December 2000

The Effects of Information Request Ambiguity and Construct Incongruence on Query Development

A. Faye Borthick
Georgia State University

Paul Bowen
University of Queensland

Donald Jones
Georgia State University

Michael Tse
Jardine Fleming

The Effects of Information Request Ambiguity and Construct Incongruence on Query Development

A. Faye Borthick[1],
Georgia State University, Atlanta GA 30302-4050

Paul L. Bowen[2]
University of Queensland

Donald R. Jones[3]
Georgia State University

Michael Hung Kam Tse[4]
Jardine Fleming

Abstract:
This paper examines the effect on users ability to develop SQL queries for information requests that are (1) ambiguous and (2) have incongruities among the information request, the query syntax, and the data representation. Based on an experiment in which participants used a relational database query language, ambiguity in information requests adversely affected accuracy and efficiency. Incongruities among the information request, the query syntax, and the data representation adversely affected accuracy, efficiency, and confidence.
The results for ambiguity suggest that organizations might elicit better query development if end users were sensitized to the nature of ambiguities that could arise in their business contexts. End users could translate natural language queries into pseudo-SQL that could be examined for precision before the queries were developed. The results for incongruence suggest that better query development might ensue if semantic distances could be reduced by giving users data representations and database views that maximize construct congruence for the kinds of queries in typical domains.

Keywords: Query development; Requirements ambiguity; Construct congruence; Web front end.

1. Introduction

Even before Web front ends to legacy data on mainframes were created, managers and staff members were experiencing the need to retrieve and analyze data from various sources. They could not rely on IS professionals for ad hoc data retrieval and analysis because of the demands on IS professionals time for developing organizational systems. With Web front ends, however, data can be accessible to anyone with a Web browser (Deck, 1999; Horwitt, 1999). In some organizations, young staff members, writing their own queries, are driving the business with their analyses (Gantz, 1999). Thus, in self defense, managers and staff members are discovering that their competitiveness depends on their ability to develop database queries.

[1] Dr. A. Faye Borthick is Professor at the School of Accountancy, borthick@gsu.edu.
[2] Dr. Paul L. Bowen is Senior Lecturer at the Department of Commerce, bowen@lorien.commerce.uq.edu.au.
[3] Dr. Donald L. Jones is Assistant Professor at the School of Accountancy, donjones@gsu.edu.
[4] Mr.Michael Hung Kam Tse is a Corporate Finance Executive with Jardine Fleming, michael.hk.tse@jfleming.com.

Web front ends may have made more data accessible and the actual querying easier, but perennial problems with database querying remain—the difficulties associated with understanding data from textual or graphical representations and mapping the meaning of a query into the interface language. This research investigates the query development difficulties associated with varying semantic distance, i.e., the distance between the information users want and the expression in the query interface language that will produce that information. The distance is manipulated in two ways: through the level of ambiguity of information requests and through the extent of congruence between constructs of the data representation and the interface language.

Potential benefits of this research include improved communication, e.g., between management and knowledge workers, and improved query support tools. Communication improvements might result from clearer statements of information requirements, more informed analysis of information requests, and greater use of divide and conquer strategies to ensure that information provided matches the information desired. Improved query support tools might entail enhancements to query front ends, e.g., to highlight potential ambiguities, and the creation of graphical data models and database views that facilitate typical queries by specific groups of end users.

This study extends prior research on end user query performance (e.g., Jih et al., 1989; Suh and Jenkins, 1992; Chan et al., 1993; Rho and March, 1997). Prior research typically compared different query languages or different forms of data representations. This research builds and empirically tests a theory to explain why ambiguity and incongruence adversely affect end user query performance. Methodological improvements over prior studies include a more realistic business setting, direct interaction between experimental participants and the computerized information system, complete interactive capture of this interaction, and a detailed analysis of the errors made by the experimental participants.

2. Ambiguity and Incongruence as Impediments to Query Development

In general, because humans are limited information processors, more effective problem solving results when less cognitive effort is required (Mitchell and Hunt, 1989). In Norman's (1986, 1988) model of user query performance, users exert cognitive effort to bridge the semantic distance between their query objectives and the way they must specify these objectives to the information system. Thus, the cognitive effort required to formulate successful queries increases with increasing semantic distance (Hutchins et al., 1985). Consistent with this model, empirical evidence indicates that information requests with shorter semantic distances require less cognitive effort and lead to better performance in terms of query correctness and query development time (Boehm-Davis, 1989; Suh and Jenkins, 1992; Rho and March, 1997). Two aspects that could affect semantic distance are the ambiguity of the information request and the congruence of the information request with the syntax of the query language and the data representation.

2.1 Model of the Query Process

Formulating a query requires transforming an information request into query components (Katzeff, 1990). It requires knowledge in three domains: knowledge of the information needed, knowledge of the database structure, and knowledge of the query language. Ineptness in one or more of these domains generates user errors (Ogden et al., 1986) that produce erroneous information and lead to inappropriate decisions (Davis, 1989; Podhorn et al., 1991).

Figure 1 illustrates a conceptualization of how end users formulate queries. First, users identify required constructs by filtering out unnecessary components from and adding missing or implied components to the statement of the information request. Second, they examine the data representation to identify the tables and attributes needed to obtain the required constructs. These two processes result in a mental model of the information request in terms of the available data. Third, end users translate their mental models into the query language syntax required to satisfy their mental model of the information requirements.

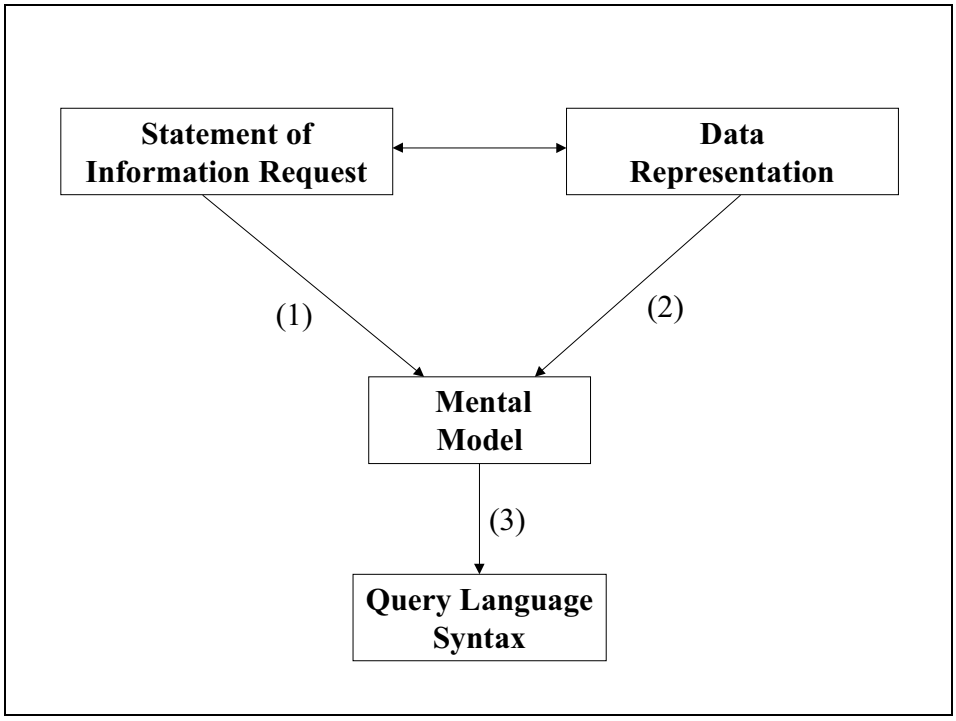

Figure 1. Model of End Users' Query Formulation Processes

Difficulties and errors result from discrepancies between the user's mental model and the other representations, i.e., the information request, the data representation, and the query language syntax. These discrepancies can be conceptualized as semantic distances. The semantic distance between query objectives and correct queries has two aspects: the information requirement distance and the data representation distance. The information requirement distance denotes the gap between the statement of the information request and the operations (operators) available in the query language (path (1) and (3) in Figure 1). The data representation distance signifies the gap between the data representation and the constructs (operands) required to formulate the appropriate query (path (2) and (3) in Figure 1). Greater cognitive effort is required when either semantic distance increases.

2.2 Ambiguity of Information Requests

One major problem in transforming a natural language request to an appropriate formulation in the query language, i.e., of traversing the information requirement distance, is the ambiguity of the information request (Parkison et al., 1977; Davidson et al., 1983; Sekine et al., 1992; Almuallim et al., 1997). Information requests that are ambiguous cause end users to be uncertain about the intent of the information request. This uncertainty, i.e., multiple possible desired outcomes of the information request, will lead to one-to-many mappings

from words in the natural language to the syntax (operators) in the query language. That is, a natural language information request may have multiple interpretations such that several query formulations may appear to be possible solutions.

End users can reduce ambiguity of information requests by translating them into a less ambiguous form, i.e., performing a stepwise refinement of the information request. Because of the smaller semantic distances (Norman, 1986; 1988), end users are likely to find it easier to resolve ambiguity in their own language than in a computer language. Consider, for example, the natural language information request for a transportation information system:

> Manager-English: Management wants to know the routes that each truck is not permitted to travel.

or the same request in language that is closer to the syntax of the query language:

> Pseudo-SQL: List all truck numbers and, where applicable, the route numbers where the truck violates height or weight constraints.

The manager-English version is the more ambiguous of the two statements because it could be mapped to several different pseudo-SQL statements (Parkison et al., 1977; Davidson and Kaplan, 1983; Sekine et al., 1992; Almuallim et al., 1997). To make the lexical transformations from manager-English to pseudo-SQL (Reisner, 1977), users would need to understand the query in a business context in which trucks are not permitted to travel on routes for which they violate height or weight constraints. The existence of multiple interpretations increases the information load (Campbell, 1988), increasing the cognitive effort users must expend to develop a correct query. For this query, the manager-English request does not explicitly state that permission to travel involves height or weight constraints or both. For this example, users of the manager-English request are more likely to make query errors of omission than users of the pseudo-SQL request.

Because the manager-English request acknowledges the existence of one or more categories of permissions that are required for a truck to travel a route, users must search their memory or consult some other source, e.g., a route table, to determine what the applicable constraints are. Then, if they want to compare their query with the manager-English request, users may create the list of constraints again. These search and compare processes take time, which would decrease a user s efficiency compared to formulating the query for the pseudo-SQL version of the information request.

Because query accuracy, efficiency, and confidence are often related, it is common to assess them jointly (Vessey, 1991). Users working with the more ambiguous manager-English request may identify multiple interpretations of the information requested and of the applicable constraints. Hence, they are likely to be less confident in their queries than users of the pseudo-SQL request. The combined correctness, efficiency, and confidence effects are the first hypothesis:

> H1: User query performance (correctness, efficiency, and confidence) will be inversely related to the ambiguity of information requests.

2.3 Incongruence

A second major problem in creating a query that correctly satisfies the information request is identifying the correct data elements, including their relationships and restrictions.

Successfully negotiating this data representation distance requires correctly mapping between real world constructs and their representation in the information system. The greater the incongruence between the real world and information system constructs, the greater the cognitive effort and the more likely the users' queries will contain errors. For example, queries requiring temporary views, missing relationships, or outer joins exhibit construct incongruence. Greater construct incongruence implies degraded Task-Technology-Fit (Goodhue, 1995), which is associated with worse performance (Vessey and Galletta, 1991; Goodhue and Thompson, 1995). As construct incongruence increases, the complexity of a query also increases, which increases the time and cognitive effort required to develop a correct query.

Users confronted with construct incongruity are likely to detect cues that alert them to inadequately developed queries. Each time users compare the information request with a newly developed query is an occasion on which they could detect mismatches between the information request and the query syntax and between the information request and the data representation. To the extent this comparison process identifies query errors, users get confirmation that they have not developed a usable query, which ought to undermine their confidence in the correctness of their queries. Thus, users' confidence in their queries should be inversely related to the degree of construct incongruence.

For example, consider two information requests. First,

> Pseudo-SQL: List the driver's name and the trip number for trips where drivers were assigned trucks that violate the height constraint for the prescribed route.

with its corresponding query (see the data structure in Figure 2),

```
select surname, firstname, trip.trip_no
from driver, trip_driver, trip, route, truck
where driver.license_no = trip_driver.license_no and
        trip_driver.trip_no = trip.trip_no and
        trip.route_no = route.route_no and
        trip.truck_no = truck.truck_no and
        height > height_constraint;
```

Second,

> Pseudo-SQL: List all truck numbers and, where applicable, the route numbers where the truck violates height constraints.

with its corresponding query (see the data structure in Figure 2),

```
create view truckviolation as
select truck_no, route_no
from route, truck
where height > height_constraint;

select truck.truck_no, route_no
from truck, truckviolation
where truck.truck_no = truckviolation.truck_no (+);
```

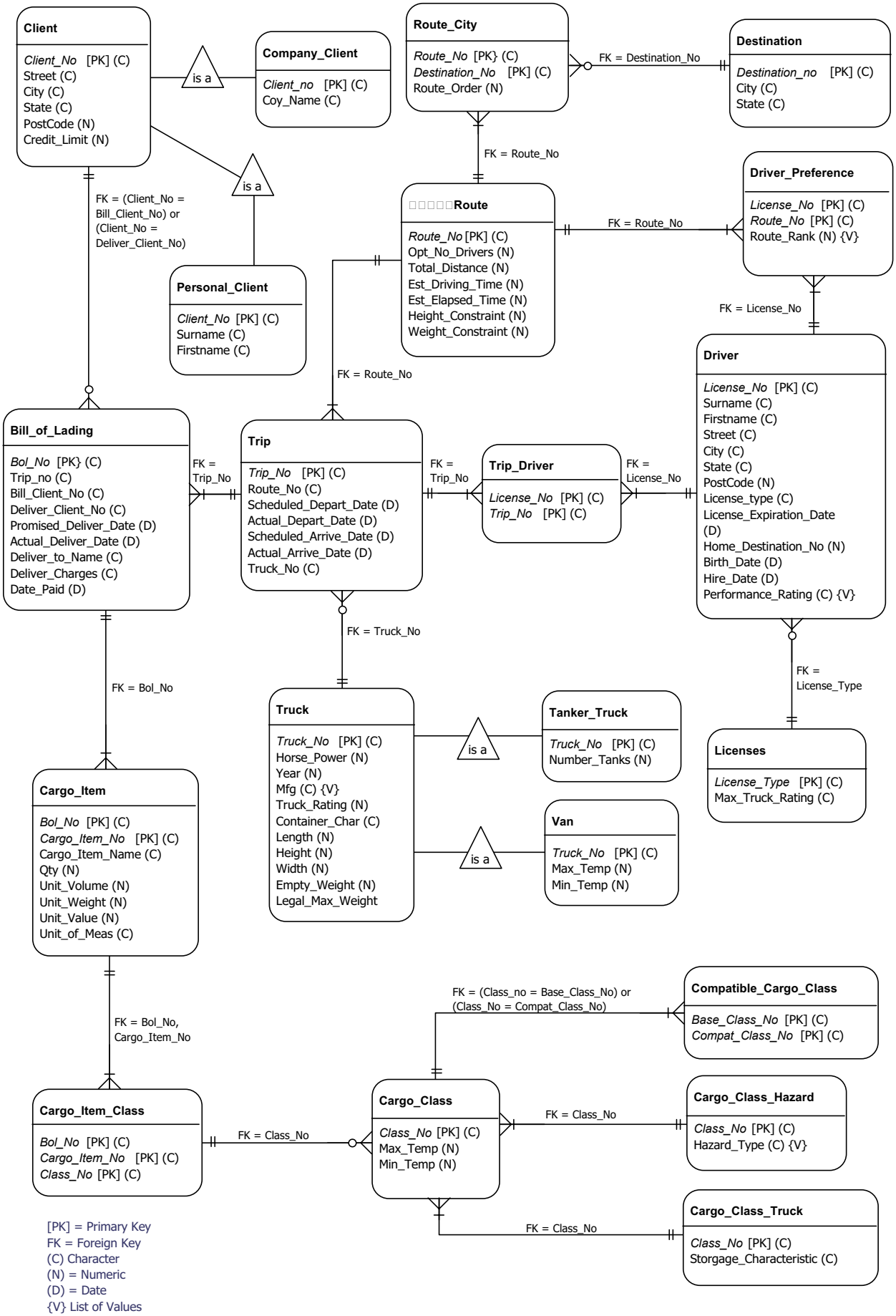

FIGURE 2: Entity-Relationship Diagram

For the first information request a relatively good task technology fit exists between the information request, the entity-relationship diagram (ERD), and the required query. For example, the first join of the query involves the tables *driver* and *trip_driver*. On the ERD, these tables are adjacent to each other and the foreign key necessary for a natural join of the two tables, *license_no*, is readily apparent. Hence, end users can relatively easily derive the required join, *driver.license_no = trip_driver.license_no*.

Conversely, the second information request does not exhibit a good task technology fit between the information request, the entity-relationship diagram (ERD), and the required query. The query contains a view, a Cartesian product, and an outer join, none of which are obvious from the ERD. End users must recognize that a view will facilitate obtaining the information requested. They must then determine what tables and attributes are required to construct the view *truckviolation* including the foreign key necessary to join the view with the *truck* table. This view is atypical in that it does not contain a join restriction but retains the full Cartesian product of the *route* and *truck* tables. Furthermore, to ensure that the query reports all trucks, a left outer join is required between *truck* and *truckviolation*. Hence, relative to the first information request, the second information request exhibits greater construct incongruence, requires more cognitive effort, and is likely to lead to poorer end user query performance.

The combined effects of accuracy, efficiency, and confidence as related to construct congruence comprise the second hypothesis:

> H2: User query performance (accuracy, efficiency, and confidence) will be inversely related to construct incongruence.

2.4 Query Complexity

The effects of task complexity have been studied extensively (Campbell, 1988; Wood, 1986). More complex tasks, e.g., more complex queries, place greater cognitive demands on persons undertaking the tasks and reduce their performance (Campbell, 1988). For computing tasks including queries, complexity is often measured using Halstead's (1977) difficulty measure (Jih et al., 1989). In this research, this measure is a function of the number of mental discriminations required to write a query. To confirm the expected relationship between performance as manifested by accuracy, efficiency, and confidence, the third hypothesis is:

> H3: User query performance (accuracy, efficiency, and confidence) will be inversely related to query complexity.

Figure 3 summarizes the hypothesized relationships of ambiguity, incongruence, and query complexity with performance.

3. Method

3.1 Design

The hypotheses were tested in a two-factor within-subjects laboratory experiment in which participants composed and executed queries in Oracle SQL. The ambiguity factor had two levels: information requests posed in pseudo-SQL (low ambiguity) or manager-English (high ambiguity). The congruence factor had two levels: information requests that were construct

congruent or construct incongruent. The traditional approach to testing query performance has been with pencil and paper (with or without an intermediary to return results) or simulated systems that did not reveal results to participants (e.g., Lochovsky and Tsichritzis, 1977; Greene et al., 1986, 1990; Jih et al., 1989; Davis, 1990; Suh and Jenkins, 1992; Chan et al., 1993). The testing approach in this experiment, however, incorporates a higher level of realism in that participants had the opportunity to work on their queries until they were satisfied with the query results from the database system.

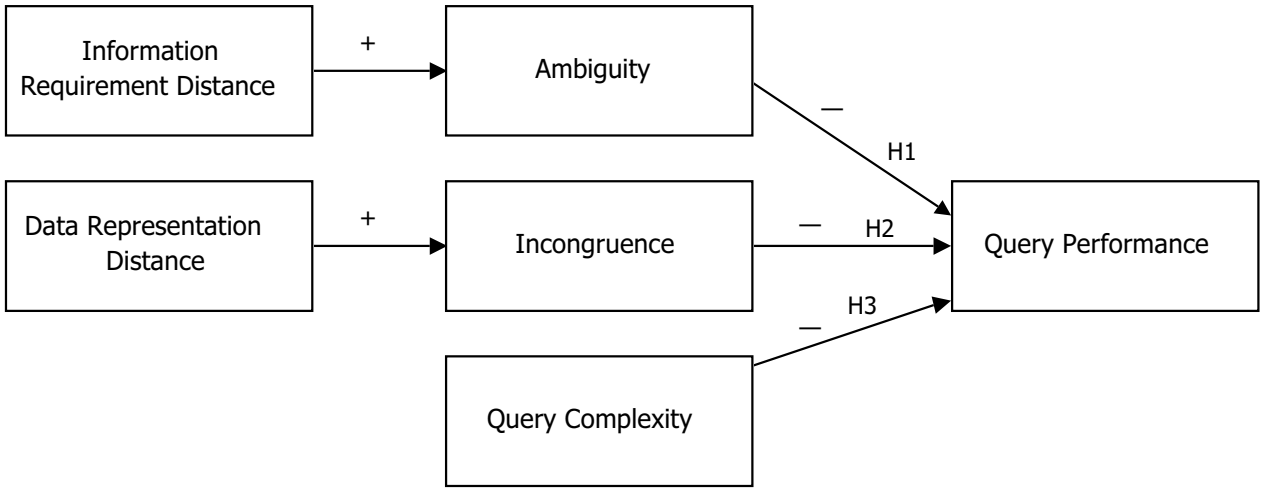

FIGURE 3: Query Performance Model

3.2 Participants

Participants were 23 graduate business students enrolled in an information systems course. Five percent of their course grade was based on performance in the experiment. Before the experiment, participants received training in interpreting entity relationship diagrams (ERDs) and formulating SQL queries. Because they were intelligent, had computing experience, and had training in query development, the participants were appropriate surrogates for casual users that are beginning to develop their own database queries. Because user characteristics are associated with query performance (Reisner, 1981), participants were divided into two equivalent groups based on their ranking determined from their information systems experience, education, and GPA. Table 1 summarizes demographic data for the two groups.

3.3. Procedure and Variables

At the beginning of the two-hour session, participants received instructions, an ERD representing the database (figure 2), and 16 information requests to satisfy with SQL queries. Odd-numbered requests had low construct congruence; even-numbered requests had high construct congruence. Each request had two versions: pseudo-SQL and manager-English. Each group received equal numbers of requests posed in each version, but for different requests, i.e., for each request, one group received the pseudo-SQL version and the other group, the manager- English version. After executing a query, participants were permitted to modify the query or move to the next query after indicating their level of confidence in the

correctness of the query. All computer interactions and time stamps were recorded in log files.

Table 1. Demographic Data

Variable	Group 1	Group 2
Gender		
Female	3	5
Male	9	6
Degree		
Graduate Information Systems	11	11
Graduate Business Administration	1	0
GPA (7-point scale)		
Mean	5.33	5.20
Standard deviation	1.0112	0.9970

Without knowledge of the identity of the participant or the version of the request the participant received, two researchers independently coded the accuracy of each query (based on the last query attempt) by reference to standard query solutions and resolved differences. Accuracy was assessed in two ways: in terms of the number of macro errors (errors involving row, column, and aggregation errors) and micro errors (the count of the minimum number of changes required to transform an actual query into a correct query). Appendix B contains an example of the accuracy coding for one question. Efficiency was measured as the total time spent on a query and as the number of query attempts. Confidence was measured as participants' self report after each query.

A complexity measure for each query was calculated as Halstead's (1977) difficulty measure as illustrated in Appendix C, and the information request pairs were arranged so that the difficulty of the congruent request of each pair was always equal to or greater than the difficulty of the incongruent request. The query performance data were analyzed in separate analysis of covariance (ANCOVA) models for macro performance, micro performance, elapsed time, number of query attempts, and confidence. In each case, performance was analyzed as a function of ambiguity (coded 0-1) as a nested component by request, incongruence (coded 0-1), and complexity (covariate). Because of the two hour time constraint, few participants attempted information requests eleven to sixteen. All statistical analyses are based on the 202 responses to information requests one to ten. The information requests and model SQL queries are shown in the appendix.

4. Results

4.1 Summary Statistics

Table 2 contains a summary of the experimental results. It verifies that the average complexity of the congruent questions is higher than that of the incongruent questions, i.e., that the information requests are biased against finding the hypothesized inverse relationship between incongruence and performance. Average performance for accuracy (micro and macro errors) show noticeable degradation when information requests are more ambiguous or incongruent. Average efficiency (time and attempts) and average confidence measures, however, show only slight degradation as ambiguity and incongruence increase.

Table 2. Summary of Experimental Results

	Mean / (Standard Deviation) by Condition			
	N = 50	N = 58	N = 62	N = 54
Measurement	**Congruent / Unambiguous**	**Congruent / Ambiguous**	**Incongruent / Unambiguous**	**Incongruent / Ambiguous**
Comp	16.2950 (8.7427)	16.0152 (8.2891)	12.0842 (6.2541)	12.3446 (6.2797)
Micro	2.1800 (5.2980)	6.6731 (10.9754)	17.3600 (17.5415)	20.4800 (20.2154)
Macro	0.3000 (0.6145)	0.8077 (1.1209)	0.7400 (0.5272)	1.1000 (0.7354)
Time	8.9000 (5.0679)	9.2692 (6.8431)	10.1400 (7.5404)	10.0200 (6.3132)
Attem	3.0200 (2.5674)	4.1346 (3.5260)	3.9800 (3.4905)	4.6800 (3.5135)
Conf	5.8600 (1.3704)	5.7692 (1.5031)	5.2400 (1.9332)	5.1400 (2.0204)

Figures 4 and 6 display the types of micro errors made by the participants for the different levels of ambiguity and incongruence, respectively. Figures 5 and 7 present micro errors by clauses for the different levels of ambiguity and incongruence.

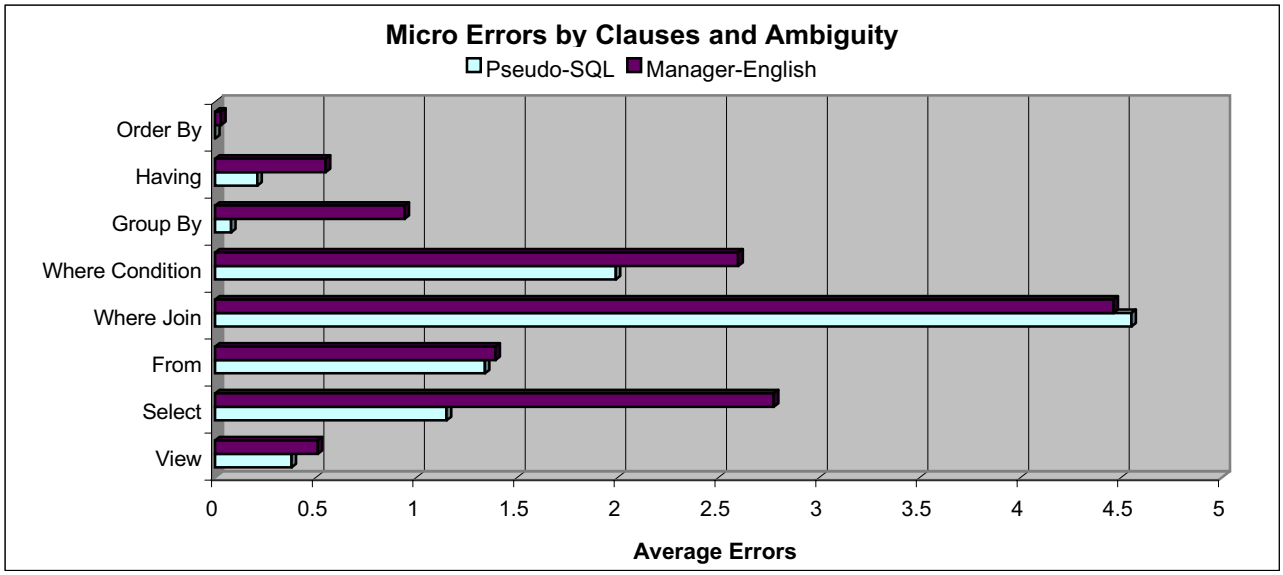

Figure 4. Types of Micro Errors (Average) and Ambiguity

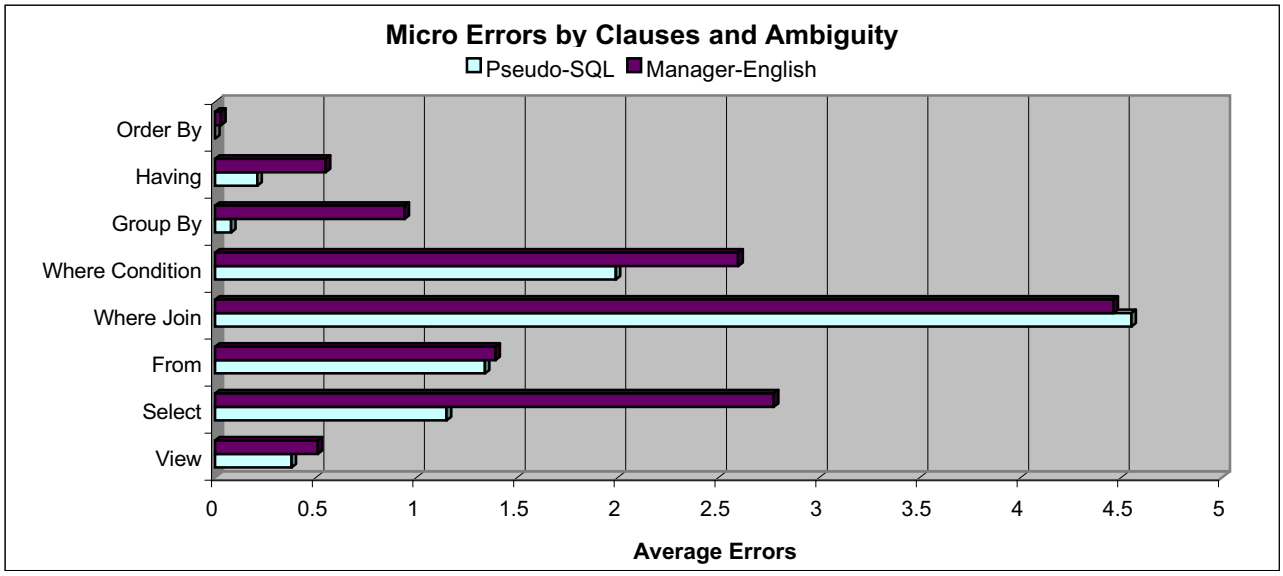

Figure 5. Micro Errors by Clauses (Average) and Ambiguity

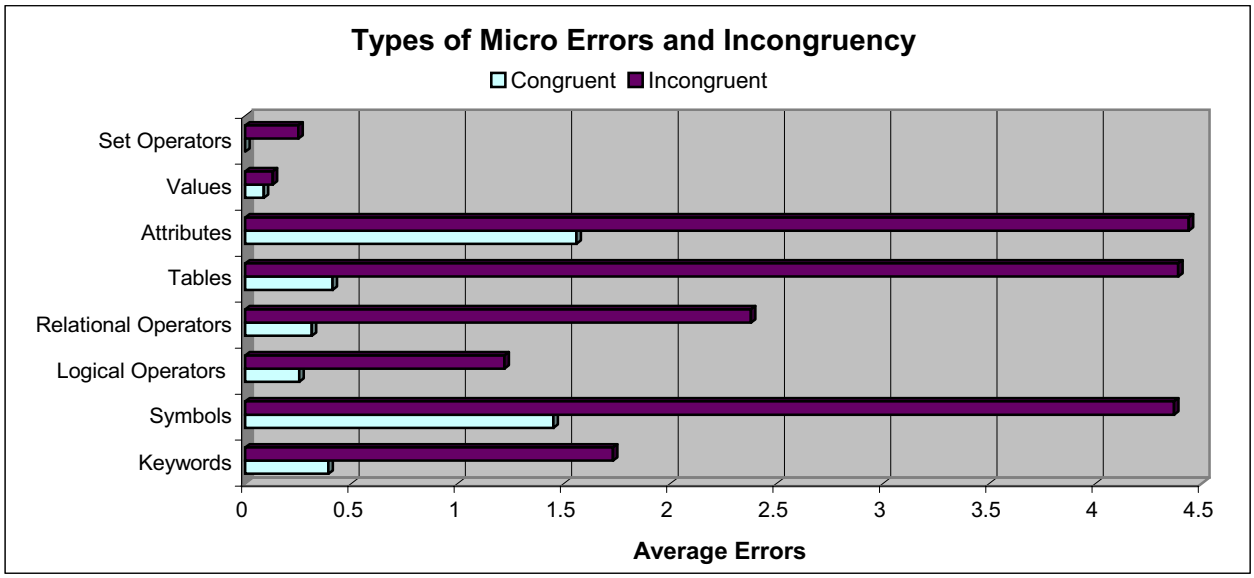

Figure 6. Types of Micro Errors (Average) and Incongruency

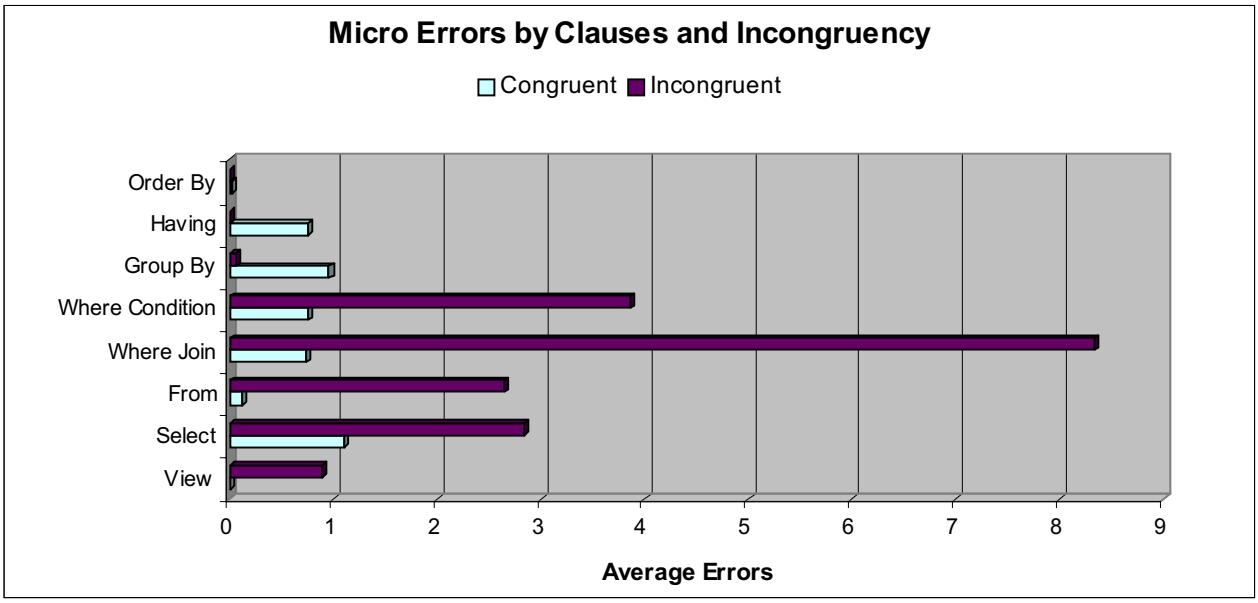

Figure 7. Micro Errors by Clauses (Average) and Incongruency

Figure 4 shows that, except for Values and Set Operators, information requests formulated in manager-English, on average, resulted in more micro errors for each category. The difference in accuracy between manager-English and pseudo-SQL is especially pronounced for Symbol and Attribute errors. Figure 5 reveals that information requests formulated in manager-English, on average, resulted in more micro errors for seven of eight different classes of SQL clauses. The difference between manager-English and pseudo-SQL is most evident for Select errors. The higher average number of Attribute and Select errors for manager-English formulations indicate that participants experienced difficulty in identifying what information was required from information requests that were more ambiguous. That is, the experimental participants experienced more difficulty identifying the correct columns than recognizing the appropriate row restrictions.

Figure 6 illustrates the effects of incongruence on types of micro errors. The negative effects of incongruence on participants' performance was especially pronounced for errors related to Attributes, Tables, Relational Operators, Symbols, and Keywords. Figure 7 shows that queries with greater construct incongruence resulted in more errors for six of eight classes of SQL clauses. Performance differences were most evident for errors in the Where Condition, Where Join, From, and Select clauses.

4.2 Tests of Hypotheses

Regression results appear in Table 3. For H1, that accuracy, efficiency, and confidence are inversely related to the ambiguity of information requests, the results support the hypothesis for accuracy as measured by micro and macro errors and for efficiency as measured by the number of attempts. Confidence was not significantly associated with ambiguity.

Table 3. ANCOVA Results for Performance Components

Source (n = 202)	df	Mean Square	F Value	p-value	Parameter Estimate	Standard Error of Estimate	R^2
Model: Accuracy Micro errors	12	2232.0830	15.55	0.0001			0.4967
Error	189	143.5789					
Intercept				0.0210	-0.3043	0.1308	
Complexity	1	1812.7555	12.63	0.0005	0.5658	0.1592	
Incongruence	1	7151.2602	49.81	0.0001	17.5626	2.4885	
Ambiguity	10	874.6809	6.09	0.0001	*	*	
Model: Accuracy Macro errors	12	7.5652	29.60	0.0001			0.6527
Error	189	.25562					
Intercept				0.0210	-0.3043	0.1308	
Complexity	1	7.7867	30.46	0.0001	0.0371	0.0067	
Incongruence	1	8.2400	32.24	0.0001	0.5962	0.1050	
Ambiguity	10	4.3835	17.15	0.0001	*	*	
Model: Efficiency Time	12	70.5968	1.76	0.0580			0.1004
Error	189	40.1697					
Intercept				0.0003	6.0487	1.6393	
Complexity	1	173.3585	4.32	0.0391	0.1750	0.0842	
Incongruence	1	90.6014	2.26	0.1348	1.9768	1.3163	
Ambiguity	10	46.0172	1.15	0.3305	*	*	
Model: Efficiency Number of attempts	12	42.7901	4.71	0.0001			0.2302
Error	189	9.0853					
Intercept				0.3762	0.6915	0.7796	
Complexity	1	115.6110	12.73	0.0005	0.1429	0.0401	
Incongruence	1	56.5460	6.22	0.0135	1.5617	0.6260	
Ambiguity	10	22.3305	2.46	0.0088	*	*	
Model: Confidence	12	7.6835	2.80	0.0015			0.1510
Error	189	2.7423					
Intercept				0.0001	7.0094	0.4283	
Complexity	1	28.1709	10.27	0.0016	-0.0705	0.0220	
Incongruence	1	19.4967	7.11	0.0083	-0.9170	0.3439	
Ambiguity	10	1.8109	0.66	0.7601	*	*	

* Ambiguity parameter estimates by information request are available from the authors.

For H2, that accuracy, efficiency, and confidence are inversely related to construct congruence, the results support the hypothesis for accuracy as assessed by micro and macro errors, for efficiency as assessed by the number of attempts, and for confidence. Even though efficiency, as measured by the number of attempts, was significant for both hypotheses, elapsed time was not significant for either one. This suggests that elapsed time may be driven by factors that are independent of a user's ability to formulate correct queries. As expected, H3, that query performance will be inversely related to query complexity, was supported by all measures of accuracy, efficiency, and confidence.

5. Discussion

The results of this study support the idea that the interaction among information requests, the query language, and the data representation affect individuals' ability to formulate correct

queries. Specifically, ambiguity in information requests adversely affects accuracy and efficiency. Incongruence among the information request, the query syntax, and the data representation adversely affects accuracy, efficiency, and confidence. Because these effects are in addition to the complexity effect, it may be appropriate to include ambiguity and incongruence as well as complexity in future research.

5.1 Implications Arising from Ambiguity Results

The results for ambiguity suggest that organizations might elicit better query development from casual users if users were sensitized to the nature of the kind of ambiguities that could arise in their business contexts and were trained to translate natural language queries into pseudo-SQL that could be examined for precision before the queries were developed.

From Figure 4, ambiguity is associated with errors in attributes, keywords, and, to a lesser extent, relational operators and tables. From Figure 5, ambiguity is associated with errors in select, where condition, group by, and having clauses. Using currently available tools and techniques, organizations can take a number of steps to reduces these ambiguity-induced errors. The user responsible for formulating a query can focus on more clearly identifying the attributes and columns needed to satisfy the request and on the conditions the data must satisfy. Users can work on improving their communications with the people making the information requests and following a more structured stepwise refinement process when converting the information requests to SQL queries. For example, users could develop pseudo-SQL representations of the information request and discuss that intermediate formulation with the information requestor.

Database owners can improve the data dictionary to enhance definitions and descriptions of both tables and attributes within the tables. Organizations can train users responsible for formulating the query to be more conscientious and insightful when developing queries. As a simple example, information requestors typically want to see the values of the attributes used in the restrictions even if they do not explicitly identify them when stating the attributes they want. Organizations can also cross train their employees. That is, given the importance to most organizations of exploiting their information system resources, information requestors need to improve their understanding of the data stored in their organization's information systems and information providers (persons responsible for formulating the queries) need to improve their understanding of the organization.

Specification languages such as VDM, Z, and B can help reduce ambiguity between information system analysts and programmers. To information requestors, these languages are likely to be even less understandable than query languages such as SQL and QBE. Future research could develop and test a specification language to facilitate clearer communications between information requestors and information providers.

5.2 Implications Arising from Incongruence Results

The results for incongruence suggest that better query development might ensue if semantic distances could be reduced by giving users data representations and database views that maximize construct congruence for the kinds of queries in their domains.

From Figures 6 and 7, incongruence is associated with substantial increases in errors for almost all types and almost all clauses. The typical way of viewing the data structure and its relationships, e.g., via an ERD, creates and reinforces a powerful mental model of an

information system. This mental model acts as an anchor that can inhibit an information provider from formulating a query that correctly satisfies the information request. In addition to the improvements recommended to reduce problems arising from ambiguity, organizations could develop or encourage their database management system provider to develop an adaptive data structure interface, e.g., an adaptive ERD interface.

Rather than displaying an ERD with the typical relationships and foreign keys, an adaptive interface could begin by listing the tables that comprise an information system. Individual tables could be expanded to show their attributes or designated as part of the subsystem necessary to satisfy the information request. After selecting the tables deemed relevant to the information request, users could then specify the foreign keys between those tables. Assuming that the underlying database management system provides domain support, the adaptive interface could inform users of possible foreign keys and issue warnings about attempts to use foreign keys with incompatible domains. Obviously, formulating queries in any query interface would be facilitated by consistent naming practices, i.e., using the same name for the same attribute in different tables and avoiding using the same name for attributes that are actually different. Even better would be using a data dictionary that resolves such inconsistencies.

Future research could investigate the desirable characteristics of a knowledge-based interface that assists users in formulating queries. Such an interface might request a natural language formulation of the information request, help users build their queries, warn users of likely errors, check the queries against natural language formulations for completeness, and suggest ways to enhance queries.

References

Almuallim, H., Akiba, Y., Yamazaki, T., and Kaneda, S., "Learning Verb Translation Rules from Ambiguous Examples and a Large Semantic Hierarchy," *Computational Learning Theory and Natural Learning Systems* (4), MIT Press, 1997, pp. 323-336.

Boehm-Davis, D. A., Holt, R. W., Koll M., Yastrop, G., and Peters, R., "Effects of Different Data Base Formats on Information Retrieval," *Human Factors* (31:5), pp. 579-592.

Campbell, D. J., "Task Complexity: A Review and Analysis," *Academy of Management Review* (13:1), 1988, pp. 40-52.

Chan, H.C., Wei, K.K., and Siau, K.L., "User-Database Interface: The Effect of Abstraction Levels on Query Performance," *MIS Quarterly* (17:4), Dec 1993, pp. 441-464.

Davidson, J., and Kaplan, S.J., "Natural Language Access to Data Bases: Interpreting Update Requests," *American Journal of Computational Linguistics* (9:2), Apr.-Jun. 1983, pp. 57-68.

Davis, J.S., "Experimental Investigation of the Utility of Data Structure and E-R Diagrams in Database Query," *International Journal of Man-Machine Studies* (32), 1990, pp. 449-459.

Deck, S., "Are You Weary of Warehouses?" *Computerworld*, May 31, 1999, 61, www.computerworld.com/home/print.nsf/all/990531AADE.

Gantz, J., "The New World of Enterprise Reporting Is Here," *Computerworld*, Feb. 1, 1999, 34, www.computerworld.com/home/print.nsf/all/9902018D36.

Goodhue, D.L., "Understanding User Evaluation of Information Systems," *Management Science* (41:12), Dec 1995, pp. 1827-1844.

Goodhue, D.L., and Thompson, R.L., "Task-Technology Fit and Individual Performance," *MIS Quarterly* (19:2), 1995, Jun., pp. 213-236.

Greene, S.L., Gomez, L.M., and Devlin, S.J., "A Cognition Analysis of Database Query Production," *Proceedings of the Human Factors Society 30th Annual Meeting* (2), 1986, pp. 9-13.

Greene, S.L., Devlin, S.J., Cannata, P.E., and Gomez, L.M., "No Ifs, ANDs, or Ors: A Study of Database Querying," *International Journal of Man Machine Studies* (32:3), March 1990, pp. 303-326.

Halstead, M. H., "Elements of Software Science," *Elsevier North Holland,* 1977.

Horwitt, E., "Webifying the Mainframe," *Computerworld*, Jan. 25, 1999, 76-79, www.computerworld.com/home/print.nsf/all/9901258B66.

Hutchins, E., Hollan, J.D., and Norman, D.A. (1985), "Direct Manipulation Interfaces," in D.A. Norman and S.W. Draper (Eds.), *User Centered System Design: New Perspectives on Human-Computer Interaction,* Hillsdale, NJ, Erlbaum Associates.

Jih, J.W.K., Braford, D.A., Snyder, C.A., and Thompson, N.G.A., "The Effects of Relational and Entity-Relationship data Models on Query Performance of End-Users," *International Journal of Man-Machine Studies* (31:3), 1989, pp. 257-267.

Lochovsky, F.H. and Tsichritzis, D.C., "User Performance Considerations in DBMS Selection," *Proceedings of ACM SIGMOD*, 1977, pp. 128-134.

Mitchell, D.B., and Hunt, R.R., "How Much Effort Should Be Devoted to Memory?" *Memory and Cognition* (17:3), 1989, pp. 337-348.

Norman, D.A., "Cognitive Engineering," in D.A. Norman and S.W. Draper (eds.), *User Centered System Design,* Lawrence Erlbaum Associates, Hillsdale, NJ, 1986, pp. 31-61.

Norman, D.A., The Psychology of Everyday Things, New York, NY: Basic Books, 1988.

Parkison, R.C., Colby, K.M., and Faught, W.S., "Conversational Language Comprehension Using Integrated Pattern-Matching and Parsing," *Artificial Intelligence* (9:2), Oct. 1977, pp. 111-134.

Reisner, P., "Use of Psychological Experimentation as an Aid to Development of a Query Language," *IEEE Transactions on Software Engineering* (SE3:3), 1977, pp. 218-229.

Reisner, P., "Human Factors Studies of Database Query Languages: A Survey and Assessment," *Computing Surveys* (13:1), Mar. 1981, pp. 13-31.

Rho, S., and March, S.T., "An Analysis of Semantic Overload in Database Access Systems Using Multi-table Query Formulation," *Journal of Database Management* (8:2), Spring 1997, pp. 3-14.

Sekine, S., Carroll, J.J., Ananiadou, S., and Tsujii, J., "Automatic Learning for Semantic Collocation," *Third Conference on Applied Natural Language Processing,* 1992, pp. 104-110.

Suh, K.S., and Jenkins, A.M., "A Comparison of Linear Keyword and Restricted Natural Language Database Interfaces for Novice Users," *Information Systems Research* (3:3), 1992, pp. 252-272.

Vessey, I., "Cognitive Fit: A Theory-Based Analysis of the Graphs Versus Tables Literature," *Decision Sciences* (22), Spring 1991, pp. 219-241.

Vessey, I., and Galletta, D., "Cognitive Fit: An Empirical Study of Information Acquisition," *Information Systems Research* (2:1), 1991, pp. 63-84.

Appendix A: Information Requests and Model SQL Queries

1. Congruent; **Halstead Difficulty = 6.06**

1m. Management wants the names of personal clients and their credit limits.

1p. List client names and credit limits where the clients are personal clients.
surname, firstname, credit_limit

```
select surname, firstname, credit_limit
from client, personal_client
where client.client_no = personal_client.client_no;
```

2. Incongruent (missing relationship); **Halstead Difficulty = 5.25**

2m. Management wants to know the names of drivers with the same post code as clients.

2p. List names of drivers where the post code of the driver is the same as the post code of a client.
surname, firstname

```
select surname, firstname
from driver, client
where driver.postcode = client.postcode;
```

3. Congruent (aggregation); **Halstead Difficulty = 10.25**

3m. Management wants the names of bill-to clients who are persons and their total delivery charges. Clients with larger charges should be listed first.

3p. List the client numbers of bill-to clients who are persons and their total delivery charges with clients having larger charges listed first.
surname, firstname, sum(deliver_charges)

```
select surname, firstname, sum(deliver_charges)
```

```
from personal_client, bill_of_lading
where  personal_client.client_no = bill_of_lading.bill_client_no
group by surname, firstname
order by 3 desc;
```

4. **Incongruent** (set operation, cartesian product, or, negation, missing relationship); **Halstead Difficulty = 8.27**

4m. Management wants to know the trucks that can travel all routes.

4p. List the truck numbers of trucks where the truck does not exceed either the height or weight constraints for any route.

truck_no

```
select truck_no
from truck
where truck_no not in
                        (select truck_no
                          from route, truck
                          where (height > height_constraint or
                          empty_weight > weight_constraint));
```

5. **Congruent** (aggregation); **Halstead Difficulty = 19.35**

5m. Management wants to know the names of clients who are persons that are over their credit limit.

5p. List the names of clients who are persons where the sum of their unpaid delivery charges exceeds their credit limit.

surname, firstname, credit_limit, sum(deliver_charges)

```
select surname, firstname, credit_limit, sum(deliver_charges)
from personal_client, bill_of_lading, client
where client.client_no = bill_of_lading.bill_client_no and
        client.client_no = personal_client.client_no and
        date_paid is null
group by surname, firstname, credit_limit
having sum(deliver_charges) > credit_limit;
```

6. **Incongruent** (missing relationships, cartesian product, outer join); **Halstead Difficulty = 16.36**

6m. Management wants to know *all* trucks and the routes that each truck is not permitted to travel.

6p. List *all* truck numbers and where applicable the route numbers where the truck violates height or weight constraints.

truck_no, route_no

```
create view truckviolation as
```

```
select truck_no, route_no
from route, truck
where (height > height_constraint or
        empty_weight > weight_constraint);

select truck.truck_no, route_no
from truck, truckviolation
where truck.truck_no = truckviolation.truck_no (+);
```

7. **Congruent**; **Halstead Difficulty = 21.13**

7m. Management wants to know the names of drivers and the trip numbers where the truck exceeded the height constraint.

7p. List the driver's name and the trip numbers where the truck's height exceeded the height constraint for that route.

surname, firstname, trip_no

```
select surname, firstname, trip.trip_no
from route, trip, truck, trip_driver, driver
where  trip.trip_no = trip_driver.trip_no and
        trip_driver.license_no = driver.license_no and
        trip.truck_no = truck.truck_no and
        trip.route_no = route.route_no and
        height > height_constraint;
```

8. **Incongruent** (outer join, missing relationship, aggregation); **Halstead Difficulty = 16.88**

8m. Management wants to know *all* the destinations (and their city and state) and the number of clients at each of the destinations.

8p. For *all* destinations (same city and state), list the destination number, city, state, and total number of clients at that city and state.

destination_no, city, state, count(client_no)

```
select destination_no, destination.city, destination.state, count(client_no)
from destination, client
where  destination.city = client.city (+) and
        destination.state = client.state (+)
group by destination_no, destination.city, destination.state;
```

9. **Congruent** (aggregation, aggregation restriction); **Halstead Difficulty = 30.37**

9m. Management wants to know the trip number, truck number, and maximum and minimum weight of the trucks for the trips where the trucks weighed more than their legal maximum weight.

9p. List the trip number, truck number, and the maximum and minimum weight of trucks for the trips where the cargo weight plus the truck's empty weight exceeds its legal maximum weight.

trip_no, truck_no, legal_max_weight, empty_weight,
sum(unit_weight * qty) + empty_weight

```
select trip.trip_no, trip.truck_no, legal_max_weight, empty_weight,
        sum(unit_weight * qty) + empty_weight
from bill_of_lading, trip, truck, cargo_item
where  trip.truck_no = truck.truck_no and
        trip.trip_no = bill_of_lading.trip_no and
        bill_of_lading.bol_no = cargo_item.bol_no
group by trip.trip_no, trip.truck_no, legal_max_weight, empty_weight
having sum(unit_weight * qty) + empty_weight > legal_max_weight;
```

10. **Incongruent** (view, missing relationship, cartesian product, outer join); **Halstead Difficulty = 23.19**

10m. Management wants to know *all* drivers and the trucks that each driver is licensed to drive as of 8 Sept 1998.

10p. Assume today is 8 Sept 1998. For *all* drivers, list license number, driver name, and truck number where the driver's license meets or exceeds the requirements for the truck and the license has not expired.

license_no, surname, firstname, truck_no

```
create view legaltruck as
select license_no, truck_no
from driver, truck, licenses
where  driver.license_type = licenses.license_type and
        max_truck_rating >= truck_rating and
        license_expiration_date > '8-Sept-1998';

select driver.license_no, surname, firstname, truck_no
from driver, legaltruck
where driver.license_no = legaltruck.license_no (+);
```

Appendix B: Example of Error Marking for Accuracy in Queries

Question 3 is used to demonstrate how the accuracy performance measure was calculated.

Accuracy:
Question 3: Management wants the names of bill-to clients who are persons and their total delivery charges. Clients with larger charges should be listed first.

Correct solution:

```
select surname, firstname, sum(deliver_charges)
from personal_client, bill_of_lading
where bill_of_lading.bill_client_no = personal_client.client_no
```

```
group by surname, firstname
order by 3 desc;
```

Sample actual solution:

```
select surname, sum(deliver_charges)
from bill_of_lading, personal_client
where bill_of_lading.deliver_client_no = personal_client.client_no
group by surname;
```

The semantic errors in the response are:

- Missing the attribute *firstname* in the *select* clause: this error is recorded at the micro level as one "*Attributes Select*" error and one "*Symbols Select*" error for the extra "," that is required. At the macro level, this error is recorded as a "*Columns*" error.
- Incorrect join the between *bill_of_lading* and the *personal_client* tables: at the micro level, to make the query correct requires the following elements to be deleted and inserted: delete *.deliver_client_no*, which is one "*Symbols Where Join*" error and one "*Attributes Where Join*" error; and insert *.bill_client_no*, which is one "*Symbols Where Join*" error and one "*Attributes Where Join*" error. At the macro level, this is error is recorded as a "*Rows*" error because an incorrect join affects the rows that are retrieved by the query.
- Missing the attribute firstname in the group by clause: this error is recorded at the micro level as one "*Attribute Group by*" error and one "*Symbols Group by*" error for the extra "," that is required. At the macro level, this in not an macro error because this error originated from the error in the *select* clause.
- Missing the order by statement: at the micro level, to make the query correct requires the following elements to be inserted: *order by 3 desc*, which is two "*Keyboards Order by*" errors and one "*Attribute Order by*" error. At the macro level, this error is recorded as a "*Rows*" error because the order of the data retrieved by the query is affected.

Performance, in terms of accuracy, was measured by the total number of micro errors and the total number of macro errors. This example had 11 micro errors and 2 macro errors. The following error counting sheet contains the micro errors and the macro errors for this example query.

Name	Question Number	Attempts

MICRO ERRORS

Keyboards

View	Select	From	Where Join	Where Cond	Group by	Having	Order by
	1						2

Symbols

View	Select	From	Where Join	Where Cond	Group by	Having	Order by
	1		2		1		

Logical Operators

View	Select	From	Where Join	Where Cond	Group by	Having	Order by

Relational Operators

View	Select	From	Where Join	Where Cond	Group by	Having	Order by

Tables

View	Select	From	Where Join	Where Cond	Group by	Having	Order by

Attributes

View	Select	From	Where Join	Where Cond	Group by	Having	Order by
			2		1		1

Values

View	Select	From	Where Join	Where Cond	Group by	Having	Order by

Set Operators

Where	Union	Intersect	Minus

MACRO ERRORS

Columns	Rows	Aggregation
1	1	

Appendix C: Example Calculation of Halstead Complexity measure

Halstead's (1977) difficulty measure was chosen as the complexity measure for this research. The formula is:

$$D = V/V^* = (N \log^2 n)/(n^* \log^2 n^*)$$

where:

N = program length = $N^1 + N^2$
N^1 = total operators
N^2 = total operands
n = vocabulary size = $n^1 + n^2$
n^1 = unique operator count
n^2 = unique operand count
n^* = potential (minimum) vocabulary = $n^{1*} + n^{2*}$
n^{1*} = potential (minimum) operator count

n^{2*} = potential (minimum) operand count

For any SQL query, n^*, the potential (minimum) vocabulary is always 5 because the minimum query is:

```
Select * From table;
```

The solution to question 3 is:

```
Select surname, firstname, sum(deliver_charges)
From personal_client, bill_of_lading
Where personel_client.client_no = bill_of_lading.client_no
Group by surname, firstname
Order by 3 desc;
```

Operators	Count	Operands	Count
Select	1	surname	2
, (comma)	4	firstname	2
sum	1	deliver_charges	1
( )	1	personal_client	2
From	1	bill_of_lading	2
Where	1	client_no	2
. (period)	2	3	1
= (equal sign)	1		
Group by	1		
Order by	1		
desc	1		
; (semicolon)	1		

$n^1 = 12$ $n^2 = 7$

$N^1 = 16$ $N^2 = 12$

$D = (28 \log^2 19) / (5 \log^2 5) = 10.25$