analytical review risk (AR) was
judged to be 61% (R = 0.5). Therefore, the risk model indicates the maximum allowable tests of details risk (TD) is 61% (R = 0.5). The product of the four risks yields an ultimate risk (UR) of 5% and the sum of the reliability factors is 3.0.

In Example B, we assume that control risk (CR) has been specified as 100% (either because controls do not exist or because the cost of compliance testing exceeds the benefits), that the STAR Program was used for analytical review with an R = 3.0 (AR = 5%) specified for the application, and that either (1) no unusual fluctuations were identified for investigation or (2) all unusual fluctuations were satisfactorily explained. Therefore, no explicit reliance is placed on tests of details in the risk model (i.e., TD = 100%). Nonetheless, we perform at least our minimum required level of substantive tests of details. As before, the sum of the four reliability factors is 3.0 and the product of the four risks is 5%.

In Example C, we assume that control risk was set at 37% (R = 1.0) (based on secondary controls judged to partially mitigate the lack of primary controls) and that analytical review risk (AR) is judged to be 37% (R = 1.0). Therefore, the risk model indicates the maximum allowable tests of details risk (TD) is 37% (R = 1.0). Again, the product of the four risks yields an ultimate risk (UR) of 5% and the sum of the reliability factors is 3.0.

As illustrated in the examples above, the auditor—by specifying an appropriate allowable level of ultimate risk and by assigning an appropriate level of risk for each component in the risk model—will have achieved a reasonably low risk of material misstatement in the
financial statements taken as a whole. The use of the risk model for measuring and controlling the achievement of specific audit objectives, together with the application of our basic audit approach with its primary and corollary tests, will provide this desired overall assurance.

The model also can be related to individual accounts or specific objectives. For example, with an objective to test revenue for understatement, we can demonstrate that the risk of a material understatement has been reduced to an appropriate level in various ways: (1) by evaluating the internal accounting control system to determine that appropriate procedures have been designed, and testing compliance with such system to ensure that all shipments are billed, at appropriate prices, and are properly summarized in the financial statements; (2) by testing details for potential understatement of the initial recording and summarization of revenue; and (3) by analytically reviewing for potential understatement of revenue. The auditor chooses, based on effectiveness and efficiency, the amount of reliance to be obtained from each component so that the ultimate risk will be controlled at the appropriate level.

After completion of our auditing procedures, if we have made reasonable or conservative estimates for the risk associated with each component of the model, we will have a lower ultimate risk than the specified 5% because of the performance of analytical review procedures or tests of details procedures that have not been factored into the model and because the actual level of inherent risk may be less than our conservative assignment of 100% to inherent risk.
DISTINCTION BETWEEN ORDINARY AND SPECIAL AUDIT RISK

From a conceptual viewpoint, all aspects of audit risk can be associated with one of the four risk model components. However, in a practical sense, we attempt to evaluate and control the risk components by establishing certain operational procedures (such as those discussed above) that are designed to be suitable for the measurement or estimation of the risk components in the typical or ordinary audit situation. In doing this, we find that there are special audit risk situations that are not susceptible to the ordinary measurement or estimation techniques, but that need to be considered in assessing and controlling ultimate audit risk. Therefore, in planning and conducting an audit, we find it helpful to identify special audit risk situations that may influence ultimate audit risk, but that are not incorporated into our ordinary audit risk measurement procedures.

Special audit risks are those attributable to (1) conditions that increase beyond a normal level the risk associated with the inherent limitations of internal accounting control (particularly those relating to collusion, forgery, and management fraud or override of existing control procedures) or (2) conditions that arise from the general business or management environment or from particular accounting or auditing problems that are not directly related to internal accounting control.

The major element in our identification and evaluation of special audit risks is our professional judgment, based upon an under-
standing of the client's business and any related accounting or auditing issues. In exercising our professional judgment in this area, we find it helpful to give special consideration to four areas of special audit risks: business environment, management environment, audit environment, and special accounting matters. For convenience, we treat business risk (described above) as a type of special audit risk because it involves increasing our audit effort over that required by generally accepted auditing standards as a matter of prudent business judgment.

It is not feasible to establish definitive procedures to be followed in response to identified special audit risks, because the appropriate procedures necessarily will depend on the circumstances in particular cases. Appropriate responses may include consultation with specialists; changes in the nature, timing, or extent of auditing procedures; qualification of our report; or other actions.

PARAMETER ESTIMATION ERROR

Auditor error in estimating the model parameters can result in errors in estimating ultimate risk. Conceptually, ultimate risk (UR) is a product of inherent risk (IR), control risk (CR), analytical review risk (AR), and tests of details risk (TD). If we correctly estimate the model parameters (IR, CR, AR, and TD), then we should be able to estimate UR correctly, provided the model components are independent—an issue discussed in the next section. However, in practical audit situations, the actual UR (or the actual value of any of the model components) may be regarded as a function of both (1) our estimates of the parameters and (2) any errors in parameter assessment.
Since the auditor is interested in preventing actual UR from exceeding an allowable level, it follows that he wishes to avoid underestimating actual UR (or the actual values of any of the model components). This can be accomplished in one of two ways. First, objective techniques could be utilized, where feasible, to eliminate or minimize the probability of parameter estimation errors (and the magnitude of such errors whenever they may occur). Second, in situations where such objective techniques are not feasible (or cost-effective), the parameter value used in the risk model could represent a conservative estimate that is a function of both the auditor's subjective estimate of the risk component and the probability of parameter underestimation error.

For example, based on these considerations, the values assigned in the DH&S risk model to CR, and to AR when the firm's STAR program is not used, are judged to be conservative risk estimates that allow for both (1) the most likely estimate of CR and AR and (2) the probability of parameter underestimation error. However, when the STAR program is used to estimate AR, or when the firm's Audit Sampling Plan is used to estimate TD, the probability of failing to detect a material misstatement is objectively determined and the probability of parameter estimation error is judged to be effectively controlled.

Several research studies have utilized simulation techniques to investigate the probability and magnitude of parameter estimation error associated with various statistical techniques for measuring TD and AR. See Neter and Loebbecke (1975) for examples concerning
TD and Kinney (1982) for an example concerning AR. Any significant parameter estimation errors associated with particular statistical techniques should also be factored into the process of assigning values to the components utilized in an audit risk model.

POTENTIAL INTERDEPENDENCE IN MEASURING COMPONENTS

Although the components of the risk model are conceptually distinct and independent, certain considerations may make it difficult for the auditor, in a practical situation, to assess them independently. Consequently, special attention needs to be devoted to minimizing or at least controlling such interdependence, since failure to do so would result in "double counting" the effect of factors that may be considered in more than one component. This section discusses the potential interdependence between various components and possible mechanisms for minimizing or controlling the adverse effects of such interdependence.

We believe that the most serious problems of interdependence are related to the assessment of inherent risk. The interdependence between inherent risk and control risk was discussed earlier in the paper. The probability of initial occurrence of a material misstatement (inherent risk) is likely to be a function of the probability that such a material misstatement would be detected (control risk). For this reason, it does not seem feasible, in most practical audit situations, to assess inherent risk separately from control risk. Rather, if we desire to recognize the different levels of inherent risk in different audit situations, we likely
would find it more practical to estimate the probability that material errors would exist after being processed through the internal accounting control system, which would represent the joint product of IR and CR.

The evaluation of inherent risk may also be interdependent with the evaluation of analytical review risk. Many of the same factors that are considered in the evaluation of IR (such as those related to the auditor's understanding of the client's business) are also considered in the evaluation of AR. However, this is not actually a problem of the basic risk model, conceptually, nor is it an indication of interdependence of the basic risk-model components. Rather, it illustrates that certain factors (such as a lack of understanding of the client's business) that may cause errors in estimating one component may also cause errors in estimating another component. The appropriate response to such a potential problem is to attempt to control such factors in an effort to reduce parameter estimation error and, where such parameter estimation error exists, to make an adequate allowance for it in the assignment of a value for the component in the risk model.

In some cases the evaluation of control risk may be interdependent with the evaluation of analytical review risk. The identification of unusual fluctuations in an analytical review requires a reasonable data base for the independent variable(s). If control risk is high with respect to an independent variable used in an analytical review, the probability of a reasonable data base for the analytical review may be decreased. On the other hand, when CR
is high, the auditor may gain satisfaction regarding the reasonableness of the independent variable in an analytical review by using other means, such as auditing the data representing the independent variable. An appropriate audit response to the potential interdependence between CR and AR is to address the question of data reasonableness directly and to require a certain minimal degree of reasonableness of the data base for the independent variable(s) in order to place reliance on the results of an analytical review.

In some situations, the evaluation of TD could also be dependent upon CR if internally generated documents are utilized for substantiation of the details (transactions or balances) being tested. For example, if sales are tested for understatement by sampling from internal shipping documents and tracing such shipments to the sales records, test of details risk will be increased if control risk related to such internal shipping documents is high. The appropriate audit response is to recognize the potential interdependence between CR and TD in such situations and either (1) to require a certain minimal level of reliability of internally generated documents used to test details (through adequate internal control or other independent tests of such documents) or (2) to make an appropriate measured allowance for the decreased reliability of such documents in assigning a value to TD in the risk model.
Combining Single Probability Estimates vs. Combining Probability Distributions--One of the limitations of current practical applications of the risk model is that errors or irregularities associated with each of the model components form a probability distribution, whereas current applications of the model utilize only a single estimate of the probability that such errors or irregularities will exceed a certain (material) amount. From a conceptual viewpoint, combining the entire probability distributions associated with model components into an overall probability distribution of undetected errors and irregularities (after completion of the audit) would provide a more appropriate measure of UR than combining only the probabilities of material misstatement from the individual components (a simplified binary representation of the complete probability distribution), as in current applications of the model.

SOME SUGGESTIONS FOR FUTURE RESEARCH

There is an abundance of research opportunities concerning the development and application of audit risk models. Readers have likely identified many future research needs related to the problems and limitations of current applications discussed in this paper. The following comments are not intended to provide an exhaustive or comprehensive listing of research opportunities concerning audit risk models, but rather to suggest a few areas in which research would be especially promising.
Some of the objectives of such future research include (1) measurement, control, and reduction of parameter-assessment error; (2) control and reduction of the adverse effects of interdependence of the risk component estimates; and (3) the utilization of reasonably representative empirical measurements or estimates of the probability distribution of misstatements associated with each of the risk components and with ultimate risk. With respect to the third objective, when a particular set of conditions has been identified in an audit, the auditor could find it very helpful to know the probability distribution of undetected errors and irregularities that, according to reliable empirical studies, are associated with that set of conditions.

It seems likely that future research concerning the measurement and control of inherent risk cannot effectively be isolated from control risk considerations. Consequently, research concerning the probability of errors and irregularities occurring and not being detected by the internal control procedures in different identifiable inherent risk situations could provide the basis for developing helpful decision aids for the auditor.

Research is needed concerning the reliability of various analytical review techniques. The reliability of various human information processing techniques could be compared with that of regression analysis and other modeling techniques to assess their relative effectiveness and efficiency in various audit situations.
The effectiveness and efficiency of various analytical review investigation rules has been addressed in one study by Kinney (1982), and the effectiveness of some human information processing in analytical review has been addressed by Kinney and Uecker (1982) and by Mock (1982).

Research and development is needed concerning the potential quantification of nonsampling risk in the assessment of TD. Some research has been conducted with respect to the validity of different methods of confirming receivables. Considerable additional research is needed to develop techniques for measuring the validity of various auditing procedures for tests of details, for assessing the resultant nonsampling risk, and for measuring and controlling the combination of nonsampling and sampling risk.

Finally, research attention needs to be directed to the development of practical techniques and operational audit software packages for combining the probability distributions for undetected errors and irregularities associated with various risk components and arriving at an empirically justifiable estimate of ultimate audit risk.

CONCLUSION

The auditor may be able to make better and more consistent judgments when risks are objectively quantified. For example, the use of statistical sampling to make tests of details (of transactions or balances) provides certain advantages because it provides a practical means for (1) expressing in quantitative terms judgments about the audit objectives, (2) measuring the degree of assurance that results from audit tests, and (3) evaluating whether the
objectives of those tests have been achieved. However, we recognize that the auditor must assess much information that is subjective and that quantifying all audit risks in an objective manner may involve costs in excess of the benefits. Consequently, a crucial element in the practical application of the risk model is the use of auditor judgment, especially in assigning values to risk-model components in situations where such components are not susceptible to measurement or where such measurement is not cost-effective.

In the past, the auditor considered all of the available evidence and intuitively combined it to reach an overall audit conclusion. Doing so involved use of some sort of implicit underlying model, however ill-structured and incoherent it might have been. Even if the model was not explicit, it did exist and its general structure was apparent. For example, when the auditor had an analytical review tool that provided reliable evidence, the extent of substantive tests of details was reduced. Similarly, when internal accounting control was nonexistent, substantive testing was increased. Even an intuitive model as simplistic and incomplete as in these examples was useful because it provided a framework to assist the auditor in assessing audit risk.

We believe that the use of an explicit, fully-articulated audit risk model that integrates all aspects of the audit process provides a more effective and efficient approach to the measurement and control of audit risk. This is consistent with recent research on
auditors' judgments, which indicates that such judgments could be improved through the use of models to aid the auditor in combining and integrating audit data. (For a review and synthesis of this research, see Ashton, 1982, and Libby, 1981.) The purpose of this paper has been to describe efforts to develop and introduce such a model into our audit practice and to integrate it with our existing audit framework and statistical procedures.
REFERENCES

American Institute of Certified Public Accountants (1980), Statement on Auditing Standards No. 31, Evidential Matter.

(1981), Statement on Auditing Standards No. 39, Audit Sampling.


A Discussion of Audit Risk Model: A Framework for Current Practice and Future Research

Lynford E. Graham
Coopers & Lybrand
Discussant's Comments by Lynford E. Graham
Partner, Coopers & Lybrand

Audit Risk Model: A Framework for Current Practice and Future Research

Introduction

I am pleased the authors have chosen to address the risk model that first appeared in our professional literature over ten years ago as Appendix B to Section 320 of SAS 1. As you are aware, this "formula" (both the early version of Section 320 and its recent refinement expressed in an appendix to SAS 39, Audit Sampling) has been criticized both by practitioners and by academic statistical experts. An open discussion and examination of this risk model's uses and limitations is necessary and long overdue, because it continues to be part of our professional literature. Furthermore, an SAS addressing the components of audit risk and our professional responsibilities toward its control is currently being studied by the ASB. In the past, discussions of the formula have been as rewarding as an open debate between traditional physicists and members of the Flat Earth Society, generating more heat than light and changing nothing.

Extraneous Material

My remarks on the paper will focus primarily on the audit risk model. Some of the sections of the paper have nothing to do with the risk model itself. For example, the policy decisions that DHAS uses to implement the model—setting a maximum 5% ultimate risk, setting a minimum 14% risk for the effectiveness of controls, always setting inherent risk at 100%, and setting a minimum 37% risk for substantive testing—are subjective judgments. Undoubtedly, these figures can be challenged. Little, if any, support for these judgments is provided in this paper. I will defend the right of any firm to set whatever policies it sees fit to help it perform what it believes is a quality service at an economic price if that firm will in turn permit me to set policy for my practice. But I would strenuously object to having these judgments adopted as standards for my practice without a great deal more research and discussion than is present in this paper.
The paper's discussion of regression analysis is related to the way DH&S applies the STAR system in practice, but this issue should also be separate from discussions of the risk model. Whether direct comparisons, ratio analysis, or regression analysis is used, the general model is not affected. I will leave it to others—qualified, independent statisticians and auditors—to discuss the problems associated with considering a time series of financial data as independent observations for a multiple regression application; the resolution capability of 36 historical data points; the accuracy and consistency of monthly data in most accounting systems; and whether the auditor has the responsibility to determine that the identified regression relationships are logical, complete, and projectable. These issues are not new, do not apply solely to the approach described in the paper, and do not affect the determination of the validity of the risk model.

Contribution

One issue I'd like to raise is whether the paper contributes to the professional literature and practice. Does it go beyond the explanation of the risk model that can be found in the appendix to SAS 39, that is expanded on in the proposed audit guide to SAS 39, and that is again embellished in the materiality and audit risk exposure draft? I'm not sure I can be impartial in that judgment since I, as chairman of the Quantitative Methods Task Force, worked with Jim Kirtland and others to achieve what I believe is a substantial clarification of the model: its usefulness, its limitations, and its judgmental nature for the SAS 39 appendix. I do think, however, that a careful reading of the risk model as it is presented in the current and proposed professional literature gives one an understanding similar to that achieved by reading this paper.

If I am wrong in this judgment and readers find this paper helpful in understanding or interpreting the audit risk model, it is worthy of study by practitioners. I think we must immediately provide the practitioner with whatever tools he needs to be able to fully and properly interpret the professional literature and requirements relating to this risk model. Past examples of abuses and misunderstandings concerning the model are evident on
paper and in practice. For example, I believe the simplistic equation of the risk factors in the model to confidence levels for compliance tests is preventable. To the extent the paper serves to communicate, whether by reinforcement or clarification, it should receive the attention of practitioners.

Alternative Models

If properly used, I believe the audit risk model presented here can be a valuable planning tool for the auditor. However, I also remain receptive to the idea that other models may ultimately provide more accurate and intuitively satisfying answers to the problem of combining subjective and objective risks. Research in the use of Bayesian statistical techniques, for example, may provide us with the next important refinement to the current model. For example, Leslie, Teitlebaum, and Anderson, in Appendix C to Dollar Unit Sampling, address one approach to using a "Bayesian" framework to analyze audit risk. Audit risk under their model is shown to differ from that calculated using the model presented in this paper. The exact patterns of risk assigned to components of the audit determine whether the computed audit risk under the two models is similar or materially different. Other researchers, including Bill Felix, John McCray, and John Neter, are currently making contributions that will undoubtedly expand our understanding of and ability to work with Bayesian formulations and alternative audit risk models. It would be wise for the profession to remain a bit flexible on this issue, recognizing that improvements in our audit technology are likely to be forthcoming.

Risk Model Sensitivity

Since some auditors may choose to use the risk model for planning and executing an audit engagement, it seems worthwhile to review the model's sensitivity to two major factors:

- The number of elements employed in the computations, and
- The subjective judgments that are assigned to each stage.
This second factor is itself sensitive to the underlying ability of a technique (e.g., sampling, regression analysis) to do its job (e.g., detect material error) and the amount of emphasis the auditor puts on that technique because of the characteristics of a particular application (e.g., sample size, sophistication of the regression model). A few numerical examples will help illustrate this sensitivity.

The more meaningful factors that can be added to the model, the more dramatic the reductions in substantive test sample sizes that are possible when the model is used to determine sample size. In the 1970s the two-factor model of Section 320 B considered internal control risk and substantive test risk. In the 1980s version, the risk factors have been increased to four. Each of the previous risks have been subdivided into two component risks, and the relationship of the risks is still multiplicative. As we further define and refine the audit risks in the model, more factors will probably be added. One that immediately comes to mind is the risk associated with those "other auditing procedures" that are alluded to in professional standards and the literature.

It can be demonstrated—even to a mathematically unsophisticated person—that adding additional factors to the model, provided they can be assessed at a risk level of anything less than 100%, will always reduce the extent of substantive detail test procedures necessary to achieve a low-risk audit. If we could string 30 factors together and justify reducing the risk associated with each factor from 100% to 90% (a small reduction, often requiring very little work), we might be able to perform an audit in which each component only carries a 10% confidence level but the overall demonstrated audit risk is less than 5%. How long it might take to achieve this model is hard to say. In ten years we increased the number of factors in the model from two to four. Have we doubled, squared, or increased by two the number of factors? I believe we have to be careful that we don't get carried away with our multiplicative risk model methodology, since there may come a point at which our simple multiplicative model fails.
The model is also sensitive to the judgments made about each risk. These judgments should be influenced by the perceived strength of the auditing procedure employed. For example, how effective is applying a regression analysis routine or a carefully designed fluctuation and ratio analysis routine in detecting material error and misstatement? It would be naive to believe that under realistic conditions either technique, even if performed by a guru in its use, will be 100% effective in, say, detecting material error. So, some allowance must be made in the judgment process for the risk that the technique can never be 100% effective. Secondly, the auditor must gauge the level of intensity or effort of the test. A simplistic linear regression analysis with few reliable data points and a lot of correlation between the independent variables has to be viewed as a weak application, although we may have rated the potential strength of regression techniques to be higher than some other techniques. A carefully structured fluctuation and ratio analysis, applied prevasively and logically, may be a strong application in which the potential power of the technique to reveal material error is fully realized. Thus, regression may not always be the superior analytical tool in an individual circumstance. These two factors (the strength of the technique and its intensity of application) must be combined when assessing each of the various risks in the model. The paper is very right in suggesting that more research is necessary if we are to be able to make consistent and logical assessments in this area.

Putting this all together, I'm not impressed with this model's ability to develop consistency in detail substantive test sample sizes when two auditors, looking at an identical set of facts, can use it to arrive at sample sizes that differ by 50% or more. Let's take the example on page 5 of the paper, in which the facts given lead the auditor to a 5% ultimate risk. Is it possible that an auditor, overwhelmed by the beauty of a crisp fall day and a windfall profit in the stock market, could subjectively assess inherent risk, control risk, and analytical review risk each to be a mere five percentage points lower than shown in the illustration? Or, after reading about a particularly unfavorable court judgment against another firm,
might not an auditor assess control risk and analytical review risk each to be a mere five percentage points higher than the illustration? If these assessments are possible, the resultant detail test sample sizes could differ by 50%. One could of course design other examples with greater or lesser impacts on detail testing levels, but the important thing is to recognize the extreme sensitivity of the risk model to its individual factors and to remember the subjectivity of the elements being multiplied in it. As a guide, the model can put the auditor in the right "ballpark," but it probably should not be used as a formula.

Some Suggestions

Because I feel the paper will have continuing interest, I would like to offer some specific suggestions for improving it and making it more beneficial to academics and practitioners.

1. We should consider whether a definitional restatement of inherent risk and control risk might possibly allow an auditor to assess them independently. As inherent risk is analyzed more fully, its distinct identity may be more fully apparent. Again, this is an area for future research, but our internal research to date leads me to believe independent assessment of these risks is possible. For example, the auditor may assess the risk of management override of existing controls by considering factors that are likely to indicate such problems. Deteriorating company financial position, a record of questionable management integrity, and a concentration of responsibility and control within a few key management people may indicate that additional caution by the auditor is warranted. If the problem of management override is considered separately from the internal control system evaluated by the auditor, its risk may have an additive relationship in the basic model, not a multiplicative one.
2. On page 29 the authors state: "The model can be related to individual accounts or specific objectives." Perhaps the statement should read instead that the model should be related to individual accounts or specific objectives. Overall assessments of inherent or control risk are unlikely to be meaningful in applying the model.

3. Because each segment of the audit is being performed using a 5% ultimate risk standard, should we discuss the practical problems in combining the individual account assessments to form overall conclusions? After all, auditors issue opinions on the financial statements taken as a whole. For example, how do we combine inventory and receivables conclusions under the model if an auditor does not use dollar unit sampling in which each account receives the same relative intensity of sampling effort? Do we need to add some explanatory guidance on setting risk and materiality criteria for the auditor who uses judgmental sampling, classical statistical techniques, or dollar unit sampling with unequal intensities of sampling effort within the framework of the audit risk model and approach described in this paper?

4. I disagree with the implication on page 34 that the basic risk model components are not interdependent. Only on the basis of a careful definition of the component risks can this interrelationship be minimized. However, I believe there remains a degree of interdependence between the components, and the next sentence in that paragraph appears to support this belief. I agree heartily with the suggestion on page 37 that future research should focus on the "control and reduction of the adverse effects of interdependence of the risk component estimates."

5. The characterization on page 39 of intuitive models as "ill-structured and incoherent" is not helpful to the paper, and the judgment is not supported. A thorough examination of past audit failures would be necessary in order to conclude that other risk
models in use are ill-structured; the preponderance of evidence simply doesn't support this contention. I will grant, however, that in the interface between practitioners and researchers, the ability of some practitioners to articulate their risk model and the ability of some researchers to extract that model have, to date, been limited.

6. The paper's assertion that other research shows that "judgments could be improved through the use of models to aid the auditor in combining and integrating audit data" ignores some significant recent findings. The assertion has intuitive appeal, but is contradicted by a recent major study. Auditing Research Monograph No. 3, by Mock and Turner, demonstrates that structured guidance and the use of tools such as statistical sampling may actually destabilize the audit environment and make the auditor's response to stimuli in it more variable. We must be more cautious when drawing what may appear to be intuitively obvious conclusions on this issue at this time.

Conclusions
I believe that the risk model described in this paper has long needed examination and open discussion by practitioners and academics. The risk model has found its way into our professional literature and is likely to be used by some auditors to demonstrate compliance with professional standards. Guidance on its proper application and limitations should be generally available. To the extent this paper or a revision thereof provides this guidance, wide circulation should be encouraged. However, we must make the distinction—which this paper does not now do—between the underlying conceptual model and the details of its implementation at DH&F. We have to be careful to separate our conclusions about the basic model from our rights to set individual firm policies that we believe lead to effective and efficient audits. As a tool and guide, the audit risk model can be useful to the profession; however, its use as a "formula" is limited.
A Discussion of
Audit Risk Model:
A Framework for Current Practice and Future Research

Wally Smeliauskas
University of Toronto
Discussion of "Audit Risk Model: A Framework for Current Practice and Future Research"

By Wally Smieliauskas

Introduction

In discussing the paper by Holstrum and Kirtland I will focus on issues concerning the audit risk model that I feel have not received as much attention as they deserve. Many issues have been identified in recent conferences and in prior research. Holstrum and Kirtland recognize there are major problems in implementing the audit risk model of SAS No. 39. They have described to us how Deloitte, Haskins and Sells has dealt with some of these problems. Such descriptions can give valuable guidance to academics in identifying important new research areas and perhaps guidance to other practitioners in implementing the model for their own practices.

I would like to consider this paper from two perspectives. First, I would like to consider some of the implications of this paper for audit practice and auditing theory as reflected by practice. Second, I would like to mention some implications of this model for research that have not been touched on in prior works, yet which I think are important.
Implications for Audit Practice and Theory

The audit risk model described in SAS No. 39 is suggested for use as a planning device. Holstrum and Kirtland indicate that this is the purpose for which the model is used in Deloitte, Haskins and Sells. The model is used to provide a more objective basis for determining audit risk. It is used to provide a framework to facilitate the integration of several different audit procedures. In this way it can assist in the planning and coordination of audit procedures to meet the overall audit objective of providing a specified level of assurance that the financial statements are not materially misstated. The specified level of assurance is simply 1 - audit risk.

The audit risk formula suggested in SAS No. 39 and used by Deloitte, Haskins and Sells is the following: audit risk = ultimate risk = UR = IR x CR x AR x TD = CR x AR x TD.

This formula implies that the audit process reliability can be represented by a parallel system structure as follows.
Each box or component represents different kinds of audit procedures, each related to a different risk component, that can be used in the audit. The expressions in the box indicate the probability these procedures have of detecting material errors under conditions specified in SAS No. 39. The audit process "fails" if it can get to point B without discovering a material error. The only way this can occur in a parallel system is if all routes to point B fail, i.e., if all audit procedures fail to detect a material error. Hence the audit process reliability can be improved by increasing the reliability of any single audit component (e.g. 1-AR) or more than one component. Conversely, audit reliability may be maintained by reducing the reliability of one component, but having a compensating increase in another component.

Note that the structure of the components says nothing about the independence or dependence of the components on one another. In fact, however, the definitions of AR, TD, and CR in SAS No. 39 make clear that the probabilities involved are conditional probabilities, conditional probabilities which imply a fixed sequencing of audit activities as follows: internal control evaluation, analytical review, and substantive tests of details. Thus, sequencing does not mean that all of these procedures must be performed or relied upon, however.

The parallel structure thus recognizes that there can be different sources of evidence supporting the overall assurance level and that these sources of evidence carry equal weight. That is, each audit component can contribute as much to overall assurance as the other components and, most importantly, the assurance provided by each of these components is
interchangeable with the other audit procedures. These implications all follow from the structure assumption.

Of course auditing theory and practice recognize significant qualitative differences among sources of evidence. For example, analytical review is not normally considered to be as reliable as physical examinations or confirmations, internal documents are less reliable than external documents. The audit risk model, however, ignores such qualitative differences in evidence. For this reason it is necessary to impose constraints in applying the model. For example, SAS No. 1, paragraph 71 prohibits reliance on internal controls that excludes other audit procedures. Deloitte, Haskins and Sells has placed a limit on the reliance assignable to internal controls by not letting CR have a value lower than 14% (p. 16). Similarly, another limit has been placed by the fact that substantive tests of details are apparently never entirely eliminated, although TD may be allowed to go to high levels — much higher than examples given in SAS No. 39. On the other hand, apparently some firms feel that a ceiling should be set on the value of TD used in statistical sampling. For example, some have set a ceiling of .2. In practice there thus appears to be a great deal of diversity in setting the allowable values for the TD variables of the audit risk model.

Carl Warren has noted this issue by suggesting that different weights be assigned to each risk component. This would be equivalent to adding additional modules to the parallel structure given earlier. In this way more weight could be assigned to such items as TD or perhaps CR relative to AR. It would appear from the second standard of fieldwork, and even from theoretical works such as Toba's, that CR should carry more weight relative
to AR than is recognized by the risk model. Certainly this is an area where more guidance in implementing the model would be useful.

As an example of the types of problems that may arise, I refer you to the table on p. 27 of the Holstrum and Kirtland paper. Example B illustrates a situation of virtually complete reliance on a single audit procedure, in this case analytical review. Such an approach may lead to problems if sufficiency of audit evidence is ever questioned. For example, one might be hard pressed to distinguish between a complete audit and a review of financial statement items if the risk model were used carelessly.

This brings me to the type of audit situations to which the risk model can be applied. It would appear from the discussion of at least some authoritative works on auditing, (e.g. Arens and Loekbecke, p. 260) that the primary purpose of tests of transactions is to assess the reliability of internal controls. Although some of this testing may be dual purpose involving compliance as well as substantive objectives, it seems that the primary purpose of the tests of transactions is nevertheless to assess the evaluation of internal controls with a compliance testing objective. Recognizing this means that a major benefit of the model would be to use it to significantly reduce the extent of direct tests of balances at year end by relying on extensive interim testing which evaluates the related transaction and control systems. Such an approach helps spread the total audit work load over a longer period and yet maintains the reliability of the audit. This consequence of the model may be one of the most important in terms of practical significance.

However, the Holstrum and Kirtland paper gives little indication that the model is used in this way by their firm. For example all the applications
they discuss relate to transaction streams. Their discussion of compliance testing within the context of a single control tends to imply transaction processing. Similarly, all the analytical review applications referred to either by them or in related papers tends to imply use of transaction stream data. They thus appear to view the major usefulness of the model for determining extent of substantive tests for transactions. Tests which some would view as primarily compliance testing to evaluate internal controls. Therefore, their illustrations would not appear to be as useful for spreading the total audit workload between interim and year end testing as a strategy which would try to reduce direct tests of year end balances.

This emphasis on determining the extent of substantive tests for transactions rather than for account balances is reflected by their discussion of internal control evaluation and the determination of CR. For example, when they discuss the evaluation of compliance tests they disclose a precision limit of .045 for compliance deviations that would allow complete reliance on the related internal control. They discuss this in the context of a single test for a single control which tends to imply they are evaluating the accuracy of a single transaction stream. However, the auditor frequently has to consider higher levels of aggregation in evaluating controls. For example, in evaluating controls for accounts receivable the auditor must consider the controls from several transaction streams (e.g. sales on account and cash collection) and somehow combine the results of these different tests. How does Deloitte, Haskins and Sells deal with this problem? Do they still use the .045 rule to evaluate the system error rate? How do they determine system error rate?
An objective solution to these error integration problems may be possible if Deloitte, Haskins and Sells follows a particular methodology in designing its compliance tests. If they find it is economical to test only one control relating to a particular error (e.g., using the concept of key controls) then it turns out there is a fairly simple structure assumption one can use in modeling the controls. Such a model could greatly facilitate the objective integration of the results of several compliance tests to support the evaluation of more complex or aggregated accounting systems. However, they make no mention of any integration model and thus I assume they evaluate sets of compliance tests totally subjectively or they do not find it useful to evaluate sets of compliance tests.

Other interesting issues are raised by their .045 compliance test evaluation rule. Do they assume that the .045 rule applies to all systems across all clients? Do they have empirical evidence supporting this rule? Is the rule applicable to attribute sampling based on Monetary Unit Selection as well as neutral selection of items (i.e., is the rule independent of the sampling method)? In developing the .045 rule, was it related in any way to materiality in dollar terms? For example, is some assumption made as to the relationship between compliance deviations and monetary errors in financial statements.

It should also be noted that unlike the linkage of AR and TD, the linkage between CR and the two substantive tests does not directly follow from the formula. Instead, Holstrom and Kirtland indicate that grades of internal control are used to determine the TD values. Hence the risk model reflects the relationship between CR and the substantive tests only in a general sense,
and is not used literally. This is somewhat puzzling since one can argue that analytical review can be just as subjective as relating compliance deviations to monetary errors and yet the model is used directly in relating AR and TD. Hence, if the risk model can be used to relate AR and TD then why make CR an exception? Perhaps the authors could clarify some of the unique problems they feel the firm has in assigning values to CR.

One issue which the risk model does not address is the efficiency with which the audit is performed. The focus on ultimate risk helps assure that the audit is effective in detecting material misstatements and thus is in fact the primary concern of the audit. However, efficiency is also an important consideration and although Holstrum and Kirtland have indicated that the cost of the audit procedures involved are an important factor in determining the mix of planned risk levels, I feel that without making these costs more explicit it is more likely the mix is non-optimal. To indicate the nature of the problem, I will use alpha risk as a measure of efficiency in the audit, and beta risk as a measure of effectiveness, as suggested by the authors and others. This characterization is very convenient because it allows the use of the concept of operating characteristic curve from statistical sampling theory to indicate the range of possibilities for these sampling risks. The risks in turn indirectly indicate some audit costs of following a particular audit strategy (i.e., a particular combination of planned values in the audit risk model).

An operating characteristic (OC) is simply a graph of the probability that a test will accept the null hypothesis. The OC curve can be used to concisely provide data on how sampling risks vary over the entire range of possible error conditions, including the point of exactly material error.
The audit risk model specifies the risk at the point of exactly material errors and ignores the risk elsewhere. Hence the OC curve gives a better appreciation of the costs ignored by the audit risk model.

In Figure 1 are plotted OC curves of two possible statistical tests of details. At the control point of material errors M both curves satisfy the control of beta risk objective at .05 even though the curves differ drastically over other parts of the range. In particular the $\alpha$ risks of curve A are much less than those of curve B and hence more efficient in the audit sense. However, the sample size associated with curve A is four times that of curve B, so this "efficiency" is gained through more extensive audit testing. However, these curves reflect only the TD component of audit procedures. One can only imagine what would happen to these curves with various combinations of audit procedures. Nevertheless the point is that the risk model only concerns itself with control of risk at point M, even though the actual amount of error is unlikely to exactly equal M. Hence for audit planning purposes it would appear that a more complete descriptive model recognizing changes in risks over a range of errors may be more useful.

Of course trying to assess or identify such OC curves may present a major research effort. However, given the potential, for example, of having available a less costly combination audit procedures that results in OC curve A rather than a more costly combination which results in OC curve B, it would seem that auditors should not ignore the more comprehensive approach using the entire OC curve.

Another issue ignored by the audit risk model is what some have referred to as the aggregation issue. The risk model relates only to tests
of individual account balances (e.g., accounts receivable or inventory) or class of transactions. The evaluation of internal controls by classes of transactions (SAS No. 39, paragraph 67) tends to reinforce this approach. A problem associated with this approach is determining how this piecemeal assurance relates to the assurance provided by the audit as a whole. This in turn is affected by how the auditor determines materiality: whether he subjectively sets individual materiality levels for audits of different account balances or classes of transactions or he identifies a materiality for the audit as a whole and allocates it to audits of individual account balances or classes of transactions. Apparently, Deloitte, Haskins and Sells follows the latter approach. However, auditing standards have not made explicit the kind of materiality concept, overall or piecemeal, that should be applied. Thus the relationship of overall assurance to piecemeal assurance is left unclear by auditing standards. This is certainly one aspect of the audit risk model that needs clarification. It should also be noted that the way total materiality is allocated may have a significant impact on the difference between planned risk of the model and actual risk, as indicated by the analytical review research that has been done recently.

The final implication with regard to practice and auditing theory I would like to mention is extending the usefulness of the model by using it to provide more informative disclosure in the audit report. A statement of the level of assurance planned by the auditor, and accompanied by a range of values related to that assurance may significantly benefit users and thus add to the value of the model in a social sense. This would assume, of course, that many of the issues raised here and elsewhere could be satisfactorily resolved.
Implications of the Model for Research

The major additional research implications that I see are as follows:

1. Behavioral research is needed to isolate specific forms of non-sampling and judgmental errors that can affect model accuracy.

2. Empirical research is needed to assess the kinds of OC curves that can result with different combinations of procedures. Note that this would require a system instead of a piecemeal approach to analyzing the audit process.

3. There is a need to develop more formal Bayesian methods for combining the various forms of audit evidence in a more coherent fashion. This is what Holstrum and Kirtland may have referred to when they discussed models using the entire probability distribution of errors rather than just the risk components. However, it should be noted that such more sophisticated models may not be "better" in terms of their actual OC curves. Hence it will be necessary to confirm such models via simulations.

4. Finally, I would like to suggest that there be more cooperation between academics and practitioner auditors in creating data bases which could be used to improve auditor decision models. Much of the current data is limited to that provided by one or two firms. Researchers need more diverse and representative data bases. For example data is needed from firms using different audit procedures or which model internal controls differently (such as users of monetary unit sampling). Evidence is needed for identifying the relationship between compliance deviations and monetary errors. Perhaps in some cases the data is available only from internal auditors. Nevertheless if further progress is to be made on useful audit decision models relevant data on audit environments is a key precondition.
Auditor Error Detection Performance

Edward Blocher
University of North Carolina

I appreciate Will Thomas' help in the data analysis and Will Galliart's help in data collection.
AUDITOR ERROR DETECTION PERFORMANCE

The objective of this study is to obtain some initial evidence about auditor performance in detecting errors in tests of internal control compliance. The results of the study have important implications for auditors because of the recent and continuing growth in emphasis on internal accounting controls. The concern for strengthening these controls as shown by Congress, the SEC, and the accounting profession is easily documented. The result is that both internal and independent auditors have become much more involved in evaluating and testing controls. The auditor's performance in the latter aspect—compliance testing—is the principal interest of this study.

EXPERIMENTAL OBJECTIVES AND HYPOTHESES

There are three principal elements which influence auditor error detection performance in the compliance test task (Figure 1).

(a) the auditor, his or her experience, knowledge and training,

(b) the task, which includes the presentation mode, task complexity, error mix, time pressure, and related task attributes, and

(c) audit management, the nature, timing, and extent of supervision and/or review provided to the staff auditor performing the detection task. The terms guidance and feedback will be used in this paper to refer to supervision and review, respectively, since the former terms are conventionally used in the literature.

(Figure 1)

The attributes of the auditor, per se, combine in some fashion to create a detection strategy for that auditor. Possible cues incorporated in this
strategy are error amount, error frequency, and error type, among others. That is, an auditor's detection strategy might be based upon the assumption that errors are most readily detected by looking for large amounts, by looking for certain error types, by using error frequency, or some other characteristic of errors. The error frequency cue could be used by considering the frequency of various error types in the prior audit year, in related audit engagements for different clients, or in the audit work for the current year up to the point when the auditor is applying the detection strategy.

The derivation and implementation of an auditor's detection strategy is an important research issue. The research in human information processing which has considered heuristics and biases in judgments seems to be particularly relevant here. This research is surveyed by Ashton [1982], Libby [1981], Holstrom [1980], and Hogarth and Makridakis [1981], among others. To illustrate how this research might apply, consider Tversky and Kahneman's [1973] report of an experiment which showed in part that different representations of the same problem can elicit different judgment heuristics. One result of the experiment was the finding that the frequency of a class of items is likely to be judged by the availability heuristic if individual characteristics are emphasized, whereas the representativeness heuristics is relevant if generic features are emphasized. This suggests the potential importance of presentation mode for auditors whose detection strategy is based at least in part on an error frequency cue.

A study of the other two elements, the audit task and audit management, will also suggest research questions. However, to better focus the present research study, this paper does not provide a comprehensive coverage of the literature which is relevant for each of the three elements. Instead, this study is viewed as a part of a multi-phase project to investigate auditor detection performance. This experiment reported in this paper studies only the
three factors—error frequency, guidance, and feedback. The next phase of the project will consider error amount and error type. In reviewing the literature, it is convenient to combine the discussions for guidance and feedback. Similarly, the discussion of the experimental design and hypotheses will be combined for feedback/guidance.

Feedback and Guidance

The first research objective is to investigate the effect of different forms of feedback/guidance on auditor detection performance. Feedback/guidance is in the form of information to the auditors to help them identify the errors in the case exercises. The objective of the study is not to show that feedback/guidance improves performance, since this has been demonstrated clearly in both the psychological and accounting literature.\(^1\) Instead, the study investigates relative performance effects of different forms of feedback/guidance for this task context.

The different types of feedback/guidance and the different approaches for measuring the relevant performance effects are as follows:

Four important types of feedback/guidance can be distinguished:

(a) **outcome feedback**—this is best described as knowledge of results. That is, the task participant receives a report of the outcome for each repetition of the task. An extensive survey of the theory and research for this type of feedback in the context of verbal learning and memory is provided by Annett [1969, pp. 76-104, 139-159].

(b) **partial feedback**—this is a type of outcome feedback wherein the knowledge of results is presented for some portion of the repeated trials of the task, rather than for each trial.

(c) **process feedback**—this type of feedback is derived from models of judgment. The task participant is periodically presented the parameters of the model which is obtained from a series of past responses of the participant.

(d) **task-properties guidance**—this type provides the task participant some generalized information about the nature of the task prior to
and/or during completion of the task. No specific knowledge of results is involved. Kessler and Ashton [1981] show that task-properties guidance is more effective than process feedback in a task wherein subjects predicted bond ratings given financial ratio data. In another study, Ashton [1981] finds no significant performance differences between an exact and a general form of task-properties guidance in a pricing task.

This study investigates outcome feedback and task-properties guidance only, since these appear to be most relevant for the task situation.\textsuperscript{2} Outcome feedback can be compared to a review of auditor's work by a supervisor, while task-properties guidance can be compared to a worksheet decision aid [Libby, 1981, pp. 107-112].

The objective of the study is to compare performance between subject groups, one with outcome feedback and one with task-properties guidance. Also, a third group of subjects will complete the experiment without feedback or guidance, to provide a benchmark for the analysis comparing feedback and guidance. Additionally, performance will be measured in two ways:

(a) \textit{initial performance effect} - the effect on detection performance in the case immediately following the receipt of feedback or guidance,

(b) \textit{continuing performance effect} - the effect on detection performance after two or more cases have been completed, following the initial and continued receipt of the feedback or guidance.

\textbf{Expectations for Performance Differences Among the Three Feedback/Guidance Groups}

Prior research in accounting and psychology provides a basis for expectations for the comparison of performance among groups. There should be a strong positive effect for outcome feedback on both initial performance and continuing performance. This has been a consistent finding in both the psychological and accounting research. For example, an accounting study by Mock [1973] involving a business game type task shows that outcome feedback has a significant effect on performance.
Prior research in task-properties guidance does not provide a clear expectation for the results for the study task. Harrell [1977] and Kessler and Ashton [1981] show that task-properties guidance is effective in an evaluation or pricing task, respectively. However, Mock and Turner's [1978] study in an audit context showed that guidance had little effect on auditors' decisions about sample size in substantive tests. Also, Blocher, et al. [1982] studied analytical review judgments with results that showed no effect on these judgments of a generalized guidance treatment. But, none of these studies involved a task which is very similar to the task of error detection. Thus, I offer no directional expectation for the performance of the feedback group relative to the guidance group.

For the group without feedback, there should be little or no initial or continuing performance effect, though both types of performance should be less than for the outcome feedback or task-properties groups.

The research objective for the three feedback/guidance groups can now be stated in the null form hypothesis:

H1: The initial detection performance effect and the continuing performance effect will not differ among the three feedback/guidance groups.

Error Frequency Effects

The concern which motivates the second objective of the study is the potential association between error frequency and auditor detection performance. Performance in detecting frequent errors might differ from that for less frequent errors within the same case to reflect this aspect of auditor detection strategy. This may be the case by intent or simply because of the characteristics of the auditor's memory. The research which has investigated the individual's process of coding and organizing items in memory will be helpful for predicting performance differences. Two error frequency levels are considered.
First Frequency Level: Dominant error, an error type is dominant if this error type represents about 50% of all the errors present in a test group of invoices.

Second Frequency Level: Not-dominant error, an error type which represents only about 10% or less of the errors present in the group of invoices.

Memory Research - Expectations for Performance

Differences Across Frequency Levels

Research in human memory has been active in psychology, though relatively unexplored in accounting and auditing. My approach in this discussion is to focus only on the specific aspects of this body of research which appear to be relevant for the task context and research objective. A comprehensive presentation of the state of the art of this research is not attempted, though the interested reader can find useful sources for this, such as the textbook by Gregg [1975, pp. 59-136] and the survey by Birnberg and Shields [1980].

An important concern of research in memory is how each item of information is encoded in memory. The concept of "chunking" was introduced by Miller [1956] to address the way in which items are encoded. This concept has had an important role in memory research, and it can provide a basis for predicting auditor performance in detecting dominant errors relative to non-dominant errors.

A chunk can be explained as a set of related items stored in memory. One theory which has been investigated in memory research asserts that a person stores information in chunks, or "categories of relatedness," each of which has a name or label. At the point of recall, the recall of items in a chunk is facilitated by naming the chunk at that time. If the chunk is not named, the subject might not remember anything from that chunk while if the chunk is named, the subject might remember several items from the chunk [Tulving and Pearlstone, 1966; Gregg, 1975, pp. 79-93]. Researchers in memory use the concepts of "available" and "accessible" to describe this behavior.
An item is "available" if it is stored anywhere in memory, while an item is "accessible" if the chunk name is cued at recall.

The concept of chunking can be applied to auditor detection performance in the following way. Error frequency can be considered a surrogate for a chunk. That is, frequent errors might be encoded in the auditor's memory as a chunk, a category of related errors, while infrequent errors are encoded individually. Then the discovery of a single dominant error could trigger the detection of several related errors by making that error type "accessible," resulting in better than average detection performance for the frequent errors. 3

With this in mind, the hypothesis relating error frequency and detection performance is stated in the null form:

H2: Detection performance will not differ for dominant errors versus non-dominant errors.

THE EXPERIMENTAL DESIGN

There are two experimental treatments—three feedback/guidance levels and two levels of error frequency. The feedback/guidance treatment is manipulated between subjects, while the error frequency treatment is manipulated within subjects. Each auditor completes three case exercises.

The subject's task is to list the errors they perceive in each case. Three cases are used to allow a measurement of the effect on later performance as well as the initial performance effect of feedback/guidance. The dominant error is a different type of error in each of the three cases, so that the effect of error frequency, if any, in one case does not bias performance in a subsequent case. The three error types are extension errors, classification errors, and cut-off errors, each of which is a common type of error for the experimental task in the study.
Subjects

The participants for this experiment were audit staff members of four offices for three Big Eight public accounting firms. Each office was asked to provide 30 to 40 staff auditors with both one-half to two years audit experience and some experiences with the invoice testing task. A total of 91 auditors participated. A summary of the participants is shown in Table 1. These auditors were selected by the firms, so there is a potential for bias due to non-random selection. Also there was some concern initially that the somewhat higher mean experience level for one of the firms' participants might distort the findings. However, subsequent analysis indicated no significant relationship between performance and experience for all participants, so this does not appear to be an important source of bias in the study.

(Table 1)

Since three audit firms participated in the study, it is possible to test for inter-firm differences. A partial test of this type was done by obtaining two groups of auditor's from one of the three firms. Of the four groups, each from a different office, two from different firms completed the experimental task under the outcome feedback condition. Each of the other two groups completed the task under one of remaining two feedback/guidance conditions. Part of the planned analysis of the experimental results is to compare the performance for the outcome feedback groups as a partial test of inter-firm differences. However, prior studies of auditor judgment show little evidence of inter-firm differences, so I do not expect a significant difference between these groups. For example, Asthon [1974] found about the same degree of decision variability among auditors as among firms, and Joyce [1976] reported a few significant findings, but overall no pattern of inter-firm differences.
The Experimental Cases

The experimental task consists of three cases adapted from an auditing case developed by Myers [1964, pp. 298-330]. Each case consists of 16-18 invoices and a voucher register for one month's activity. Errors were introduced into the three cases to achieve the desired mix. The cases were then reviewed by an experienced auditing instructor who subsequently prepared a solution for each case.

Also, in order to assess the extent of the perceived realism of the cases and the clarity of the instructions, the auditors were asked to evaluate the cases and their perception of their performance upon completion of the experiment. Sixty of the 91 auditors indicated that they were satisfied with their performance and had no significant criticism of the case exercises. Nine did not respond, and there was no pattern to the 17 remaining responses, except that many said they became bored with the exercise. Additionally, some auditors pointed out that the very high number of errors in the cases was unrepresentative of the typical audit situation.

Procedure

The cases were administered in a controlled setting in each firm's office. All auditors in each group received the same instructions, and were allowed the same amount of time (30 min.) per case. Only four of the participants were unable to complete the cases in the time allowed, and these cases were discarded, leaving 91 usable responses. Also, the order of the cases was randomized among participants in order to avoid the confounding effects of case sequence and the practice effect.
Each participant in the two outcome feedback groups was given the appropriate case solution after he or she finished a case. The solution noted each error in the case. Then, the participant handed in the case answers and the solution before going on to the next case. The group receiving task-properties guidance was given ten minutes to study the checklist of errors and deficiencies prior to beginning the case exercises. The checklist listed each error that occurred in the cases, though no indication was given of the expected frequency of the errors.

RESULTS

The auditors' scores are summarized in Table 2. An initial study of these results shows that the two outcome feedback groups (O1 and O2) have a similar pattern of performance. Also, the task-properties group (T) and no-feedback group (N) have similar results. The findings of a statistical analysis of the results follows.

The first step in the analysis is to compare groups O1 and O2 to determine whether there might be an inter-firm difference. Univariate t-tests were applied to each of the nine pairs of mean scores reported in Table 2, with the result that O1 and O2 mean scores were significantly different for two of the nine pairs, at p = .05. These two pairs are the mean scores for dominant errors--first case and not-dominant errors--second case. Apart from these differences, the two groups have very similar scores. Because of these differences, there is a chance that the results for feedback/guidance are confounded for inter-firm differences. This must be considered when interpreting the results of further analysis presented below.
Because of the O1-O2 differences these two groups are not pooled for subsequent analyses. Rather, only the results for group O1 are used. This group is chosen because by inspection of Table 2 it appears that the use of O1 is less likely to give significant results for the two experimental treatments. So, the results reported below can be viewed as a conservative analysis of the data.

An additional consideration for a proper analysis of the data is that the three feedback/guidance treatment levels are so different in nature that a simultaneous analysis of the three levels is inappropriate. Thus, rather than to analyze the data in a single 3x2 ANOVA, the data will be analyzed by pairwise comparisons of feedback/guidance levels in a series of three 2x2 ANOVAs. This approach is shown in Table 3 which summarizes the ANOVA results. The analyses investigate both the initial and continuing performance effects. Also, there is manipulation check which analyzes first case performance for group O1 versus group N. These scores should not be significantly different, since the O1 group received no guidance for the first case. Also the analysis of the initial effect uses second case performance for group O1, since this group received feedback for the first time after completing the first case.

ANOVA Results

The results summarized in Table 3 show that the main effects for feedback/guidance, error frequency, and the interaction effect are all strongly significant in the O1 tests for initial effect. This is due to the strong performance improvement in the second case for the O1 group in contrast to lack of performance differences in the N or T groups. The interaction effect appears to come from the more than proportional effect of outcome feedback on
The detection of dominant type errors. This group's performance in detecting dominant errors is the best for any of the group-frequency cells.

The manipulation check analysis shows a significant main effect for the frequency treatment, due to the very favorable first case performance in the 01 group for detecting dominant type errors. One could question whether there is a meaningful difference here, in part because of the strong difference between the dominant error-first case performance for the 01 and 02 groups. Also, the auditors' performance for dominant error-first case shows very high variability among groups relative to any of the other treatment level combinations. Thus, the significant effect for the manipulation check ANOVA may simply be an artifact of greater auditor performance variability among feedback/guidance groups for dominant type errors.

DISCUSSION

The interpretation of these results is subject to the well-known limitations of the research design, and to the potential confounding effect of inter-firm differences noted above. Other noteworthy limitations include the limited realism of the cases, in which were embedded a higher rate of error than the auditors would ordinarily encounter. Also, some auditors complained of boredom, which could affect third case and perhaps second case performance. The slight decline in performance second to third case, for groups N and T, could be attributed to boredom, especially since these groups were not stimulated by feedback between cases.

Another limitation is the absence of a motivational variable such as a payoff for good performance on the cases. The objective of a variable such as this would be to have the auditors attending to the cases in as nearly as
possible the same manner they approach an actual audit. This would also obviate the problem of boredom on later cases reported by some auditors. Unfortunately, the typical monetary payoff given in laboratory exercises is not practical in this case since the auditors were anonymous, the cases were administered in the accounting firms' offices, and because a small monetary reward would probably not be very meaningful for these auditors.

Interpretations

The outcome feedback treatment is effective, as expected. However, it is not easy to interpret the practical significance of the result. It would be difficult for an auditor to supply full outcome feedback in an actual audit without duplicating the audit work already performed. Perhaps one answer, and one most consistent with current practice, is a partial feedback approach wherein the reviewing auditor tests the staff auditor's work on a sample basis [Jaenicke, 1980; p. 75]. The effectiveness of this approach has not been examined experimentally, and it would appear to be a worthwhile extension of this study.

On the other hand, it appears that task-properties guidance for this task context is essentially equivalent to no feedback. This result is consistent with the studies reported by Mock and Turner [1978] and Blocher, et al. [1982]. The finding of a relative lack of effectiveness for task-properties feedback indicates that an audit supervision approach using either a suggestive checklist or generalized information might not produce the desired benefit. The issue deserves further study, however. In particular, further research could investigate the effectiveness of an approach based on strengthening task-properties feedback. For example, the feedback could be incorporated into the audit working papers by requiring the auditor to indicate by a check
mark or an initial whether a certain investigation or task was completed. This approach has been used by auditors in practice, but it has not been shown experimentally to be effective.

Another approach which could be investigated is to have the auditor perform a self-review at the completion of each significant audit step. For example, a structured working paper which elicits responses about the completion of certain tasks could be used in this way. A potential advantage of this approach is that it has some of the characteristics of outcome feedback, since the self-review might cause the auditor to redo a task and thereby detect his or her own error or omission.

The interaction effect of outcome feedback and error frequency is subject to varied interpretations. One plausible inference for this is that feedback highlighted the dominant errors; their frequency attracted attention, while the less frequent errors were so diverse the feedback was less effective. Alternatively, the quick paced, repeated nature of the task might have enhanced the effect of error frequency for the feedback group by increasing these auditors' sensitivity to the occurrence of frequent errors. Or, the effect might simply be the artifact of the apparently greater variability of auditor performance for dominant errors relative to non-dominant errors, as noted above.

Because of the significance of the interaction effect, it is inappropriate to draw strong conclusions about either main effect. However, the results of an analysis of the cell means using univariate t-tests indicates that the feedback effect is significant, apart from the influence of error frequency, whereas the main effect for error frequency is not. We cannot therefore conclude that error frequency is a significant aspect of auditor detection strategy for this task, though there is an apparent tendency for frequent errors to be more readily detected.
A final general observation from these results is that there seems to be high variability in detection performance among auditors, considering the standard deviations of mean scores reported in Table 2. This high dispersion is due in part to the small sample sizes and to the scaling of the scores since the base for the percentage performance score is a relatively small number. Yet, considering the homogeneity of education, audit experience, and training for the participating auditors, one would expect less variability in performance than observed here, apart from the influence of sample size and scaling. Perhaps the apparent variability in detection performance is a general characteristic, as is the variability in auditor decision making observed in prior studies (Mock and Turner, 1978; Ashton, 1974; Joyce, 1976). Then, an important question for further research is to investigate the potential association of selected individual and situational variables with detection performance variability, such as the two variables which were examined in this study.

Also, further research could address questions concerning the error search process of auditors. What heuristics are employed in searching for errors and deficiencies? What tends to attract an auditor's attention when scanning documents for errors? What thresholds and strategies for further investigation are implicit in the auditor's selection of a test item, or in scanning documents? These questions are important in understanding auditor performance in a variety of audit tasks—detail test procedures, working paper review, and analytical review.
FOOTNOTES


2 The study does not address the judgment biases potentially related to feedback, such as the so-called "gambler's fallacy." For a discussion see Hogarth and Makridakis [1981].

REFERENCES


Gregg, Vernon, Human Memory, Methuen and Co., Ltd. (London 1975).


Figure 1

Overview of Factors Influencing Auditor Error Detection Performance

Three Elements of Auditor Detection Performance

(1) The Auditor
- Experience
- Knowledge
- Training

Detection Strategy
Possible Cues:
- Error Amount
- Error Frequency
- Error Type

(2) The Detection Task
- Error Mix
- Presentation Mode
- Time Pressure
- Task Complexity

(3) Audit Management
- Supervision (Guidance)
- Review (Feedback)

Management Strategy

AUDITOR DETECTION PERFORMANCE
TABLE 1

Characteristics of Participating Auditors

<table>
<thead>
<tr>
<th>Experimental Group</th>
<th>Group 01</th>
<th>Group 02</th>
<th>Group T</th>
<th>Group N</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big Eight Firm</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>Number of participating auditors</td>
<td>29</td>
<td>31</td>
<td>15</td>
<td>16</td>
<td>91</td>
</tr>
<tr>
<td>Audit experience (months)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>3-22</td>
<td>6-26</td>
<td>5-54</td>
<td>8-25</td>
<td>3-54</td>
</tr>
<tr>
<td>Mean</td>
<td>12.0</td>
<td>15.9</td>
<td>22.3</td>
<td>16.0</td>
<td>15.7</td>
</tr>
<tr>
<td>Median</td>
<td>14</td>
<td>16</td>
<td>12</td>
<td>13</td>
<td>14</td>
</tr>
</tbody>
</table>
TABLE 2
Percentage of Errors Not Detected

<table>
<thead>
<tr>
<th></th>
<th>All Errors</th>
<th>ERROR FREQUENCY</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Dominant Type Errors</td>
</tr>
<tr>
<td><strong>Group 01 (outcome feedback, n = 29)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First case in sequence</td>
<td>x 52.0</td>
<td>40.8</td>
</tr>
<tr>
<td></td>
<td>s (16.2)</td>
<td>(34.6)</td>
</tr>
<tr>
<td>Second case in sequence</td>
<td>36.7</td>
<td>19.6</td>
</tr>
<tr>
<td></td>
<td>(18.9)</td>
<td>(15.8)</td>
</tr>
<tr>
<td>Third case in sequence</td>
<td>36.2</td>
<td>22.5</td>
</tr>
<tr>
<td></td>
<td>(15.7)</td>
<td>(20.9)</td>
</tr>
<tr>
<td><strong>Group 02 (outcome feedback, n = 31)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First case in sequence</td>
<td>57.3</td>
<td>69.4</td>
</tr>
<tr>
<td></td>
<td>(16.3)</td>
<td>(27.6)</td>
</tr>
<tr>
<td>Second case in sequence</td>
<td>34.7</td>
<td>21.0</td>
</tr>
<tr>
<td></td>
<td>(19.6)</td>
<td>(19.5)</td>
</tr>
<tr>
<td>Third case in sequence</td>
<td>37.8</td>
<td>30.3</td>
</tr>
<tr>
<td></td>
<td>(18.8)</td>
<td>(25.3)</td>
</tr>
<tr>
<td><strong>Group T (task-properties guidance, n = 15)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First case in sequence</td>
<td>49.9</td>
<td>42.2</td>
</tr>
<tr>
<td></td>
<td>(14.7)</td>
<td>(24.3)</td>
</tr>
<tr>
<td>Second case in sequence</td>
<td>49.5</td>
<td>40.0</td>
</tr>
<tr>
<td></td>
<td>(18.3)</td>
<td>(33.7)</td>
</tr>
<tr>
<td>Third case in sequence</td>
<td>51.4</td>
<td>52.1</td>
</tr>
<tr>
<td></td>
<td>(14.2)</td>
<td>(35.0)</td>
</tr>
<tr>
<td><strong>Group N (no feedback, n = 16)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First case in sequence</td>
<td>55.1</td>
<td>52.1</td>
</tr>
<tr>
<td></td>
<td>(14.0)</td>
<td>(32.8)</td>
</tr>
<tr>
<td>Second case in sequence</td>
<td>50.9</td>
<td>44.8</td>
</tr>
<tr>
<td></td>
<td>(16.0)</td>
<td>(30.2)</td>
</tr>
<tr>
<td>Third case in sequence</td>
<td>49.3</td>
<td>52.1</td>
</tr>
<tr>
<td></td>
<td>(11.7)</td>
<td>(25.1)</td>
</tr>
</tbody>
</table>
### TABLE 3
Summary of ANOVA Results

Pairwise Comparison of Feedback/Guidance Groups
(Group - Case in sequence)

<table>
<thead>
<tr>
<th>Initial Effect</th>
<th>Main and Interaction Effects (p &lt; .05)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Feedback/Guidance</td>
</tr>
<tr>
<td>1. 01-2d, N-1st</td>
<td>F = 20.48, p = .0001</td>
</tr>
<tr>
<td>2. 01-2d, T-1st</td>
<td>F = 11.25, p = .0017</td>
</tr>
<tr>
<td>3. N-1st, T-1st</td>
<td>NS</td>
</tr>
</tbody>
</table>

Continuing Effect

| 4. 01 (2d to 3d), T (1st to 2d) | NS                        | NS                        | NS                        |
| 5. 01 (2d to 3d), N (1st to 2d) | NS                        | NS                        | NS                        |
| 6. N (1st to 2d), T (1st to 2d) | NS                        | NS                        | NS                        |
| 7. N (1st to 3d), T (1st to 3d) | NS                        | NS                        | NS                        |

Manipulation Check

| 8. 01 (1st) vs. N (1st)         | NS                        | F = 4.92, p = .0318 | NS                        |

* "NS" means not significant for p < .05.
A Discussion of
Auditor Error Detection Performance

George W. Krull, Jr.
Alexander Grant & Company
Auditor Error Detection Performance:  
A Discussion  
George W. Krull, Jr.  
Alexander Grant & Company

Introduction

I am pleased to have received an invitation to participate in the fifth Symposium on Auditing Research and to contribute as a discussant of Dr. Blocher's paper entitled "Auditor Error Detection Performance." My comments will be presented in three parts; first, an assessment of the need for this area of research and the possible implications of Dr. Blocher's study; second, a discussion of the research design, method and analysis used; and, third, some concluding comments and remaining questions.

Research Issues and Implications

The major focus of the research, as noted in the abstract of Dr. Blocher's paper, is, "... a study of auditors' ability to detect errors in performing tests of compliance for

The views expressed are those of the author and do not necessarily represent the views of Alexander Grant & Company. I gratefully acknowledge the assistance of Dr. David A. Frisbie, University of Iowa, and my colleagues in Alexander Grant & Company.
internal control." Specifically, he focused on how well outcome feedback (review) and task-properties guidance (supervision) interact with an auditor's detection strategy to influence the auditor's detection performance. The conclusion of his study is that "the outcome feedback treatment is effective, as expected." There was a significantly greater effectiveness of outcome feedback to task-properties guidance in the detection of errors for this laboratory task. However, Dr. Blocher recognizes the inherent practical or real-world auditing limitations of outcome feedback as a review process. Thus, he suggests that a partial feedback approach may be the type of feedback/guidance treatment that could be used in improving an auditor's error detection performance. Unfortunately, the partial feedback approach was not examined in this study.

Empirical studies of the auditor's decision making processes and the exercise of judgment in realistic settings are necessary to continue the improvement in auditors' abilities to rely on internal controls and to increase audit efficiencies. The research that is in process and that is summarized in Ashton [1982] and Libby [1981], among others, should prove fruitful in improving the auditing processes if the research implications can be implemented in real-world settings. Empirical research by Ashton and Hylas [1980] was used to support specific audit guidance for one major public accounting firm. The Blocher results are too tentative and preliminary to provide the same degree of guidance or to suggest changes in audit procedures or
approaches. They do contribute to the empirical body of studies which signal possible needed changes in the audit review process and how and when superiors provide guidance and/or feedback. Thus, while the Blocher study makes a limited contribution, for reasons which I will discuss under research method and analysis, the area of human information processing as related to auditing is a fruitful area for empirical research. I encourage Dr. Blocher and others to continue their quest to help us better understand our decision-making processes. We know too little about our own judgments.

However, I believe that there are a number of concerns in the study design, method and analysis which affect the interpretation of the results and their usefullness. This next discussion addresses those issues.

Research Design, Method and Analysis

The study had two research objectives or questions. First, a comparison was done of error detection performances among three treatment groups: outcome feedback, task-properties guidance and no feedback or guidance. The influence of these independent variables was measured by initial performance effect and by continuing performance effect. Dr. Blocher noted that one would expect from prior research that "there should be a strong positive effect for outcome feedback on both initial performance and continuing performance." However, he then states on page 5, "Thus, I offer no directional expectation for the performance of
the feedback group relative to the guidance group." Unfortunately, that statement makes me question whether this particular study, or possibly the approach taken to the research question, should have been done. Task-properties guidance is the primary treatment of interest, because as Dr. Blocher states, there is plenty of evidence that outcome feedback is effective. He does not provide a theoretical rationale for why task-properties guidance should be as effective, or more or less effective, than outcome feedback. Beyond the fact that outcome feedback is effective, no case is made that the forms of feedback/guidance which were studied can be applied in practical audit situations. If they cannot be applied or it is likely that they cannot be used, then why study the relationships?

The second research objective concerned the potential association between error frequency and auditor detection performance. The concept of "chunking", as used in research in memory, and its possible basis for predicting auditor performances in detecting dominant errors relative to non-dominant errors, seems highly speculative. Possibly, error types are grouped in memory according to the severity of consequences, i.e., those errors, regardless of frequency, which have the greatest potential audit implications are grouped together and those with only minor implications are stored individually or in smaller groups. Dr. Blocher's use of the "chunking" concept seems inappropriate in an auditing context. It is the relative frequency of the errors within a certain type of error that is most important rather than
dominant versus non-dominant errors. "Most frequency errors" in one audit context are not necessarily the same as those in another audit context. Thus, there may not be a generalizable set of "most frequency errors" that can be stored in memory.

The discovery of a single dominant error would likely lead to discovery of additional errors of the same type. Related errors might be discovered, but the nature of the relationship is probably something other than "frequency of occurrence." It probably relates to the cycle or process which is being audited and the auditor's understanding of how that internal control system or the process works.

The experimental task for the second research objective consisted of three cases. The dominant error was a different type, either extension, classification or cutoff, in each of the three cases which each subject auditor completed. The randomization of the order of the cases amongst the three feedback/guidance levels mitigated potential bias.

The hypothetical cases were adapted from the Thare Manufacturing Company case appearing in Myers [1964, pp. 298-330]. Dr. Blocher did not provide a copy of at least one of the three cases as an exhibit to the paper. So, we must assume from my review of the Thare Manufacturing Company case and Dr. Blocher's prior experience in case design and simulation that the case instructions were sufficient, among other possible concerns, to adequately guide the subject auditors. Each case consisted of 16-18 invoices and a voucher register for a month's activity.
Dr. Blocher introduced errors into each case to achieve the desired mix of dominant and non-dominant errors. We need to know more about what the "desired mix" is. We need to know how dominant and non-dominant errors were distributed in each case. For example, if there are ten errors in a given case, we need to know what percentage of those errors are dominant and what percentage were non-dominant so that we know how large a difference there is between dominant and non-dominant types. This would help the reader assess the percentage of errors not detected as shown in Table 2.

Also, it is not clear to me, if there were six cases, two from each condition used, or if there were really only three cases as stated, i.e., case 1 had dominant errors and non-dominant errors or case 1 had dominant errors and another form of case 1 had non-dominant errors.

Dr. Blocher tested for inter-firm differences using the two outcome feedback treatment groups, even though he did not expect a significant difference between the groups. The test for inter-firm differences was a side issue. His assumption in the research design was that there were no inter-firm differences. Given the small or borderline sample sizes for each outcome feedback firm, it may be difficult to draw conclusions about inter-firm differences.

Notwithstanding the sample sizes, the results indicate that there are inter-firm differences which Dr. Blocher noted. However, the concern is not whether there is a difference between
the two outcome groups. The concern is whether the combination of group 01 and group 02 is different than the task-properties group and the non-directed group, initially. There was a loss of statistical power, the ability to detect a difference, by dropping group 02. Dr. Blocher could have used both groups. The 01 and 02 difference is not critical for combining the two groups. It is initial group differences between groups of different treatment that should be of most concern.

The generalizability of the results was severely hampered by the method of sample selection. For convenience, the auditors were grouped by firm, rather than randomly assigned. If the inter-firm differences were an appropriate concern, then randomization of the subject auditors across firms or a matching of subject auditors based on attributes (e.g., months of experience, formal education, highest degree completed, passed certification) should have been attempted.

Dr. Blocher states on page 9, "This group's (01) performance in detecting dominant errors is the best for any of the group - frequency cells." There is good reason for this. The 01 group was told that there is a high concentration of "some" type of error in each subsequent case. Unfortunately, we do not know what percentage of all the errors in a case is represented by the dominant one.
Concluding Comments

I will first address the realism of the experimental task as evidenced by Dr. Blocher's comments and the subjects' feedback and performance. The limitations which Dr. Blocher notes on page 12 and the questions he asks on page 15 are all serious and bring into question both the internal and external validity of the study. A more fruitful approach with much greater potential for realism would have included partial feedback in the study. The task-properties guidance group had ten minutes prior to the experiment to study a checklist of errors and deficiencies. The checklist was a mere list of occurring errors without giving the expected frequency of errors. The checklist listing errors specific to the case seems unrealistic. The list should have been "all potential errors," not limited only to those occurring in the cases. This is also not appropriate task guidance for simulating actual auditor performance. A checklist of procedures, techniques and items to be done (vis a vis an audit guide or program) which could be consulted during and especially after the exercise (in the review stage), could have a more positive effect and would have better approximated actual conditions.

Receiving the "appropriate case solution" is not the same as supervisor feedback. This lacks realism. An auditor has no "schoolbook solution;" rather, an auditor has review notes and discussions with a supervisor throughout the audit review process.
Given the specific guidance provided, the high percentage of non-detected errors is surprising. It would seem more appropriate in describing the results in Table 2 to use percentage of errors detected, since the focus is on auditor performance in detecting errors in tests of internal control compliance rather than non-detection. The results would not differ, however.

I suggest, as does Dr. Blocher (page 12), that the participants' lack of motivation greatly reduces the internal and external validity of the study. How was it determined whether the case study exercise was boring or not boring? If a straightforward question were asked, "Was this exercise boring to you?" versus an open-ended questionnaire asking participants to respond to whatever they had on their minds, more participants probably would have addressed the issue of boredom than was indicated in the results on page 9. Probably more participants became bored than so stated.

To conclude, the use of the subject auditors and their results gives me significant concerns about the discipline in which the auditors approached the experimental tasks. The lack of realism in the study hampers the usefulness of the results. Based on my experience in the data collection phase of one major empirical research study using audit seniors in a controlled laboratory setting, I have some reservations about their motivation to provide valid performances on experimental tasks. The lack of a motivational variable may have hindered the error
detection performance. There were no consequences for achieving a poor result. Yet, such consequences could have been built into the experimental design.

I appreciate the difficulties of simulating reality in experimental design. However, the thrust of my critique is that a relatively modest effort at more closely simulating the conditions of the audit and the concerns of the auditor would have greatly increased the value of this study.

REFERENCES


A Discussion of

Auditor Error Detection Performance

William P. Wright

University of Minnesota
Organization of this Discussion

Human Information Processing Research in Auditing Contexts
Design of the Experiment
Discussion of the Results
General Discussion
References

Human Information Processing Research
In Auditing Contexts

The focus of Professor Blocher's research is on auditor error detection performance. In general, to pursue this type of information processing research, one should first develop one (or more) contextual models of the task. This model of information search and processing would yield predictions that could be tested given the data obtained in the experiment.

A relatively complete cognitive error detection and interpretation model might consist of individual, task and situation components. Theory and results from psychology and behavioral decision theory could be used to postulate an organization of the auditor's knowledge and the cognitive processes used to generate a judgment in the situation (Payne, 1976). Other characteristics of the auditor such as his or her decision style (Driver and Mock, 1975) might be incorporated. Task characteristics (Payne, 1982) that might be considered include the total number of errors and the different kinds of possible errors, the framing of the error detection problem (Tversky and Kahneman, 1981), the response mode, and the type of feedback to be available to the subject if learning is to be evaluated in the study. The third major component is the situation, with possible aspects being the degree of time pressure,
the decision aids to be made available, and the extent of any intrinsic and extrinsic incentives for performance by the subjects.

Of course, a large amount of work is required to develop a model which encompasses even a small number of the parameters I have mentioned. But a reasonable amount of effort is essential to derive precise predictions and efficiently design an experiment. Without a minimum amount of theoretical specification, the statistical power of the design will be low relative to the unnecessarily general research questions being addressed; it may also be difficult to interpret the results of the study and this is the situation we have with the present study.

Design of the Experiment

While the relationship among the subjects, task and situation is discussed in the paper, the lack of a task-oriented model makes it very difficult to interpret and learn from the results. Figure 1 in the paper is a listing of possibly relevant variables however the very limited theory development in the paper fails to suggest convincingly which variables are important and why they are important.

Specific cognitive error detection strategies are not developed in the paper. The author states that "The attributes of the auditor, per se, combine in some fashion to create a detection strategy for the auditor. Possible cues incorporated in this strategy are error amount, error frequency, and error type, among others" (pp. 1-2). Some discussion is provided but a model is not developed to indicate why auditors may concentrate on particular types of errors, error amounts or error frequencies in this context.
One of the major hypotheses here is differential percentage error detection for errors occurring with different relative frequencies, regardless of the number of errors. Surely most error detection strategies would predict that types of errors that occur more frequently would be detected more frequently but why should the percentage detection rates be different for errors having relative frequencies of 30% ("dominant" errors) vs. 10% ("non-dominant" errors)? Specifically, it is not clear why relative (not absolute) error frequency, independent of error type and amount, was selected as a major aspect of this research. Again a cognitive error detection model is essential here.

The concept of chunking in memory is suggested as a possible basis for differential percentage detection of dominant vs. non-dominant errors. This memory organization concept of related items or attributes being stored as an integral unit in memory is discussed appropriately (pp. 6-7) yet applied incorrectly in the paper. Conceptually-related characteristics or attributes of a stimulus may be stored as a unit in memory and retrieved as a unit for processing, but why should the relative frequency of errors, independent of error type and other characteristics, be treated as a chunk of information in memory?

A major research objective here was to investigate the effects of outcome feedback and task-properties guidance on auditor detection performance. Recent research on the difficulty of learning from outcome feedback (Einhorn, 1980) suggests that learning from outcomes is context-dependent and surprisingly difficult. The auditor's error detection strategy will be a major determinant of the extent and speed of any learning from outcome feedback.
Discussion of the Results

The criterion variable in the study is the percentage of errors detected in the different frequency/feedback conditions. The total number of dominant and non-dominant errors in the materials is never indicated. The issue of using percentages is important for two reasons: (1) serious misinterpretations can result based on percentage detection scores (vs. the actual number of errors detected) and (2) the error scores may not conform adequately with the assumptions inherent in the ANOVA model used to analyze the data. The sensitivity of the results to the use of percentage errors detected is not reported although the author mentions the issue in the conclusion (p. 15).

The dominant error is different in each of the three sets of invoices (p. 7). The three error types mentioned here are extension, classification and cut-off errors. A data analysis assumption for this paper seems to be that differential error detection performance for the three error types is not an important issue and/or the effects of dominant and non-dominant errors and kinds of feedback are not contingent on error type. It seems likely that, independent of relative error frequency, some kinds of errors will be searched for, and detected, more frequently than others. Simultaneous consideration of error detection results for error types and error frequencies seems to be necessary here.

The expected two-factor ANOVA's (for three feedback/guidance conditions and two relative error frequency levels) for the initial and continuing effects were not provided. The explanation is that "the
three feedback/guidance levels are so different in nature that a simultaneous analysis of the three levels is inappropriate" (p. 11). Two-by-two ANOVA's are employed to analyze the data. It is not clear why this was done. Levels of a factor are not assumed to exhibit ratio, interval or even ordinal measurement properties; they are just supposed to be distinct related conditions. Linear contrasts among subsets of levels of a factor can be used to test different hypotheses. If the no feedback, outcome feedback and task-properties guidance conditions were thought to be so different that they could not be analyzed simultaneously as a factor, the design of the study becomes questionable.

The results reported in Tables 2 and 3 indicate: (1) significantly lower "percentage failure to detect errors" scores for the outcome feedback vs. no feedback and outcome feedback vs. task-properties guidance conditions on the initial measurement, (2) no significant effect of task-properties guidance in any of the treatment conditions and (3) no effect of the feedback/guidance and error frequency factors on the degree of improvement at the second measurement. However two major data interpretation considerations must be addressed. First, as discussed above, the error detection scores are analyzed as percentages. Apart from the use of percentages given the assumptions of the ANOVA model (see below), the results could simply be an artifact of the percentage measurement score, particularly for the dominant and non-dominant errors since we do not know how many errors are involved for each percentage. Second, indications of the magnitude and importance of the statistically reliable effects are not provided so we do not know, e.g., what percentage of the variance in the error scores is
accounted for by the factors (however the significant interactions will complicate this analysis).

The degree of conformity of the data with the underlying assumptions of the ANOVA model is not discussed. Also a standard ANOVA table format detailing the allocation of the sums-of-squares is not utilized. Given this incomplete presentation of the data, the reader cannot determine the extent to which the computed ANOVA results can be relied upon for inference purposes. For example, concerning the normality assumption, if the distributions of percentage detection scores are not at least symmetric and unimodal, the results can be very misleading. Also Table 2 suggests very unequal cell error variances.

General Discussion

A distinction might be made between error detection and error analysis, i.e., collection of evidence vs. interpretation and evaluation of audit evidence. Error analysis is where patterns of errors are analyzed for inferences concerning the degree of reliance upon the system. Analysis of last year's work papers would provide a basis for determination of this year's audit program. The audit program would specify a sample size and a list of possible errors. The error detection strategy is therefore built into the audit program. At this point, a junior auditor would simply go out and mechanically check for the possible errors. Perhaps the specific task in the current study is too mechanical to be a setting where one even needs cognitive error detection strategies.

Another issue here is the apparently very high number of errors in the materials. Some auditors actually mentioned that the number of
errors was unrepresentative of actual audit situations (p. 9). An unfortunately very simple explanation for the effect of the outcome feedback may be that, after discovering the large number of errors actually in the materials, the subjects dramatically increased their error expectations and improved their performance. While the limited ecological validity of the task and modest incentives for subject performance (see below), serious questions can be raised concerning the external validity of the results. The conclusion most favorable here is that the study provides only a very limited test of the hypotheses.

A result here was that task-properties guidance provided no improvement in percentage error detection rates. The guidance was a simple list of errors that apparently would appear in the materials about to be completed by the auditors. It should be noted that (1) a convincing task-oriented prediction concerning why this guidance should have an effect was not provided (e.g., did the auditors already know what errors to expect in the materials?) and (2) a list of errors is a limited form of task-properties guidance. The effects of other more comprehensive forms of task-properties guidance in audit situations should be investigated.

While it may be worthwhile to compare subjects having outcome feedback with subjects having task-properties feedback, an alternative design would have compared the performance of subjects having outcome feedback with subjects having both outcome feedback and task-properties guidance. This design would permit assessment of the marginal benefit of having task-properties guidance. Interpretation of the ANOVA results would have been simplified since there would have been a natural
progression from no feedback to outcome to outcome plus task-properties guidance. The difficulty of learning from outcome feedback (Einhorn, 1980) suggests that task-properties guidance could provide a useful decision aid.

As noted by the author, no performance-oriented incentives were provided for the subjects in the experiment. The issue of intrinsic and extrinsic incentives, and therefore an overall level of subject motivation, is important here since the task is very demanding and performance-oriented (Wright and Abouelezz, 1983). The subjects were required to list each error they detected in each of the 16-18 invoices in each of the three cases. The subjects were asked to search for as many as 11 possible errors per invoice. Thirty minutes were allowed for each case with short rest periods after each case, implying over two hours to complete the materials. The subjects reported significant boredom during the study and 4 of the 95 subjects failed to finish the materials. The impact of the results of relatively unmotivated subjects is an unanswered question (e.g., this may explain the lack of any significant results for the continuing performance measure).

The author indicated the difficulty of providing performance-based financial rewards in the experiment, however, other methods exist to provide increased extrinsic incentives such as reporting a subset of the most accurate judges within the firm, etc. While the impact of financial incentives on performance of cognitive tasks is just now being clarified (Grether, 1980, 1981--see Payne, 1982), Eger and Dickhaut (1981) and Wright and Abouelezz (1983) have shown that the presence of financial incentives can result in improved judgments.
A basic general criticism here is the lack of sufficient a priori theory development to suggest why the study was conducted. A core behavioral decision theory literature (Anderson, 1980; Einhorn & Hogarth, 1981; Kahneman, Slovic & Tversky, 1982; Payne, 1982; Slovic, Fischhoff & Lichtenstein, 1977) is now beginning to evolve. Researchers in the area of judgment/choice in accounting and auditing can now provide better theoretical process models and more justification for their hypotheses. Research in this area should therefore become more informative and cumulative in its implications. We may even expect more in terms of descriptive models of real world behavior.
References

Anderson, J. R., Cognitive Psychology and Its Implications (W. N. Freeman, 1980).


The proceedings of the past Symposia are available from:

The Center for International Education and Research in Accounting
University of Illinois at Urbana-Champaign
1206 South Sixth Street
Champaign, Illinois 61820

The Department is grateful for the financial support provided by the Arthur Andersen & Co. Foundation