Auditing:
A Journal of Practice
& Theory
Vol. 19, No. 1
Spring 2000

# The Relation between Consensus and Accuracy in Low-to-Moderate Accuracy Tasks: An Auditing Example

**Elizabeth B. Davis, S. Jane Kennedy, and Laureen A. Maines**

**SUMMARY**

This study examines the relation between consensus and accuracy using an error frequency estimation task for which auditors' overall accuracy is known to be low to moderate. We also investigate whether experience moderates the relation between consensus and accuracy for the three industries examined: manufacturing, natural resources, and banking. We find that accuracy is positively related to consensus for all auditors in manufacturing and for auditors with more than 12 (36) months of experience in natural resources (banking). For banking and natural resources, we provide evidence that auditors with little experience in these industries use a heuristic consistent with manufacturing error frequencies as an "educated guess" for the specialized industries' error frequencies. This heuristic leads to consensus among auditors, but results in low accuracy. The results are important for auditing practice and research since reliance on high consensus as a surrogate for accuracy may prove inefficient or, worse, ineffective. The results also demonstrate the need for further investigation of the determinants of audit knowledge and performance across multiple industries and tasks.

**Key Words:** Consensus, Accuracy, Error frequency estimation, Industry knowledge.

**Data Availability:** The data are the property of Alison H. Ashton. Please address requests to her at the Fuqua School of Business, Duke University.

Auditors often check their judgments and decisions by seeking advice from colleagues. For example, when evaluating a potential client, an auditor may ask another auditor for an opinion about the likelihood of material errors in the company's financial statements. If the two auditors disagree, they know that at least one of them is incorrect. However, if the two agree, they typically will presume that they are correct, even though they may be agreeing on the wrong answer. The strength of the auditors' presumption may depend on their auditing experience; two partners who agree may be more confident of their accuracy than two staff auditors who agree.

Agreement, or consensus, among auditors is an important topic because accuracy is unobservable for many auditing tasks (Wright 1987; Ashton 1985; Libby 1981). In the absence of an accuracy criterion, one of the best justifications for a course

*Elizabeth B. Davis is an Associate Professor at Baylor University, S. Jane Kennedy is an Associate Professor at the University of Washington, and Laureen A. Maines is an Associate Professor at Indiana University.*

*We thank Alison Ashton for her data and valuable input. We thank Robert Ashton, Jim Jiambalvo, and participants at the University of Washington Accounting Workshop for helpful discussions of this research. Financial support from the Hankamer School of Business, Baylor University, the KPMG Peat Marwick Foundation, and the Accounting Development Fund, University of Washington are gratefully acknowledged.*

of action is to demonstrate that others would make the same choice (Solomon and Shields 1995). Even when an accuracy criterion is ultimately available, it is known only after a decision has been made, so it is only useful for similar future decisions, whereas consensus can be an input to any decision on an *ex ante* basis. Given the potential usefulness of consensus for inferring accuracy in audit decision making, an important research issue is the extent to which consensus is related to accuracy of audit judgments.

Research to date has found that consensus is a fairly good surrogate for accuracy in accounting tasks; however, this research has examined only relatively high-accuracy tasks where consensus is a necessary (mechanical) result of high accuracy (Keasey and Watson 1989; Ashton 1985). We extend this research by examining the relation between consensus and accuracy in an auditing task where auditors are known to achieve only low-to-moderate accuracy. We use Ashton's (1991) data in which auditors provided judgments of financial statement error frequencies for several industries, including manufacturing and two specialized industries, natural resources and banking.

We also add to prior research by investigating whether experience affects the relation between consensus and accuracy and whether the effect of experience differs by industry. We expect that auditors without requisite experience in an industry may rely on their general accounting knowledge to make error frequency judgments (Hogarth 1991). Since most auditors' general accounting knowledge is based on manufacturing firms, we predict that less-experienced auditors' error frequency judgments will be consistent with error frequencies in the manufacturing industry. This heuristic leads to fairly high consensus among less-experienced auditors for all industries, but high accuracy for only the manufacturing industry. In specialized industries, such as natural resources and banking, this heuristic is less effective, resulting in consensus being unrelated or even negatively related to accuracy for auditors with low industry experience. As auditors gain experience in specialized industries, we expect that their agreement is driven less by use of the manufacturing heuristic and more by assessment of industry-specific error frequencies, leading to a positive relation between consensus and accuracy for auditors with greater industry experience. We investigate whether different amounts of experience are needed in different industries to achieve this positive relation between consensus and accuracy.

The remainder of the paper is organized as follows. The next section discusses the relation between consensus and accuracy and develops a framework for when high consensus can exist in audit judgment given low-to-moderate accuracy. The third and fourth sections describe our hypotheses and method, respectively, and the fifth section reports results. The final section discusses these results and concludes the paper.

## THE RELATION BETWEEN CONSENSUS AND ACCURACY

Consider a group of individuals who each make a series of judgments. Each individual's accuracy can be measured as the correspondence of his/her judgments with a criterion or benchmark, while an individual's consensus (agreement) with the group can be measured as the average correspondence between the individual's judgments and each of the other members' judgments (Ashton 1985).[1] To examine the relation between consensus and accuracy, we partition the group's overall consensus and their overall accuracy into two regions: (1) low-to-moderate and (2) high. The intersection of these constructs can be represented in a $2 \times 2$ matrix with four quadrants but only three possible outcomes, as illustrated in Figure 1.

The relation between high overall accuracy and consensus is easy to predict. In the extreme, only one outcome is possible: individuals who are perfectly accurate must agree (Cell 4 in

[1] Consensus, as defined in this paper, considers whether individuals come *independently* to the same conclusions. We neither examine whether individuals would "come to a consensus" if allowed to confer with each other, or whether these individuals would come to a different collective judgment as a group, nor do we examine composite judgment (e.g., the average of the individuals' judgments or a model of their judgments). These other notions of consensus have merit and have been studied, but they are not the definition of consensus investigated in this paper.

**FIGURE 1**
**2 × 2 Matrix of Accuracy and Consensus**

| | | Overall Accuracy: Low-to-Moderate | Overall Accuracy: High |
|---|---|---|---|
| Overall Consensus | Low-to-Moderate | Cell 1 Research Issue | Cell 2 Impossible Result |
| | High | Cell 3 Research Issue | Cell 4 Mechanical Result |

Figure 1). A distribution of judgments that reflects high overall accuracy/high overall consensus is provided in Panel A of Figure 2. As shown, individuals' judgments that are accurate (very close to the criterion) are necessarily also close to each other. Thus, high consensus is a mechanical result of high accuracy, and high accuracy/low-to-moderate consensus (Cell 2 in Figure 1) cannot occur.

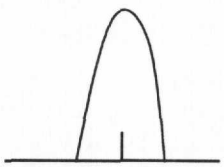
In contrast, either high or low-to-moderate overall consensus can occur when overall accuracy is low to moderate. Thus, it is the left half of the matrix in Figure 1 that merits investigation. In some cases, low-to-moderate overall accuracy is associated with low-to-moderate overall consensus (Cell 1 in Figure 1), as shown by the distribution of judgments in Panel B of Figure 2. However, low-to-moderate overall accuracy may also be associated with fairly high consensus (Cell 3 in Figure 1), as shown by the distributions in Panels C, D, and E of Figure 2.

The relation between individuals' accuracy and consensus is not easily predicted in the low-to-moderate accuracy case. Even though overall accuracy may be low to moderate, it is still possible for the relation between accuracy and consensus to be positive. For example, although overall consensus is low in Panel B of Figure 2, the individuals in the distribution who are closest to the criterion also show more overall agreement with other group members than the less-accurate individuals at either end of the distribution. A second example is shown in Panel C of Figure 2, where the peaked portion of the distribution (high-consensus portion) is closer to the criterion (higher accuracy) than judgments in the tail of the distribution. In contrast, there may be cases where the individuals who agree are not very accurate, suggesting no relation or a negative relation between consensus and accuracy. Panels D and E of Figure 2 demonstrate these situations, respectively. In Panel D, the peaked portion of the distribution is equidistant from the most and least accurate judgments, suggesting no relation between consensus and accuracy. In Panel E, the peaked portion of the distribution is furthest from the criterion suggesting a negative relation between consensus and accuracy.

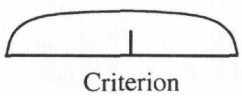
Empirical accounting research to date has examined the relation between consensus and accuracy in tasks for which accuracy was fairly high (Cell 4 in Figure 1). Ashton (1985) examined the relation between consensus and accuracy in two tasks involving sales forecasting and going-concern predictions, and found a strong positive relation (average correlation of 0.82) between consensus and accuracy in these tasks. Individuals exhibited fairly high accuracy at these tasks, suggesting that the strong positive relation between consensus and accuracy might be a mechanical result of high accuracy. The average correlation between subjects' judgments and the criterion for the sales prediction task was 0.74, and subjects correctly predicted 85 percent of the 40 going-concern cases. Subsequent research in this area also used tasks in which individuals exhibited fairly high accuracy (Keasey and Watson 1989).

**FIGURE 2**
**Distributions Representing Different Relations Between Accuracy and Consensus**

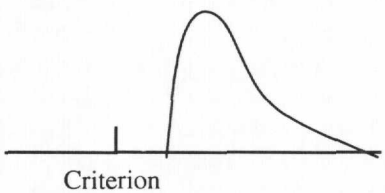
**Panel A: High Accuracy and High Consensus**



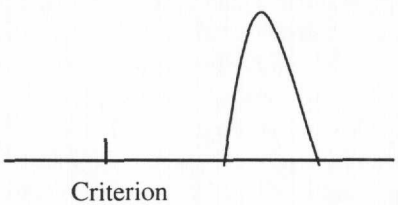

**Panel B: Low-to-Moderate Accuracy and Low Consensus**



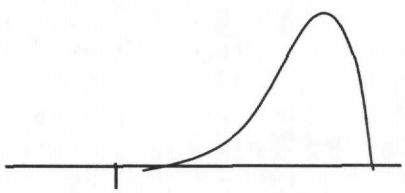

**Panel C: Low-to-Moderate Accuracy and High Consensus—Positive Relation Between Consensus and Accuracy**




**Panel D: Low-to-Moderate Accuracy and High Consensus—No Relation Between Consensus and Accuracy**




**Panel E: Low-to-Moderate Accuracy and High Consensus—Negative Relation Between Consensus and Accuracy**



Prior empirical research leaves open the question of whether consensus is a good predictor of accuracy when overall accuracy in an accounting task is low to moderate. Additionally, analytic research that has investigated the effect of low-to-moderate accuracy on the relation between consensus and accuracy has examined only tasks with dichotomous choices (Pincus 1990). She found that consensus was not a good surrogate for accuracy when the probability of making a correct choice is less than 0.5. We further this line of research by investigating empirically the relation between consensus and accuracy in low-to-moderate accuracy tasks in which the judgments are continuous variables. We also extend prior research by examining how the extent of industry-specific audit experience affects the relation between consensus and accuracy.

## CONSENSUS AND ACCURACY IN A LOW-TO-MODERATE ACCURACY AUDITING TASK

### Error Frequency Judgments

We investigate the relation between consensus and accuracy in an auditing task for which auditors are known to have low-to-moderate accuracy (Ashton 1991). Specifically, we examine auditors' judgments of the relative frequencies of errors in financial statement accounts in three industries: manufacturing, natural resources, and banking. As noted by Butt (1986), judgments of error frequency are relevant for assessing the risk of misstatement of financial statement accounts. These judgments affect the choice of audit procedures and, therefore, have implications for the efficiency and effectiveness of the audit. Several studies, including Wright and Ashton (1989), have estimated actual relative error frequencies using errors detected in a sample of audits. These estimated error frequencies provide a criterion against which to measure the accuracy of auditors' judgments of relative error frequencies.

Auditors' judgments of error frequencies may be formed based on direct or indirect experience (Nelson 1994; Butt 1988) or causal reasoning (Hogarth 1991). Auditors may base their estimates of error frequencies in an industry on their direct experience with errors in previous audits in that industry. They may also use indirect industry experience, such as industry-specific experiences of other auditors and historical error summaries. Alternatively, instead of relying on direct or indirect experiences with industry-specific error frequencies, auditors may base judgments of error frequencies on causal reasoning. In a discussion of Ashton (1991), Hogarth (1991) suggests that auditors use their causal understanding of accounting systems to generate educated guesses about error frequencies. Although such an understanding is developed through auditors' experiences, it is more general in nature than industry-specific experience. Auditors initially develop general audit experience as a result of education and training, with additional specific experience developed through directly performing a task (Marchant 1990).

### The Relation Between Consensus and Accuracy in Error Frequency Judgments

In order to assess the relation between consensus and accuracy in the error frequency task, it is necessary to consider how experience affects consensus and accuracy. Theoretical developments in psychology suggest that common experiences develop consensus between two individuals. Kenny (1991) identified several factors that positively influence consensus, including overlap and shared-meaning systems.[2] Overlap is the degree to which individuals observe the same information, while shared-meaning systems refers to the extent to which two individuals who observe the same information interpret this information in the same way. Overlap and shared-meaning systems develop through both indirect and direct experience. For example, accounting education provides common indirect experiences for accounting students (overlap) and also develops shared-meaning systems. Direct experience may also generate

[2] Kenny's (1991) model identifies six factors that relate to agreement among judges with respect to person perception in social psychology. Although person perception is not a professional judgment task, the research on consensus in this area may provide insights for consensus in professional judgment. We focus on two of Kenny's (1991) six factors (overlap and shared-meaning systems) that are most relevant for professional judgments.

observation of the same information and similar interpretation of the information. For example, Meixner and Welker (1988) find that consensus among auditors increases with the time spent with the same work group, under the supervision of the same supervisor.

Although common experiences are the basis for consensus, such experiences do not necessarily lead to accuracy. Accuracy (performance) in auditing requires task-specific knowledge developed from indirect or direct experience (Libby and Luft 1993). The degree to which a person's direct or indirect experience has provided this knowledge depends on both the quantity and the nature of that experience. For example, an auditor who rarely audits banks may have insufficient experience to assess error frequencies in banks' financial statement accounts. The extent to which greater direct experience leads to appropriate error frequency judgments depends on characteristics of the industry environment that affect the speed of knowledge acquisition. For example, an auditor who has audited many banks may have difficulty estimating relative error frequencies since errors are rare in the banking industry. Additionally, research indicates that auditors must have knowledge about the categories to which errors relate before they can properly encode errors (Bonner et al. 1997). This finding suggests that errors encountered early in an auditor's experience with a specialized industry may not be encoded properly. Finally, auditors with direct experience in several industries may confuse industry-specific error frequencies, leading to an "averaging" of error frequencies across industries (Nelson 1994).

This discussion of consensus and accuracy in error frequency judgments suggests that the relation between consensus and accuracy may depend on the industry experience of the auditor, and that the effect of this experience may differ by industry. In order to examine the effects of experience within different industries, we selected three of the industries examined in Ashton (1991). First, we chose the manufacturing industry since, through university and firm-specific training, it forms the foundation for most auditors' general knowledge of accounting systems and the auditing of those systems. Second, we chose two specialized industries, natural resources and banking, which differ in both the empirical frequency of errors per audit and the similarity of financial statement accounts to those in manufacturing. Maletta and Wright (1996), using the Wright and Ashton (1989) data, document that natural resources companies have a mean of 5.6 errors per audit, while banks have a mean of 2.9 errors (statistically different at $p < .05$ using Tukey pairwise comparison tests).[3] Auditors in natural resources have more opportunity to experience errors than do auditors in banking. Further, Figure 3 indicates that the financial statement accounts in natural resources are nearly identical to those in manufacturing, while accounts in banking are different. Given findings in Bonner et al. (1997), auditors in banking may need more time than those in natural resources to gain knowledge about financial statement accounts before they can begin to encode errors in memory.

We propose that auditors with little experience in an industry will form educated guesses about error frequencies by using causal reasoning based on their general accounting and auditing knowledge, which is developed by university education, firm training, and audit experience in other industries.[4] We believe that this approach will lead auditors to estimate error frequencies that are similar to error frequencies in the manufacturing industry for the following reasons. First, auditors' causal understanding of accounting systems is initially developed in university courses, which

[3] Ashton (1991) also collected error frequency judgments for the merchandising and insurance industries. We omit the merchandising industry since it similar to manufacturing. The correlation between financial statement account error frequencies between the merchandising and manufacturing industries is .90. The insurance industry was omitted because we wanted to choose two industries that differed in terms of the number of errors per audit. The mean of 4.3 errors per audit for the insurance industry is not statistically different from either the mean of 5.6 for natural resources or 2.9 for banking (Maletta and Wright 1996).

[4] Hogarth (1991) uses the phrase "educated guess" to represent a way auditors may make judgments given little or no experience with a particular issue. We do not consider random guessing a likely heuristic since auditors have both knowledge and incentives to perform better than randomly in real practice situations. We expect that auditors in the experiment also had incentives to perform better than randomly, since the participation was voluntary and was requested by the director of research for the firm.

**FIGURE 3**
**Financial Statement Accounts, Relative Error Frequencies, and the Correlation of Manufacturing Account Frequencies with the Other Industry Account Frequencies[a]**

| Financial Statement Account | Manufacturing | Natural Resources | Financial Statement Account | Banking |
|---|---|---|---|---|
| Accounts receivable (net) | 9.3 | 14.0 | Loans (net) | 11.1 |
| Inventory | 14.4 | 2.8 | Interest-bearing deposits with banks, Investment securities, Trading account securities, Federal funds sold and securities purchased | 13.9 |
| Prepaid expenses and other current assets | 3.3 | 4.7 | Cash on hand and due from banks | 8.3 |
| Property, plant and equipment (net) | 4.7 | 11.2 | Premises, leasehold improvements, and furniture and fixtures | 3.7 |
| Deferred charges and other noncurrent assets | 1.4 | 5.6 | Other assets | 12.0 |
| Accounts payable | 6.0 | 9.3 | Domestic demand deposits, Domestic time deposits, Foreign deposits, Federal funds purchased and securities sold, Acceptances outstanding | 8.3 |
| Accrued liabilities | 5.1 | 3.7 | Accrued taxes and expenses | 6.5 |
| Other current liabilities | 3.7 | 3.7 | Other liabilities | 9.3 |
| Deferred taxes | 0.5 | 3.7 | | |
| Stockholders' equity | 4.7 | 6.5 | | |
| Revenue | 7.9 | 6.5 | Interest and fees on loans, Other operating income | 6.5 |
| Cost of goods sold | 13.0 | 6.5 | | |
| Selling expenses | 5.1 | 0.9 | | |
| General and administration expenses | 14.4 | 7.5 | Other operating expenses, income tax expense | 11.1 |
| Other accounts[b] | 6.5 | 13.4 | Other accounts[b] | 9.3 |
| Total[b] | 100.0 | 100.0 | Total[b] | 100.0 |
| Correlation with manufacturing criterion | NA | .19 | Correlation with manufacturing criterion | .30 |

[a] Relative error frequencies for each industry are from Wright and Ashton (1989, Table 2).
[b] These frequencies were provided for the auditors.

typically use manufacturing firms as examples.[5] Second, most auditors have some direct experience with manufacturing errors. In our sample, manufacturing is the industry with which the greatest number of auditors has experience (383 auditors out of 453 in our sample). For auditors with low experience in specialized industries, manufacturing is the industry with which they have the most experience (see Table 1). Third, even if auditors don't have direct experience in manufacturing, their indirect experience may lead to greater knowledge of manufacturing errors than of other industry errors because there are more manufacturing audits and more errors are detected in manufacturing than in any other industry (Wright and Ashton 1989). Further, litigation alleging audit failure in manufacturing companies enhances the salience of manufacturing error frequencies. Manufacturing clients represented 47 percent of the auditor litigation cases in Stice (1991) and 57.8 percent of the cases in Palmrose (1988).

As auditors gain experience in specialized industries, we expect that they will reduce their reliance on their general audit knowledge and make frequency judgments based on industry-specific factors. These factors include direct experience with industry-specific errors, as well as indirect experience through other auditors within the industry. Improvement in industry-specific knowledge will develop as auditors gain more experience in an industry. However, the effect of both direct and indirect experience on improvement is dependent on both the idiosyncrasy of the industry's financial statement accounts and the frequency of errors in an industry. The more idiosyncratic the financial

[5] Textbooks typically develop the accounting model in the context of a merchandising firm, then expand the example to include a discussion of inventory in a manufacturing firm. See, for example, Stickney and Weil (1997, chaps. 3 and 4). Wright and Ashton (1989) found the distribution of relative error frequencies to be similar in manufacturing and merchandising firms (correlation of 0.90).

**TABLE 1**
**Mean Months of Total and Industry-Specific Audit Experience by Industry-Experience Group**

| | Industry-Experience Group | | | | | | |
|---|---|---|---|---|---|---|---|
| | Manufacturing | | Natural Resources | | Banking | | |
| Months of Experience[b] | High[a] (n = 220)[c] | Low[a] (n = 163) | High (n = 49) | Low (n = 63) | 36 mo. (n = 41) | 12 mo. (n = 117) | Low (n = 152) |
| Total audit experience | 84.15 | 44.97 | 86.96 | 63.30 | 108.95 | 83.21 | 45.94 |
| Manufacturing | **48.95**[d] | **5.20** | 42.40 | 26.36 | 25.38 | 22.36 | 16.60 |
| Natural Resources | 4.16 | 5.60 | **38.72** | **4.09** | 0.81 | 4.05 | 3.68 |
| Banking | 9.97 | 10.02 | 7.33 | 5.64 | **73.99** | **39.33** | **4.24** |

[a] High and Low refer to experience groups. Auditors in the high-industry-experience group have at least 12 months of experience in the particular industry, while auditors in low-industry-experience group have 1–11 months of industry experience. Once experience groups were determined for each industry, months of total and industry-specific audit experience were computed.

[b] Auditors determined their months of industry experience by first indicating the number of clients they had audited in the industry, estimating the months spent for each client, and then summing these months.

[c] A total of 453 auditors were in Ashton's (1991) sample. Auditors with no experience in an industry were excluded from the industry-experience groups. Seventy subjects were excluded for manufacturing, 341 subjects for natural resources, and 184 subjects for banking.

[d] Numbers in bold reflect the mean months of experience relevant for the particular industry.

statement accounts for an industry (i.e., the more they differ from prototypical accounts for industries such as manufacturing and merchandising), the longer it will take auditors to understand these accounts and properly encode errors. Further, if errors in an industry are rare or if the distribution of errors is company-specific, an auditor will not have the opportunity to learn realistic frequency judgments that will generalize to other companies. Therefore, even if industry experience does affect consensus, it may not improve auditors' error frequency judgments, as found by Ashton (1991). This suggests that the relation between consensus and accuracy may differ between industries.

For the manufacturing industry, if auditors with little experience in manufacturing rely on general accounting and auditing knowledge, they will be using knowledge that is based on the manufacturing industry. The use of general knowledge will lead to both consensus among less-experienced auditors and accuracy in their judgments, resulting in a positive relation between consensus and accuracy. Auditors with manufacturing experience have both general knowledge and their direct experience to use in forming their error frequency judgments, since errors occur relatively frequently in manufacturing audits (a mean of 5.5 errors per audit as reported in Maletta and Wright [1996]). Assuming that these errors generalize across companies, experience will also lead to a positive relation between consensus and accuracy. These arguments lead to the following hypothesis.

**H1:** Consensus among auditors' error frequency judgments in manufacturing is positively related to accuracy for all auditors, regardless of experience.

For the natural resources industry, we expect that auditors with little industry experience will tend to use their general audit experience to form educated guesses about the errors in natural resources leading to an error distribution similar to that for manufacturing. Notice in Figure 3 that manufacturing and natural resources have identical accounts, facilitating the association between the two industries. This will result in consensus among auditors with little experience in natural resources, but will not result in very accurate judgments since the correlation between the distribution of errors in manufacturing and those in natural resources is only .19 (see Figure 3). In contrast, at least some auditors with experience in natural resources should have had the opportunity to experience sufficient errors to develop empirical error distributions given the relatively high frequency of errors in natural resources documented by Maletta and Wright (1996). The fact that financial statement accounts in natural resources reflect prototypical accounts (i.e., those in manufacturing) also should aid auditors in quickly encoding these errors properly. Therefore, agreement among experienced auditors in natural resources will be driven at least in part by the accuracy of their error frequency judgments. Our hypotheses for natural resources are as follows:

**H2a:** Consensus of auditors' error frequency judgments for natural resources is not positively related to accuracy for auditors with little natural resources experience, but is positively related to accuracy for auditors with more natural resources experience.

**H2b:** Consensus of auditors' error frequency judgments for natural resources is positively related to actual manufacturing error frequencies for auditors with little natural resources experience, but is not positively related for auditors with more natural resources experience.

Consistent with our predictions for natural resources, we predict that auditors with limited experience in banking will rely on their general audit experience and estimate banking error frequencies that are similar to manufacturing error frequencies. Although manufacturing and banking are not directly related industries, they have similarities that allow for a match of accounts. For example, accounts receivable in manufacturing is similar to loans in banks (see Figure 3 for the manner in which the accounts were matched).[6] Figure 3 indicates that the correlation between

[6] Wright and Ashton (1989) had audit partners match banking accounts to manufacturing accounts. These matches are consistent with those reported in Figure 3.

error frequencies for manufacturing and banking is 0.30. Therefore, the use of general audit knowledge by inexperienced auditors should result in high consensus but fairly low accuracy and there should not be a positive relation between accuracy and consensus.

Compared to manufacturing or natural resources, banking auditors may need more experience to form appropriate estimates of errors in the banking industry. Banking has the fewest number of errors per audit of our three industries, perhaps due to the fact that the banking industry is highly regulated and audited on a frequent basis by governmental auditors in addition to banks' internal audit staff and external auditors. Audit procedures for banks are different from those for manufacturing and natural resources, as auditors tend to rely heavily on internal controls. Also, the uniqueness of financial statement accounts for banks may hinder auditors' initial ability to encode errors. As a result, auditors may need substantial experience before they can develop error distributions consistent with the industry averages. Until this experience is obtained, auditors' error frequency judgments are more likely to be driven by their general audit experience (i.e., more related to manufacturing error frequencies). Although we predict that the relation between consensus and accuracy will eventually become positive for experienced banking auditors, we have no basis on which to predict exactly how long it will take auditors to develop this knowledge. Therefore, we investigate different levels of experience for the banking industry to assess whether consensus eventually is driven by accuracy as auditors gain more experience in the banking industry. Our hypotheses for banking are as follows:

**H3a:** Consensus of auditors' error frequency judgments for banking is not positively related to accuracy for auditors with little banking experience, but is positively related to accuracy for auditors with more banking experience.

**H3b:** Consensus of auditors' error frequency judgments for banking is positively related to actual manufacturing error frequencies for auditors with little banking experience, but is not positively related for auditors with more banking experience.

## METHOD

The data analyzed in this study were collected by Ashton (1991). In her task, auditors from KPMG Peat Marwick estimated the relative frequencies with which errors affected several financial statement accounts for clients in five industries, three of which are used in this paper (see Figure 3). Specifically, auditors allocated 100 points among financial statement accounts based on their estimate of the percentage of total discovered errors that affected the account categories for that industry for clients of the firm. The accuracy of auditors' judgments was evaluated against criterion error rates found in a large sample of the financial statement accounts of the same firm's clients from 1984–1985 (Wright and Ashton 1989).[7] Auditors were instructed to consider the September 1984 to March 1985 time period used in the Wright and Ashton (1989) study when making their error frequency judgments. Auditors also provided demographic information, including months of experience with the firm and with clients in particular industries.[8] See Ashton (1991) for further details about the task and procedures.

[7] Wright and Ashton (1989) had audit managers respond to a highly structured questionnaire, providing data on 368 proposed adjustments (792 detected errors) in 186 U.S. audits of KPMG Peat Marwick clients with year ends from September 1984 to March 1985. Only errors that equaled or exceeded 20 percent of planning materiality (gauge) were included. Routine adjusting entries were excluded. Their study replicated and extended similar work and the results were consistent with prior results from another KPMG Peat Marwick sample (Hylas and Ashton 1982), as well as a client sample from another large accounting firm (Kreutzfeldt and Wallace 1986). Actual error rates, which are unknowable, are necessarily surrogated by known or discovered error.

[8] Auditors determined their months of industry experience by first indicating the number of clients they audited in the industry, estimating the months spent for each client, and then summing these months. This approach takes into account that an auditor working on two audits in the same industry simultaneously has two opportunities to observe errors.

**Measures of Accuracy and Consensus**

To compare accuracy and consensus, it is necessary to have comparable measures. Traditional measures of accuracy and consensus are not comparable because accuracy is measured for each individual, while consensus is measured for pairs of individuals. The counterparts of these traditional measures, pairwise accuracy and individual consensus, were developed by Ashton (1985). The present analysis relies on her definitions of individual accuracy and individual consensus.

Individual accuracy was calculated for each auditor who provided frequency estimates for account categories in an industry and who reported at least one month of experience in that industry.[9] Each auditor could have up to three accuracy scores, one per industry, if the auditor provided estimates for all three industries and met the one-month experience requirement for each industry. Specifically, accuracy was computed as the Pearson correlation between actual error frequencies (as established by Wright and Ashton [1989]) and an auditor's estimate of those frequencies. Individual consensus also was calculated for each auditor who provided frequency estimates for account categories in an industry, again yielding up to three scores per auditor. First, pairwise consensus was computed as the Pearson correlation between the error frequency estimates made by each pair of subjects. Then, following Ashton (1985), individual consensus was computed for each subject as the mean of the pairwise consensus scores across all auditor pairs in which the auditor was included—either all participants in an industry (Table 2) or all participants in an industry who have the same level of experience (Tables 3 and 4).[10] Although absolute accuracy and consensus could be used instead of correlational measures, a correlation better captures the relative nature of the auditing task. That is, although an auditor might not know that accounts receivable and inventory error frequencies in manufacturing are 9.3 and 14.4, respectively, he/she may know that inventory errors are more common than accounts receivable errors and score the accounts in that manner.

**Measures of Experience**

To examine the hypotheses related to experience, we split auditors into two groups based on the amount of experience they have in each industry, measured by months of industry-specific auditing experience.[11] Use of a continuous measure of experience is prohibited since consensus must be measured relative to others in a peer group. Although any split is somewhat arbitrary, we chose to split the two groups initially at one year of experience, which we believe is the minimum necessary to have developed an understanding of specialized industries and to have exposure to financial statement errors. We term auditors with less than one year of experience in an industry as "low experience" and auditors with 12 or more months

[9] All tables report results for auditors who have at least one month of experience in the industry to be consistent with Ashton (1991). We expected that at least some industry experience was necessary for auditors to recognize account titles in specialized industries. Additionally, practical considerations suggest that it is unlikely that an auditor with no experience in an industry would be used as a source to corroborate other auditors' judgments.

[10] All mean correlations were computed using Fisher's z transformation (Glass and Stanley 1970). This transformation corrects for skewness in the distribution of the sample correlation coefficient that occurs for nonzero correlations (Winkler and Hays 1975, 652–654). Strube (1998) shows that transformation prior to averaging correlations results in negligible bias for set sizes of ten and samples of 30. Our set sizes range from 10–14 financial statement accounts (see Figure 3) and our samples range from 41–220. As sample size and set size increase, the bias with and without transformation converge. Thus, transformation is at least as good as, and often better than, no transformation. Additionally, transformation corrects for heterogeneity problems that exist without the transformation.

[11] Ashton (1991) also gathered experience data on the auditor's rank within the firm and the number of clients each auditor had in an industry. We believe using months of industry experience is the best measure from Ashton's (1991) data to capture any experience-related effects. First, rank in the firm is not a good measure of experience with errors in specific industries. Second, we believe using months of experience is more liable to capture the likelihood of observing errors than number of clients. For example, it is unlikely that an auditor with two banking clients and five months of banking experience would have the same opportunity to learn error frequencies as would another auditor with two banking clients and 17 months of experience. Finally, Ashton (1991) found very similar results for the relation between accuracy and experience when experience was defined either as months of industry experience (as in our paper) or as number of clients.

of experience in an industry as "high experience." We also investigate an alternative cutoff for high experience of at least 36 months in the banking industry because error frequency knowledge may be acquired more slowly in that industry. The mean number of months of experience for each group by industry is reported in Table 1. Consensus in Tables 3 and 4 is measured as the correlation of an auditor's error frequency judgments with only those auditors in the same experience group.

## RESULTS AND DISCUSSION

Panel A of Table 2 presents descriptive statistics for consensus and accuracy of all auditors who responded for each industry, as well as the correlation between consensus and accuracy. Auditors exhibit low-to-moderate overall accuracy for the error frequency estimation task (r = .49, r = .22, and r = .22 for manufacturing, natural resources, and banking, respectively.) Auditors exhibit moderate-to-high overall consensus for both manufacturing and banking (r = .65 and r = .67, respectively) and moderate consensus for natural resources (r = .37). The correlation between consensus and accuracy is statistically positive for all three industries (.62, .25, and .18 for manufacturing, natural resources, and banking, respectively); however, the squared correlations (.38, .07, and .03) suggest that only in manufacturing is accuracy even moderately predictable from the degree of consensus among auditors. Overall, these results suggest that consensus may not be a good predictor of accuracy for low-to-moderate accuracy tasks.[12]

[12] Recall that Ashton (1991) included merchandising and insurance in her analyses. Although we are not directly interested in these two industries for reasons previously noted, we computed the relation between consensus and accuracy for these two industries for completeness. The results for merchandising are similar to those in manufacturing, and the results in insurance are fairly consistent with those in banking, although there is a stronger correlation between consensus and accuracy in insurance than in banking.

**TABLE 2**
**The Relation Between Consensus and Accuracy by Industry**

**Panel A: Descriptive Statistics—Mean (Standard Deviation)**

| | n[a] | Accuracy[b] | Consensus[c] | r(ACC,CON)[d] |
|---|---|---|---|---|
| Manufacturing | 383 | .49 (.19) | .65 (.13) | .62*** |
| Natural Resources | 112 | .22 (.27) | .37 (.13) | .25*** |
| Banking | 269 | .22 (.17) | .67 (.19) | .18*** |

**Panel B: $CON_i^c = \alpha + \beta(ACC_i^b) + \varepsilon_I$ – coefficient (t-statistic)**

| | α | β | $R^2$ |
|---|---|---|---|
| Manufacturing | .44*** | .43*** | .38 |
| | (30.28) | (15.36) | |
| Natural Resources | .34*** | .13*** | .07 |
| | (21.29) | (2.76) | |
| Banking | .62*** | .20*** | .03 |
| | (33.25) | (2.94) | |

*** Significant at less than .01.

[a] n is the number of auditors in each industry group. An auditor had to have at least one month of experience in an industry to be included in an industry group.

[b] Accuracy ( ACC) is the average of correlations of auditors' frequency judgments and the criterion vector from Wright and Ashton (1989).

[c] Consensus (CON) is the average of correlations of every auditor pair within each industry.

[d] r is the correlation of ACC and CON.

**Tests of Experience Effects on the Relation Between Consensus and Accuracy**

Panel A of Table 3 shows the mean accuracy, mean consensus, the correlation between consensus and accuracy and the number of subjects per experience group. Consistent with H1, consensus in manufacturing is positively related to accuracy in both the high- and low-experience groups ($r = .58$, $p < .01$, and $r = .67$, $p < .01$, respectively). In natural resources, the correlation between consensus and accuracy is statistically positive for the high experience group ($r = .61$, $p < .01$), but it is not statistically significant for the low experience group ($r = -.04$, $p > .10$), as predicted by H2a. Finally, for banking, there is a small, but statistically significant correlation between consensus and accuracy for both experience groups ($r = .15$, $p < .10$ and $r = .18$, $p < .05$ for the high- and low-experience groups, respectively). Although this is not statistically consistent with H3a, the magnitude of the correlation is so low that it suggests little ability to predict accuracy from consensus in banking for either experience group.

**TABLE 3**
**The Relation Between Consensus and Accuracy by Industry and Experience Level**

**Panel A: Descriptive Statistics—Mean (Standard Deviation)**

| Industry | | High Experience[a] | Low Experience[a] |
|---|---|---|---|
| Manufacturing | ACC[b] | .49 (.20) | .50 (.17) |
| | CON[c] | .64 (.13) | .66 (.13) |
| | r[b] | .58*** | .67*** |
| | n[b] | 220 | 163 |
| Natural Resources | ACC | .32 (.24) | .15 (.27) |
| | CON | .44 (.19) | .40 (.18) |
| | r | .61*** | −.04 |
| | n | 49 | 63 |

| Banking | | ≥ 12 months | ≥ 36 months | |
|---|---|---|---|---|
| | ACC | .23 (.18) | .24 (.19) | .21 (.17) |
| | CON | .68 (.20) | .66 (.25) | .67 (.18) |
| | r | .15* | .43*** | .18** |
| | n | 117 | 41 | 152 |

**Panel B: $CON_i = \alpha_1 + \alpha_2(EXP) + \beta_1(ACC_i) + \beta_2(ACC_i*EXP) + \varepsilon_1$ – coefficient (t-statistic)**

| | $\alpha_1$ | $\alpha_2$ | $\beta_1$ | $\beta_2$ | $R^2$ |
|---|---|---|---|---|---|
| Manufacturing | .42*** | .05 | .50*** | −.12*** | .39 |
| | (17.14) | (1.47) | (10.83) | (-2.16) | |
| Natural Resources | .41*** | −.12*** | −.02 | .50*** | .18 |
| | (16.90) | (-2.56) | (−.30) | (3.93) | |
| Banking (high EXP ≥ 12 mo.) | .63*** | .01 | .19*** | −.02 | .03 |
| | (25.45) | (.16) | (2.05) | (−.14) | |
| Banking (high EXP ≥ 36 mo.) | .63*** | −.10* | .19** | .36** | .08 |
| | (25.48) | (−1.84) | (2.06) | (1.98) | |

***, **, and * indicate significantly different from zero at less than .01, .05, and .1, respectively.

[a] Experience (EXP) is high (low) if the auditor has greater than or equal to 12 months of experience in the industry (between 1 and 11 months of experience in the industry). In the regression equation of Panel B, EXP is a dummy variable coded 1 for high experience and 0 for low experience. For banking, results are also reported for high experience defined as greater than or equal to 36 months.

[b] ACC and r are defined in Table 2 and n is the number of auditors in each experience group for each industry.

[c] CON is the average of correlations of every auditor pair within each industry experience level.

Our discussion of banking suggested that auditors might need substantial experience in banking before the relation between consensus and accuracy is positive. We also examined the correlation between consensus and accuracy for auditors with at least 24 months of banking experience and at least 36 months of banking experience. The correlation for auditors with at least 24 months of experience was .23 ($p < .05$), and, as shown in Panel A of Table 3, the correlation between consensus and accuracy for auditors with at least 36 months of banking experience is 0.43 ($p < .01$).

Hypothesis 1 predicts a positive relation between consensus and accuracy in manufacturing for all experience levels, while H2a and H3a predict that the relation between consensus and accuracy should differ between the low- and high-experience auditors in natural resources and banking, respectively. To test statistically whether the relation between consensus and accuracy differs between the two experience groups in each industry, we use regressions modeling consensus as a function of experience and accuracy. The model uses a dummy variable to allow for differences in consensus between experience groups and an interactive dummy variable which allows the relation between consensus and accuracy to differ between the two groups (Pindyck and Rubinfeld 1981, 111–116).

$$CON_i = \alpha_1 + \alpha_2(EXP) + \beta_1(ACC_i) + \beta_2(ACC_i{*}EXP) + \varepsilon_1 \quad (1)$$

where:

$CON_i$ = individual correlational consensus for individual i;
EXP = 1 if the auditor is high experience, 0 otherwise;
$ACC_i$ = individual correlational accuracy for individual i;
$\varepsilon_i$ = individual error term.

Although our hypotheses are related to differences in the coefficient on accuracy (ACC) between the low- and high-experience group, we also allow the intercepts to differ between the two groups. In the regression, $\alpha_1$ is the intercept for the low-experience group. For the high-experience group, $\alpha_2$ is the difference from the low-experience group for the part of consensus that is unrelated to accuracy; therefore, $\alpha_1 + \alpha_2$ is the intercept for the high-experience group. $\beta_1$ measures the relation between accuracy and consensus for the low-experience group. $\beta_2$ is the incremental effect of experience for the relation between consensus and accuracy, with $\beta_1 + \beta_2$ measuring the relation between consensus and accuracy for the high-experience group. Hypotheses 1, 2a, and 3a are tested by examining $\beta_1$ and $\beta_2$.

Panel B of Table 3 shows the results of the separate regressions for manufacturing, natural resources, and banking.[13] As predicted by H1, the relation between consensus and accuracy in manufacturing is positive for both experience groups ($\beta_1 = .50$, $\beta_1 + \beta_2 = .38$), although the relation for the high-experience group is statistically smaller than that for the low-experience group ($\beta_2 = -.12$, $p < .01$). Finally, the adjusted $R^2$ of the regression in Table 3 (.39) is nearly identical to that in Table 2 (.38), suggesting that splitting auditors by experience does not add explanatory power. Overall, these results are consistent with auditors' accounting education and training providing both consensus among auditors and a moderately strong foundation for assessing appropriate error frequencies. Experience in manufacturing does not appear to improve, and may even diminish, the relation between consensus and accuracy.

For natural resources, consensus is not related to accuracy for the low-experience group as the coefficient relating consensus and accuracy is not significantly different from zero ($\beta_1 = -.02$, $p > .10$); however, consensus is positively related to accuracy for the high-experience group as indicated by the sum of the coefficients $\beta_1$ and $\beta_2$ ($-.02 + .50 = .48$, $p < .01$). Consistent with the significance of $\beta_2$, the $R^2$ of the equation also increases when auditors are

[13] Our dependent and independent variables (consensus and accuracy, respectively) in our regression are correlations that do not conform to the distributional assumptions of OLS. To test the validity of our results, we also performed regressions using the approximate randomization tests, making no assumptions about the distribution (Noreen 1989). We find the significance levels for the β coefficients tested in our regressions under the randomization procedure to be almost identical to the p-values in the standard regressions, providing assurance that our results from the OLS regressions are reliable, despite violations of OLS regression assumptions.

split into two experience groups, from .07 in Table 2 to .18 in Table 3. These results are consistent with H2a and suggest that consensus among less-experienced auditors in natural resources is unrelated to the accuracy of their error frequency assessments. In contrast, at least some of the consensus among auditors with high experience in natural resources is due to these auditors learning correct error frequency assessments.

Finally, the regression in Table 3 for banking indicates that allowing the relation between accuracy and consensus to vary between low- and high-experience groups does not improve upon the overall model in Table 2 ($R^2$ = .03). The relation between consensus and accuracy for the low-experience group is statistically positive, but small in magnitude ($\beta_1$ = .19, $p < .01$). In addition, the magnitude of the relation does not differ between low- and high-experience groups ($\beta_2$ = –.02, $p > .10$).

Since it is possible that auditors learn error frequencies in banking only after substantial industry experience, we redefined the high-experience group as auditors who have at least 36 months of banking experience and re-tested H3a. As shown in Table 3, Panel B, the new $\beta_2$ coefficient indicates a significantly stronger relation between consensus and accuracy for auditors who have at least 36 months of experience than for auditors who have less than one year of experience in banking ($\beta_2$ = .36, $p < .05$). Correspondingly, the $R^2$ of .08 for this regression also shows some improvement over the $R^2$ of .03 for the regression using a 12-month cutoff. Overall, these results are similar to results found for the natural resources industry using the 12-month cut-off for low and high experience. These results support the belief that generating consensus among auditors due to agreement on correct error frequencies takes longer in banking due to the less-frequent observation of errors.[14]

### Tests of "Educated Guess" Heuristics

Our results for the low-experience groups in natural resources and banking suggest that consensus among auditors with little experience in these industries is not primarily a result of auditors' accuracy. We predicted that less-experienced auditors use their general audit experience to assess error frequencies in specialized industries, and that this "educated guess" heuristic will lead to error frequency distributions similar to those in manufacturing. To test H2b and H3b, we performed a second analysis in which the actual manufacturing error frequencies were used as the criterion variable rather than the actual industry error frequencies in natural resources or banking. We computed the correlation between auditors' frequency judgments and the manufacturing error criterion, and termed this correlation MFACC. The measurement of consensus is unchanged. Descriptive statistics for MFACC and the correlation between MFACC and consensus are shown in Panel A of Table 4, while the corresponding regressions using MFACC in place of ACC are shown in Panel B of Table 4. Panel C shows regressions including both ACC and MFACC, which demonstrate whether MFACC provides an incremental contribution to ACC in explaining consensus.

For natural resources, Panel A of Table 4 shows that, on average, error frequencies of low-experience auditors were closer to the manufacturing error criterion (MFACC = .31) than the natural resources error criterion (ACC = .15 in Table 3). In addition, those auditors who agreed with each other did so because they made educated guesses consistent with the manufacturing error distribution (r(MFACC,CON) = .82, $p < .01$). In contrast, the error assessments of the high-experience auditors were not consistent with the manufacturing error frequencies. The correlation between their judgments and the manufacturing

[14] We tested the sensitivity of our results for H1, H2a, and H3a for all three industries by using two additional cutoff points for high experience: at least 24 months' experience and at least 36 months' experience. In general, except as noted for banking, the results of these regressions were qualitatively similar to the results reported in Table 3 where high experience is defined as at least 12 months of experience. In manufacturing, the $\beta_2$ coefficient for the high-experience group remained significantly negative for these higher cutoff levels. In fact, the magnitude of the coefficient increased slightly, from –.12 for 12 months to –0.17 (–29) for the 24- (36-) month cutoff. For natural resources, the $\beta_2$ coefficient increased slightly from 12 to 24 months (0.50 to 0.65), then decreased to 0.47 for the 36-month cutoff. As noted in Table 3, the $\beta_2$ coefficient for banking was significantly positive for the 36-month cutoff (0.36). The coefficient for the 24-month cutoff was insignificantly different from zero, similar to the results reported for the 12-month cutoff in Table 3.

**TABLE 4**
**The Relation Between Consensus and the Manufacturing Criterion (MFACC) by Industry and Experience Level**

**Panel A: Descriptive Statistics—Mean (Standard Deviation)**

| Industry | | High Experience[a] | Low Experience[a] |
|---|---|---|---|
| Natural Resources | MFACC[b] | .07 (.28) | .31 (.29) |
| | CON[a] | .44 (.19) | .40 (.18) |
| | r[c] | −.29*** | .82*** |
| | n[a] | 49 | 63 |

| | | ≥ 12 months | ≥ 36 months | |
|---|---|---|---|---|
| Banking | MFACC | .32 (.22) | .32 (.22) | .30 (.21) |
| | CON | .68 (.20) | .66 (.25) | .67 (.18) |
| | r | .42*** | .34** | .37** |
| | n | 117 | 41 | 152 |

**Panel B: $CON_i = \alpha_1 + \alpha_2(EXP) + \beta_1(MFACC_i) + \beta_2(MFACC_i*EXP) + \varepsilon_1$ – coefficient (t-statistic)**

| | $\alpha_1$ | $\alpha_2$ | $\beta_1$ | $\beta_2$ | $R^2$ |
|---|---|---|---|---|---|
| Natural Resources | .25*** | .21*** | .50*** | −.70*** | .41 |
| | (9.27) | (6.19) | (8.07) | (−7.21) | |
| Banking (high EXP ≥ 12 mo.) | .58*** | −.03 | .31*** | .08 | .15 |
| | (23.33) | (−.70) | (4.60) | (0.81) | |
| Banking (high EXP ≥ 36 mo.) | .58*** | −.04 | .31*** | .08 | .13 |
| | (23.35) | (−.70) | (4.40) | (0.51) | |

**Panel C: $CON_i = \alpha_1 + \alpha_2(EXP) + \beta_1(ACC_i)^d + \beta_2(ACC_i*EXP) + \beta_3(MFACC_i) + \beta_4(MFACC_i*EXP) + \varepsilon_1$ – coefficient (t-statistic)**

| | $\alpha_1$ | $\alpha_2$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | $R^2$ |
|---|---|---|---|---|---|---|---|
| Natural Resources | .25*** | .05 | −.01 | .46*** | .50*** | −.58*** | .55 |
| | (9.87) | (1.29) | (−0.25) | (4.69) | (9.13) | (−6.53) | |
| Banking (high EXP ≥ 12 mo.) | .49*** | −.09 | .32*** | .12 | .38*** | .17 | .26 |
| | (14.93) | (−1.67) | (3.82) | (0.95) | (5.74) | (1.61) | |
| Banking (high EXP ≥ 36 mo.) | .49*** | −.24*** | .32*** | .52*** | .38*** | .30** | .31 |
| | (14.99) | (−3.21) | (3.83) | (3.08) | (5.76) | (2.04) | |

***, **, and * indicate significantly different from zero at less than .01, .05, and .1, respectively.
[a]Experience, CON, and n are defined in Table 3.
[b]MFACC is the correlation between an auditor's frequency judgments and the manufacturing criterion vector.
[c]r is the correlation of MFACC and CON.
[d]ACC is defined in Table 2.

error criterion is low (MFACC = .07, $p > .10$) and the correlation between consensus and MFACC for the high-experience auditors is negative (–.29, $p < .01$).

The regression in Panel B of Table 4 confirms the results of Panel A, with a statistically significant coefficient for the relation between CON and MFACC for the low-experience group ($\beta_1 = .50$, $p < .01$), and a negative coefficient for the high-experience group (.50 –.70 = –.20). The regression in Panel C also demonstrates that MFACC provides an incremental contribution to ACC in explaining consensus for only the low-experience group. In the regression, the $\beta_1$ coefficient relating consensus and accuracy for the low-experience group is not statistically different from zero, while the $\beta_3$ coefficient relating consensus and MFACC is statistically positive ($\beta_3 = .50$, $p < .01$). In contrast, for the high-experience group, the coefficient relating consensus and ACC is statistically positive ($\beta_1 + \beta_2 = .45$, $p < .01$), while the coefficient relating consensus and MFACC is not statistically different from zero ($\beta_3 + \beta_4 = -.08$, $p > .10$). These results are consistent with H2b.[15]

The results for the banking industry are not as clear as the results for natural resources. Panel A of Table 4 shows that the correlation between auditors' error frequency judgments and the manufacturing error criterion appears somewhat higher than the corresponding correlation with the banking error criterion (.30 vs. .21 for the low-experience group and .32 vs. .23 for the ≥ 12 months-experience group). The correlation between consensus and MFACC in Table 4 is also higher than the correlation between consensus and ACC in Table 3 (.37 vs. .18 for the low-experience group and .42 vs. .15 for the ≥ 12 months-experience group). We also examined whether increased experience (≥ 36 months) in banking reduced the relation between auditors' judgments and the manufacturing criterion. The correlation between consensus and MFACC is .34 ($p < .05$), which is lower than both the 12-months experience result for MFACC and the correlation between consensus and ACC in Table 3 for the ≥ 36 months-experience group ($r = .43$).

Our statistical tests in Panels B and C demonstrate that MFACC does play a role in explaining consensus for both the low- and high-experience groups. In Panel B, the $\beta_1$ coefficient for the low-experience group is statistically positive ($\beta_1 = .31$, $p < .05$), while the $\beta_2$ coefficient is not significantly different from zero for either the 12-months or 36-months high-experience group ($\beta_2 = .08$ for both groupings). This result suggests that there is a positive relation between consensus and MFACC for the high-experience banking auditors ($\beta_1 + \beta_2 = .30$, $p < .01$), and that the magnitude of this relation is not statistically different from the low-experience group. This result is not consistent with our predictions in H3b for the high-experience group.

Panel C also demonstrates that MFACC provides an incremental contribution beyond ACC in explaining consensus for both the low- and high-experience groups. The regression with high experience defined as ≥12 months indicates that both ACC ($\beta_1 = .32$, $p < .01$) and MFACC ($\beta_3 = .38$, $p < .01$) are significantly related to consensus for the low-experience group. These relations also hold for the group with ≥ 12 months of banking experience, and are of the same magnitude as that for the low-experience auditors ($\beta_2 = .12$, $p > .10$, and $\beta_4 = .17$, $p > .10$). When high experience is redefined as ≥ 36 months of banking experience, the relation between consensus and both ACC and MFACC becomes stronger for the high-experience auditors than for the low-experience auditors ($\beta_2 = .52$, $p < .01$, and $\beta_4 = .30$, $p < .05$). Again, this result is not consistent with H3b, and suggests that auditors with high experience in banking may be using a heuristic consistent with MFACC.

Finally, given the proposition that auditors may average industry-specific error frequencies (Nelson 1994), we also investigated the possibility that auditors' error frequency judgments correspond better with the "average" errors across

[15] We tested the sensitivity of our results for H2b by using two additional cutoff points for high experience: at least 24 months' experience and at least 36 months' experience for the regressions in Panel B of Table 4. The results of these regressions were qualitatively similar to the results reported in Table 4 where high experience is defined as at least 12 months of experience. The $\beta_2$ was significantly negative and of similar magnitude for all three cutoff levels.

industries (instead of the manufacturing criterion), by using the total error frequency distribution in Wright and Ashton (1989). The regressions using total error frequencies were very similar to those reported in Panel B of Table 4, although the $R^2$s were smaller. This suggests that total error frequencies are not a better heuristic than manufacturing error frequencies. It should be noted, however, that this test does not measure the average errors experienced by each auditor based on his or her personal industry experience.[16]

## CONCLUSION

This study extends Ashton (1985) by examining the relation between accuracy and consensus in an error frequency estimation task for three industries in which overall accuracy is lower than for the tasks examined in Ashton (1985). We find that the strong relation between consensus and accuracy found in Ashton's (1985) sales and going-concern prediction tasks is not found in this error frequency task. At best, we find that accuracy is moderately predictable from consensus for all auditors in manufacturing, for high-experience auditors (12 or more months of experience) in natural resources, and for very-high-experience auditors (36 or more months of experience) in banking. We propose that these industry differences are due to at least four factors: (1) the initial training that auditors receive in university and firm programs that focus on the manufacturing industry; (2) the specialized nature of accounts in industries such as banking that may delay proper encoding of errors; (3) the relative number of mean errors per audit found in manufacturing (5.5) and natural resources (5.6) vs. the number found in banking (2.9); and (4) the nature of the control environment and audit procedures used for each of these industries. These factors affect the accuracy with which auditors can perform the error frequency task, and may affect the degree to which auditors can develop overlap (observing the same information) and shared-meaning systems (interpreting the observed information in the same way), both of which form a basis for consensus about industry-specific error frequencies.

An interesting finding of this study is that when auditors do not have the requisite knowledge to judge specialized industry error frequencies, their frequencies are consistent, at least somewhat, with the use of manufacturing error frequencies even though there is a very low empirical correlation between the distribution of errors in specialized industries and the manufacturing industry. For natural resources, the use of this heuristic was better at explaining consensus among low-experienced auditors than the use of the true error frequencies. For banking, consensus among auditors in all experience groups was explained by agreement with both the manufacturing and banking error frequencies.

The results are subject to certain limitations and should be interpreted in light of these. It can be argued that auditors neither perform this exact task in practice nor do they have any reason to know national error frequencies. Although this may explain why their accuracy in the task is low, it does not address why consensus is high. The auditors were apparently doing something systematic to have demonstrated significant levels of consensus in their error frequency judgments. Either they all were basing their error frequencies on similar experiences that differ from national averages (an unlikely scenario), or they all were using similar "educated guess" heuristics to form their error frequency judgments, such as the use of manufacturing error frequencies. The manufacturing criterion has an obvious relation to natural resources (the accounts match identically); its relation to banking is less obvious, but not unreasonable. Otherwise, there should have been no greater explanatory power of the manufacturing criterion as a heuristic if the frequencies had no meaning to the auditors. Our goal was not to identify the precise heuristic that

[16] We performed one additional sensitivity analysis for specialized industries. Although we predict that low-experience auditors will have knowledge about the manufacturing criterion from either direct or indirect experience, it is possible that indirect experience is not enough for auditors to know the manufacturing criterion. Therefore, we re-ran the analyses presented in Table 4 eliminating the subjects with no manufacturing experience. In Natural Resources, 8 (10) of the 49 (63) high- (low-) experience auditors had no manufacturing experience. In Banking, 19 (20) of the 117 (152) high- (low-) experience auditors had no manufacturing experience. The results were nearly identical to those reported in Table 4.

auditors were using, but to offer one based on general knowledge as a possibility. The investigation of other heuristics such as using the error frequencies of an auditor's most recent client (availability heuristic) or the error frequencies of an auditor's primary industry is a possible area for future research.

Our study has several implications for audit practice. First, our study suggests that confidence in audit judgments due to substantial consensus among auditors may be misplaced for complex tasks where overall accuracy is likely to be low. Although greater confidence may be warranted for auditors with experience in an industry, our results indicate that consensus among even highly experienced auditors does not guarantee judgment accuracy. One consequence of relying on a poor surrogate for accuracy is that auditors are unlikely to recognize the inaccuracy of unaided decisions and are therefore unlikely to demand or accept new approaches that offer improvements in accuracy. In an increasingly litigious and competitive environment, the failure to recognize when confidence in audit judgment quality is misplaced could be costly.

Second, our paper has implications for decisions related to risk assessments in auditing, particularly for new clients. Assessments of error frequencies are an input to estimating inherent and control risk for companies. While risk assessments for continuing clients may be based on historical client-specific errors, industry base rates for errors are likely to play a bigger part in risk assessments for new clients. Risk assessments have an important role in client-acceptance decisions (Johnstone 2000) and audit planning (AICPA 1990). Our study suggests that auditing firms may have more confidence than is warranted in client-acceptance decisions that are based in part upon agreement about expected error frequencies for the client. This may be particularly true if the potential client is in an industry in which the firm does not have extensive experience. Additionally, audit procedures based on agreed-upon error frequencies may be inefficient or even ineffective. Auditors may spend excess time auditing an account they all believed would have more errors, while not spending sufficient time auditing an account they all expected to be relatively error-free. It is possible that such underauditing might allow errors to escape detection, leading to possible future litigation against the firm.

Finally, our paper highlights that relationships documented in one industry may not extrapolate to other industries. Specifically, we found that the moderately strong relation between consensus and accuracy found for all experience levels in manufacturing did not hold for specialized industries. In natural resources and banking, moderate-to-very high experience was needed before consensus was even moderately related to accuracy. This finding has several implications for both audit practice and research. First, it affirms the importance placed on industry considerations found in the professional auditing standards and industry specialization within most auditing firms (Maletta and Wright 1996). Second, it highlights the need for audit firms to focus industry training on relevant category structures (e.g., financial statement accounts or transaction cycles) of specialized industries as early as possible (preferably before going to the field) so that less-experienced auditors can properly encode information from their direct experiences.

With respect to audit research, our results suggest that researchers should exercise caution in using consensus as a surrogate for accuracy, and in extrapolating results found for manufacturing or merchandising to specialized industries. Much of the audit research has been done using a manufacturing context (Solomon and Shields 1995). Our findings indicate that the results of audit research using manufacturing contexts may not have direct implications for other industries. Additionally, our results suggest that the industry environment may impact the speed of knowledge acquisition such that different amounts of industry experience may be needed across industries before benefits of industry specialization can be realized. This result may be helpful in the development of a theory that explicates how experience is translated into knowledge.

This study provides other promising avenues for future research on these issues. For example,

we found agreement among auditors that was often unrelated to both professional experience and judgmental accuracy. It would be useful to examine the causes of agreement among auditors, and how these causes relate to the accuracy of judgment. Kenny's (1991) model suggests important dimensions of consensus that interact. Experimental investigations of these dimensions could be fruitful and theory-based, in contrast to most investigations of consensus to date. Future studies could also examine other tasks and institutional characteristics that may moderate the relation between accuracy and consensus. This research could address the individual and interactive effects of institutional factors identified in this study, including training, specialized nature of accounts, empirical error frequencies, and nature of control environments.

## REFERENCES

American Institute of Certified Public Accountants (AICPA). 1990. *Audit Risk and Materiality in Conducting an Audit.* Statements on Auditing Standards AU Section 312. New York, NY: AICPA.

Ashton, A. H. 1985. Does consensus imply accuracy in accounting studies of decision making? *The Accounting Review* 60 (April): 173–185.

———. 1991. Experience and error frequency knowledge as potential determinants of audit expertise. *The Accounting Review* 66 (April): 218–239.

Bonner, S. E., R. Libby, and M. W. Nelson. 1997. Audit category knowledge as a precondition to learning from experience. *Accounting, Organizations and Society* 22 (July): 387–410.

Butt, J. 1986. Frequency judgments in an auditing-related task. Doctoral dissertation, University of Michigan.

———. 1988. Frequency judgment in an auditing-related task. *Journal of Accounting Research* 26 (Autumn): 315–330.

Glass, G., and J. C. Stanley. 1970. *Statistical Methods in Education and Psychology.* Englewood Cliffs, NJ: Prentice Hall.

Hogarth, R. M. 1991. A perspective on cognitive research in accounting. *The Accounting Review* 66 (April): 227–290.

Hylas, R. E., and R. H. Ashton. 1982. Audit detection of financial statement errors. *The Accounting Review* 57 (October): 751–765.

Johnstone, K. 2000. Client-acceptance decisions: Simultaneous effects of client business risk, audit risk, auditor business risk, and risk adaptation. *Auditing: A Journal of Practice & Theory* (Spring): 1–26.

Keasey, K., and R. Watson. 1989. Consensus and accuracy in accounting studies of decision-making: A note on a new measure of consensus. *Accounting, Organizations and Society* 14: 337–345.

Kenny, D. A. 1991. A general model of consensus and accuracy in interpersonal perception. *Psychological Review* 98 (April): 155–163.

Kreutzfeldt, R. W., and W. A. Wallace. 1986. Error characteristics in audit populations: Their profile and relationship to environmental factors. *Auditing: A Journal of Practice & Theory* 6 (Fall): 20–43.

Libby, R. 1981. *Accounting and Human Information Processing: Theory and Applications.* Englewood Cliffs, NJ: Prentice Hall, Inc.

———, and J. Luft. 1993. Determinants of judgment performance in accounting settings: Ability, knowledge, motivation, and environment. *Accounting, Organizations and Society* 18 (July): 425–450.

Maletta, M., and A. Wright. 1996. Audit evidence planning: An examination of industry error characteristics. *Auditing: A Journal of Practice & Theory* 15 (Spring): 71–86.

Marchant, G. 1990. Discussion of determinants of auditor expertise. *Journal of Accounting Research* 28 (Supplement): 21–28.

Meixner, W. F., and R. B. Welker. 1988. Judgment consensus and auditor experience: An examination of organizational relations. *The Accounting Review* 63 (July): 505–513.

Nelson, M. 1994. The learning and application of frequency knowledge in audit judgment. *Journal of Accounting Literature* 13: 185–211.

Noreen, E. W. 1989. *Computer Intensive Methods for Testing Hypotheses: An Introduction*. New York, NY: Wiley.

Palmrose, Z-V. 1988. An analysis of auditor litigation and audit service quality. *The Accounting Review* 63 (January): 55–73.

Pincus, K. V. 1990. Audit judgment consensus: A model for dichotomous decisions. *Auditing: A Journal of Practice & Theory* 9 (Spring): 150–166.

Pindyck, R., and D. Rubinfeld. 1981. *Econometric Models and Econometric Forecasts*. New York, NY: McGraw-Hill.

Solomon, I., and M. D. Shields. 1995. Judgment and decision-making research in auditing. In *Judgment and Decision-Making Research in Accounting and Auditing*, edited by R. H. Ashton and A. H. Ashton, 137–175. New York, NY: Cambridge University Press.

Stice, J. D. 1991. Using financial and market information to identify pre-engagement factors associated with lawsuits against auditors. *The Accounting Review* 66 (July): 516–534.

Stickney, C. P., and R. L. Weil. 1997. *Financial Accounting: An Introduction to Concepts, Methods, and Uses*. Fort Worth, TX: The Dryden Press.

Strube, M. J. 1988. Averaging correlation coefficients: Influence of heterogeneity and set size. *Journal of Applied Psychology* 73: 559–568.

Winkler, R. L., and W. L. Hays. 1975. *Statistics: Probability, Inference and Decision*. New York, NY: Holt, Rinehart and Winston.

Wright, A., and R. H. Ashton. 1989. Identifying audit adjustments with attention-directing procedures. *The Accounting Review* 64 (October): 710–728.

Wright, W. 1987. Audit judgment, consensus and experience. In *Behavioral Accounting Research: A Critical Analysis*, edited by K. Ferris, 305–327. Columbus, OH: Publishing Horizons, Inc.