APPLYING CALIBRATION TO IMPROVE UNCERTAINTY ASSESSMENT

A Thesis

by

MARK EDWARD FONDREN II

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Duane McVay
Co-Chair of Committee,	Eduardo Gildin
Committee Member,	Yuefeng Sun
Head of Department,	Dan Hill

August 2013

Major Subject: Petroleum Engineering

ABSTRACT

Uncertainty has a large effect on projects in the oil and gas industry, because most aspects of project evaluation rely on estimates. Industry routinely underestimates uncertainty, often significantly. The tendency to underestimate uncertainty is nearly universal. The cost associated with underestimating uncertainty, or overconfidence, can be substantial. Studies have shown that moderate overconfidence and optimism can result in expected portfolio disappointment of more than 30%. It has been shown that uncertainty can be assessed more reliably through look-backs and calibration, i.e., comparing actual results to probabilistic predictions over time. While many recognize the importance of look-backs, calibration is seldom practiced in industry. I believe a primary reason for this is lack of systematic processes and software for calibration.

The primary development of my research is a database application that provides a way to track probabilistic estimates and their reliability over time. The Brier score and its components, mainly calibration, are used for evaluating reliability. The system is general in the types of estimates and forecasts that it can monitor, including production, reserves, time, costs, and even quarterly earnings. Forecasts may be assessed visually, using calibration charts, and quantitatively, using the Brier score. The calibration information can be used to modify probabilistic estimation and forecasting processes as needed to be more reliable. Historical data may be used to externally adjust future forecasts so they are better calibrated. Three experiments with historical data sets of predicted vs. actual quantities, e.g., drilling costs and reserves, are presented and

demonstrate that external adjustment of probabilistic forecasts improve future estimates. Consistent application of this approach and database application over time should improve probabilistic forecasts, resulting in improved company and industry performance.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. McVay, for motivating me in my research to strive for excellence. Thank you to my committee members, Dr. Gildin, and Dr. Sun for serving on my committee.

My time as a graduate student at Texas A&M has been a rewarding experience, both academically and culturally. Thanks to my colleagues and friends from around the world, for providing me with knowledge and respect of other cultures while encouraging me to better understand petroleum engineering. Anton Padin, Brett Gilbert, Houda Hdadou, Juan Lacayo, Martin Saint-Félix, Matt Bell, Sergio Gonzales, Raul Gonzalez and others thanks!

Finally, thanks to my mother and father, Christine and Mark Fondren, for their love, encouragement, and financial support. I am grateful that they encouraged me to pursue my graduate education. Thank you to my two sisters Meagan and Michelle, for their love and support during my graduate studies, and being great roommates. Love you guys!

TABLE OF CONTENTS

Page

ABSTRACT ii

ACKNOWLEDGEMENTS iv

TABLE OF CONTENTS v

LIST OF FIGURES vi

LIST OF TABLES vii

1. INTRODUCTION AND BACKGROUND 1

1.1 Introduction 1
1.2 Literature Review 1
1.3 Objectives 8
1.4 Methodology 8

2. DATABASE MODEL 10

2.1 Relational Database 10
2.2 Database Structure 10
2.3 Scoring Rules and Queries 16
2.4 Importing and Exporting Data 21
2.5 Plotting Data 23

3. RESULTS 25

3.1 Introduction 25
3.2 Barnett Shale Drilling Cost Predictions 25
3.3 Barnett Shale Reserves and External Calibration 29
3.4 Student Estimates 34

4. CONCLUSIONS 38

5. FUTURE WORK 39

REFERENCES 41

LIST OF FIGURES

		Page
Fig. 1	Calibration plot	5
Fig. 2	Fields in the assessor table	13
Fig. 3	Fields in the question table	14
Fig. 4	Fields in the assessment table	14
Fig. 5	Index button in design view of database	14
Fig. 6	Form for running the Brier score and its components	20
Fig. 7	External data tab in Microsoft Access	22
Fig. 8	Template headers for importing into assessor table	22
Fig. 9	Template headers for importing into question table	22
Fig. 10	Template headers for importing into the assessment table	22
Fig. 11	Assessment form for manually inputting assessments	23
Fig. 12	Calibration plot in database	24
Fig. 13	Calibration plot of uncalibrated reserves estimates	30
Fig. 14	Calibration plot of externally calibrated reserves estimates	33
Fig. 15	Student estimates for the 2012 Texas A&M football season	36

LIST OF TABLES

Page

Table 1 Drilling cost prediction results 27

Table 2 Drilling cost predictions for 2011 sorted by quarter 28

Table 3 Calibration of Barnett shale reserves estimates using Jochen and Spivey model 30

Table 4 External calibration for Barnett reserves forecast using Jochen and Spivey model 32

Table 5 Calibration of 2012 Texas A&M football estimates 36

1. INTRODUCTION AND BACKGROUND

1.1 Introduction

The oil and gas industry is full of uncertainty. In addition to significant subsurface uncertainty and uncertainty in oil and gas prices, there are other risks, e.g., political, that contribute to uncertainty in oil and gas projects. The problem, as suggested by Capen (1976), is that the industry routinely underestimates uncertainty, often significantly. Underestimation of uncertainty (overconfidence) is almost a universal tendency. The cost associated with underestimating uncertainty can be substantial. According to McVay and Dossary (2012), moderate overconfidence and optimism can result in expected portfolio disappointment of more than 30%, and greater average disappointment has been experienced by the industry. Capen (1976) and other authors have pointed out that uncertainty can be assessed more reliably through look-backs and calibration, i.e., comparing actual results to probabilistic predictions over time. While many recognize the importance of look-backs, calibration is seldom practiced in the industry. We believe a primary reason for this is lack of a systematic process and lack of appreciation for the cost of underestimating uncertainty.

1.2 Literature Review

Capen (1976) demonstrated the tendency to underestimate uncertainty through several experiments. The first was a ten-question survey that required participants to provide 80% confidence intervals to general-knowledge questions. The actual

confidence intervals participants provided were too narrow, and actually only 32% on average. What Capen demonstrated is the tendency to be overconfident in assessment causes ranges to be too narrow. There were several other experiments conducted by Capen which yielded similar results, displaying underestimation of uncertainty. One explanation for the narrow estimate ranges is the failure to include all possible outcomes, specifically the ones that are unknown (Capen, 1976). Participants' ranges were too narrow because they are unable to envision all possibilities. One way to improve probability ranges, as Capen demonstrated, is to understand that ranges are generally too narrow, and to use prior knowledge of this fact to scale the ranges appropriately.

To quantify the cost of underestimating uncertainty, McVay and Dossary (2012) analyzed the effect overconfidence has on portfolio values. They found "for moderate amounts of overconfidence and optimism, expected disappointment was 30-35% of estimated NPV for industry portfolios and optimization cases…" (McVay and Dossary, 2012). The expected disappointment percentage is equal to the estimated portfolio value minus the realized portfolio value, divided by the estimated portfolio value. It was shown that reducing overconfidence should result in a more reliable portfolio NPV estimate and lower expected disappointment. While McVay and Dossary assessed the cost of underestimating uncertainty, they did not fully address how to better assess uncertainty.

The Brier score is a proper scoring rule that is commonly applied in other industries for assessing forecasts and was initially developed to assess weather forecasts (Brier, 1950). Lichtenstein and Fischhoff (1977) summarized the background for the

Brier score and its components. The Brier score ranges from 0 to 1, providing a way to rank probabilistic estimates. The Brier score is negatively oriented; a Brier score of 0 is a perfect score. The Brier score was intended to assess forecasts for events with a binary outcome. The initial application of this scoring rule was to assess the occurrence of a future event. The Brier score has also found application in assessing knowledge by assessing the proportion of correct responses (Bjorkman, 1992). The difference between the two types of assessments is detailed in a paper by Bjorkman (1992). In the first case, the Brier score is assessing an event and its corresponding uncertainty of occurrence; in the second case, the score is assessing the knowledge of a subject. The equations for both cases are the same, but perspective of the components in the Brier score change depending on what is being assessed.

There are three components to the Brier score - calibration, resolution, and knowledge. Calibration is a measure of how close an assessor's assigned probabilities match with the proportion of correct responses. Lichtenstein and Fischhoff (1977), consider the case where assessments are binary, and may be reduced to a situation of true or false. This will allow one to gauge if the probabilities assigned match the actual distribution. As stated in the paper by Lichtenstein and Fischhoff (1977), "The perfectly calibrated judge assigns probabilities so that, for all propositions assigned the same probability, the proportion true is equal to the probability assigned." The equation for calibration is defined as,

$$\text{calibration}=\frac{1}{N}\sum_{t=1}^{T} n_t(r_t-c_t)^2 \quad \text{.....(1)}$$

In Eq. 1, N is the total number of assessments, T is the total number of different response percentages, n_t is the number of assessments for the response percentage r_t, as stated r_t is the percent assigned to an assessment, and c_t is the percent correct. The percent correct is the proportion of times a binary assessment is correct for a specific response percentage r_t. In other words, the percent correct is the sum of the binary assessments, 1 or 0, divided by the number of assessments in the percentage category. It may be seen in Eq. 1 that when the forecasts response percentage, r_t, equals the percent correct c_t, the calibration is zero. The worst calibration score possible is equal to one. An additional way to test the calibration of an assessor is through calibration plots. Calibration plots are generated by plotting c_t vs. r_t . For a perfectly calibrated assessor the graph will display a linear correlation with unit slope. An example of a calibration plot is provided in Fig. 1. When an assessor is under confident, the response percentage is less than the percent correct resulting in a point above the unit slope line. Likewise when an assessor is overconfident, the percent assigned is greater than the percent correct, which results in a point below the unit slope line.

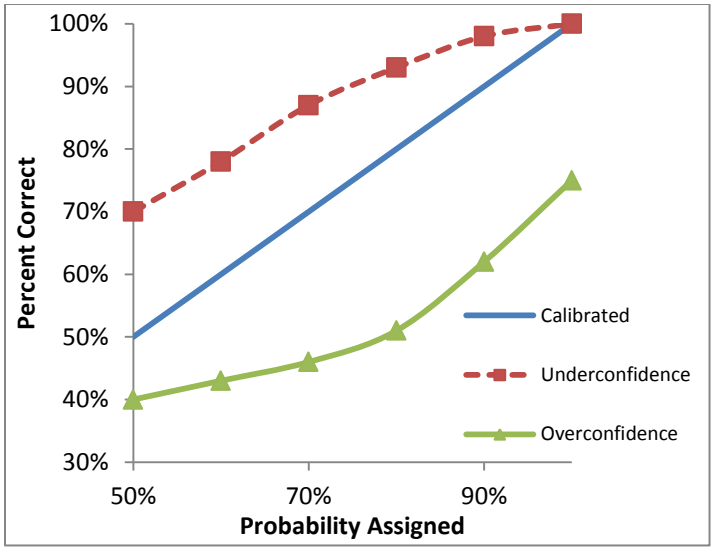

Fig. 1 Calibration plot

Resolution ranks the range of varying percentages that are used to describe the forecast. When the individual percent correct varies from the overall percent correct, the resolution term will be large. The resolution is proportional to the squared difference between the individual proportion correct and the total proportion correct. The equation for the resolution component is

$$\text{resolution} = \frac{1}{N}\sum_{t=1}^{T} n_t(c_t - c)^2 \quad \text{.....(2)}$$

The only new component in Eq. 2 that differs from Eq. 1 is c, the overall percent correct. The overall percent correct is the same as the percent correct, except the sum of the binary assessment is not segmented by response percentage. The overall percent correct, *c*, is the number of times the binary assessment is correct divided by the total number of responses. In a perfect forecast the resolution component equals the knowledge

component. Since the resolution is a negative component in the Brier score, the larger it is the better. The final component, knowledge, is related to the overall percent correct associated with a forecast. The knowledge component is defined in Eq. 3,

$$\text{knowledge} = c(1-c) \quad (3)$$

In Eq. 3, c is the overall percent correct, which is multiplied by its compliment. The knowledge score ranges from 0 to .25, and is maximized when the overall percent correct is 50%. Like calibration, knowledge is also negatively oriented, with a score of zero being perfect. According to Bjorkman (1992), a high knowledge score results from an assessor's lack of knowledge, guessing on all assessments. Viewing the knowledge component from the event perspective, the knowledge component is actually measuring the uncertainty associated with an event, not the knowledge of an assessor. An event that happens frequently or infrequently will have a low knowledge score. The equation for the Brier score containing the three components is,

$$\text{Brier score} = \text{calibration} - \text{resolution} + \text{knowledge} \quad (4)$$

The Brier score is just one proper scoring rule; there are others that may be applied to rank predictions. One other scoring rule worth mentioning is logarithmic. Bickel (2010) compared the logarithmic scoring rule with the spherical and quadratic. The Brier score is negatively oriented and is equivalent to 1 minus the quadratic score. Bickel (2010) showed that the logarithmic scoring rule depends primarily on the percent

assigned to the correct answer. Each scoring rule may provide a different perspective on the way that it ranks estimates.

Uncertainty assessment is an area in need of improvement in the petroleum industry. Bickel and Bratvold (2007) conducted a survey to determine the status of uncertainty quantification and decision making in industry. The paper addresses the need to move from merely quantifying uncertainty to using it to improve decisions. The survey was made up of 494 participants from SPE chapters and technical groups. In two separate questions participants identified the major obstacles in improving uncertainty quantification and decision making. The responses stated that uncertainty quantification was limited by a lack of time, and decision making was limited by a lack of management understanding. In an open-ended portion of the survey, participants were asked to provide the aspect in need of greatest improvement in quantifying uncertainty and decision making in their organizations. The two main responses were speed and consistency. Another paper that analyzes techniques to handle uncertainty in industry is Wolff (2010). Wolff concludes that in order to deal with uncertainty effectively in our industry, a common process should be developed that is easy to audit, which agrees with the findings of Bickel and Bratvold.

In conclusion, there is a need to promote and improve uncertainty assessment in the oil and gas industry. Capen (1976) showed that there is a universal tendency to underestimate uncertainty and McVay and Dossary (2012) showed that the cost associated with underestimating this uncertainty can be large. While McVay and Dossary provided an understanding of the cost of underestimating uncertainty, there

remains a need for methodology to facilitate look-backs on forecasts to assess their reliability. More reliable forecasts, ones that reliably quantify uncertainty, are expected to have a lower disappointment (McVay and Dossary, 2012). Proper scoring rules such as the Brier score provide a method to rate assessors, and quantify the quality of their forecasts. Additional shortcomings of uncertainty assessment in industry were brought to light by Bickel and Bratvold (2007) through surveying industry professionals. Participants in the survey indicated a need to develop a quick and consistent method to assess uncertainty. Based on prior research, there is a need for methods to apply calibration and proper scoring to improve uncertainty assessment.

1.3 Objectives

The objectives of this research are to:

1. Develop a widely applicable method for tracking and improving probabilistic estimates over time, by storing the estimates, performing look-backs, assessing proper scoring, and applying external calibration.
2. Demonstrate the utility of the method by providing application examples from the petroleum industry.

1.4 Methodology

I have created a database that will allow users to store probabilistic estimates and improve them over time. The database allows users to input new assessments and assign 10%, 50%, or 90% probabilities to their estimates. The true values corresponding to the estimates are stored in the database when available. The Brier score and its

components, mainly calibration, are used to evaluate the estimates, once the true values are available. Using historic estimates in the database and the corresponding percent correct values, external calibration may be applied to new estimates. External calibration is a method of externally modifying and improving new estimates using knowledge of the percent correct values from previous estimates.

2. DATABASE MODEL

2.1 Relational Database

I developed a relational database for tracking and improving estimates using Microsoft Access (2010). A relational database is able to reduce redundancy of data by storing information in one table and linking it, as required, to other tables. A relational database improves the efficiency of updating data. If a user updates a data entry in one place it will be applied throughout the rest of the database where linked. The relational database is able to sort and process data using queries, which manipulate the data based on assigned criteria. I developed queries in the database to calculate the Brier score and its components from assessments. Forms may be used to interact with queries and display results. The database application I developed uses a form to specify the criteria to run different scoring rules. After the criteria are specified, the scoring rules may be run by linking to the appropriate queries. Forms may also be used to display graphs of tables and queries.

2.2 Database Structure

When designing the database, the first thing that I considered was its structure. It is important to understand primary keys and their function when designing a relational database. A primary key acts as a unique identifier for the entries in the table. A unique identifier is necessary in order to establish relationships between tables and perform

queries. Microsoft Access may automatically generate a primary key field, by assigning an increasing unique number to additional entries in a table.

When developing the structure of the relational database I followed the normalization rules that were developed by Dr. E. F. Codd, a former employee at IBM. A summary of Dr. Codd's rules for relational databases may be found in a book by Balter (2010). The three rules of normalization are referred to as the first, second and third normal forms, which serve as a guide to developing the database structure. The first normal form requires that each field in a table is reduced to its simplest possible form. For example, in order to maintain first normal form you cannot store first and last name in the same field (Balter, 2010). The first normal form also requires that fields have no repeated information. In order to agree with the repetition rule of the first normal form, there should not be multiple columns of data with the same type of information. An example that demonstrates repeated information would be columns that list out multiple orders such as order 1, order 2, and order 3. This structure of the database would limit the addition of a 4^{th} order and cause similar information to be repeated in a different field. The fields in tables should be reduced so that fields do not repeat similar information.

The second normal form requires that each field in a table must be dependent on the primary key. The second normal form only applies if there are composite primary keys, which are made up of two or more fields. The database I designed has only one primary key in each table, which is a unique increasing number for each additional row. For example, in an orders table the primary key, call it Order ID, is a unique increasing

number for each order. It is possible to have a primary key that is made up of multiple fields, in which case it is important for all other fields to be fully dependent on the primary key fields.

The third normal form requires that each field is independent of the other fields in a table. The third normal form prevents storing in a field the results of calculations involving data from other fields in the table, which would establish a dependency between the fields. For example, you would not store total cost in a table, which is a function of unit price and quantity. Instead, queries should be used to manipulate the data to perform calculations. The three rules of normalization are the guidelines that I followed when constructing my relational database.

There are three main tables in the database structure, the Assessor table, the Question table, and the Assessment table. The Assessor table contains the names and identification of the people and/or entities making estimates. The Assessor table holds the information for the First Name, Last Name, and Group ID. The primary key for the Assessor table is the ID field, which is an increasing unique number. The Assessor table and its fields are shown in Fig. 2.

The Question table stores the information about what is being assessed. The Question table holds the fields for the Assessed Quantity, True Value (of the quantity assessed), Date of True Value, Category, and Units. The Assessed Quantity field describes what is being assessed. The True Value field is the actual value for the quantity, and Date of True Value is the date the actual value was obtained or will be available. The Category and Units fields are used to describe and sort the types of

questions. The primary key for the Question table is the Question ID field, which is an increasing unique number for each of the question entries. The Question table and its fields are shown in Fig. 3.

The Assessment table stores the assessments made by assessors for a specific question. The Assessment table contains the percentiles assigned in the field Probability Assigned, along with their corresponding values in the Value Assigned field. The intent is to ultimately allow a user to assign any percentile for their estimate, but currently the database only works for P10, P50, and P90 percentiles. The Assessment table also contains the Date of Assessment field, which contains the date that the assessment was made. The Assessor and Question tables link through ID fields to the Assessment table so that information is not redundant, and the appropriate assessor and question are linked to the assessment. A Details field was added to the Assessment table so that additional information about the model or method used to make the assessment may be described. The primary key for the Assessment table is the Assessment ID field, which is an automatically generated increasing number for each new assessment. The fields in the Assessment table are shown in Fig. 4.

tblAssessor

ID	Last Name	First Name	Group ID
1	Fondren	Mark	RA
2	Pervez	Agwan	501-2012
3	Andrade	Edward	501-2012
18	Zahrah	Al Marhoon	501-2011

Fig. 2 Fields in the assessor table

tblQuestion

Question ID	AssessedQuantity	True Value	Date Of True Val	Category	Units
52	Well 1 Cost	1,533,976	1/1/2011	Well Cost Prediction	Dollars
53	Well 2 Cost	1,401,309	1/1/2011	Well Cost Prediction	Dollars
54	Well 3 Cost	1,748,119	1/1/2011	Well Cost Prediction	Dollars
55	Well 4 Cost	1,730,701	1/1/2011	Well Cost Prediction	Dollars

Fig. 3 Fields in the question table

tblAssessment

Assessment ID	Question ID	Assessor ID	Probability Assigned	Value Assigned	Date of Assessment	Details
6286	298	122	10	1025733.4280473	4/24/2013	2009 Data
6287	299	122	10	1377990.01004777	4/24/2013	2009 Data
6288	300	122	10	1237811.40649528	4/24/2013	2009 Data
6289	301	122	10	1210344.03812976	4/24/2013	2009 Data

Fig. 4 Fields in the assessment table

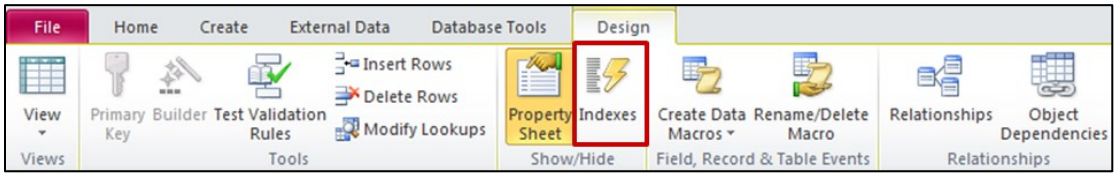

Fig. 5 Index button in design view of database

When designing the database structure I set up the tables so that some fields are required, and some are unique so that record redundancies are not permissible. In addition to the primary fields there are indexed and required fields that reduce record duplicates. For example, in the Assessor table I decided to have the combination of the First Name, Last Name, and Group ID set as unique. This is accomplished by creating a unique index with the three fields. The Index button is shown in the database in Fig. 5. This prevents assessors from being duplicated in the database. If there are two assessors with the same first and last name, a user may alter the Group ID to distinguish between

the two assessors. In the Assessor table the First Name, Last Name, and Group ID are required fields.

When considering how the Question table records should be unique, I decided that the fields Assessed Quantity, Date of True Value, and Category would provide the unique combination. The logic behind this is that some events, for example quarterly earnings, repeat every year but the unique field is the date. Including the Category field in the indexing provides additional flexibility in the naming convention used for Assessed Quantity. For example, if the Assessed Quantity is a well name, the Category may be modified to specify a unique attribute of the well, such as production or facilities cost, for the same Date of True Value. This indexing prevents duplication of the same Assessed Quantity with the same Date of True Value and Category fields. This structure allows questions to be filtered by Assessed Quantity to display all events. When setting the required fields for the Question table, I specified Assessed Quantity, Category, and Units to be required. The Category and Units fields are required to maintain the integrity of the records in the database; if a quantity happens to be unitless it may be specified in the Units field. It is important to note that if the Date of True Value is not specified for a question, the database may allow an Assessed Quantity name to be duplicated. Care should be taken when naming an Assessed Quantity and the Date of True Value should be entered when available.

In the Assessment table the fields Probability Assigned, Value Assigned, and Date of Assessment are required. In order to keep the entries in the assessment table unique, I indexed the Question ID, Assessor ID, Probability Assigned, Date of

Assessment, and Details. This will prevent an assessment from being made for the same question, by the same person, and with the same probability and approach (Details) more than once on the same day.

2.3 Scoring Rules and Queries

After developing the structure of the database I programmed the Brier score and its components. In order to maintain the normalization of the tables, queries were used to manipulate the data. The equations that were programed in the database are Eqs. 1–4, which may be found in the literature review above. It is important to clarify that Eqs. 1–4 are for a continuous cumulative distribution; the binary assessment described by Lichtenstein and Fischhoff (1977) is modified slightly. The binary assessment for the percent correct, according to the cumulative convention, is correct if the True Value is less than the assessment value. The percent correct, which is the ratio of correct responses to the number of responses in a percentage category, is modified in the equations when the binary assessment for the cumulative distribution convention is used. Likewise, the overall percent correct is affected by the cumulative distribution convention.

The knowledge component as defined by Lichtenstein and Fischhoff (1977) does not have the same significance for the continuous distribution convention being used here. Knowledge is meant to gauge an assessor's ability to make correct binary assessments. In order for an assessor to have a high knowledge score, the estimates for a continuous distribution require either the True Value to be less than the Value Assigned 100% of the time, indicating perfect knowledge, or 0% of the time, indicating perfect

lack of knowledge. The knowledge score does not reveal much about well-calibrated continuous distribution estimates. A well-calibrated continuous distribution would require the proportion of times the True Value is less than the Value Assigned to match the Probability Assigned. Since the probability values assigned are currently limited to 10, 50, and 90 percent, a well-calibrated continuous distribution would have a knowledge value of .25, which is the worst knowledge score possible.

Resolution ranks the range of varying percentages that are used to describe a forecast. The resolution is proportional to the squared difference between the individual proportion correct and the total proportion correct. The resolution term does not reveal as much as it would about the assessments if the probability values were not restricted. The resolution for these assessments is not as meaningful because the majority of assessments provided include 10 and 90 percent estimates. These estimates at the extremes have a larger effect on the resolution than the 50-percent estimate since the overall percent correct tends to be centralized. The overall percent correct may be calculated by a weighted average of the relative percent correct values, and graphically should be near the vertical center of the points on a calibration plot (Wilks, 2011).

The calibration component of the Brier score, which measures how close an assessor's assigned probabilities match the proportion of correct responses, still makes logical sense for the continuous distribution convention used for the binary assessment. Due to the convention, calibration will be the primary component of the Brier score used for evaluating estimates.

In order to program the percent correct in the database I used an IIF statement in Microsoft Access, which works just like a logical If statement. The IIF statement assigns a 1 if the True Value of the assessed quantity is less than the Value Assigned. If the True Value is not less than the Value Assigned then a null value is assigned. The percent correct is equal to the number of times the True Value is less than the Value Assigned divided by the total number of assessments. When programming the resolution and calibration components I added an additional IIF statement that returned a value of zero if there are no assessments in a percentage category. The extra IIF statement prevents an error statement in Access from a division by zero in the percent correct. Since both the resolution and calibration are multiplied by zero if the number of assessments is zero, the extra IIF statement does not alter the equations, but only prevents the database from returning an error.

When using queries in the relational database it is possible to make calculated fields. To use calculated fields as part of an equation requires saving the query and referencing it in an additional query. In Microsoft Access it is not possible to reference calculated fields without saving them as a separate query. This causes the number of queries used for simple equations to become large. When building the queries used to calculate the components of the scoring rules I had to create queries for the separate percentages assigned. The calibration and resolution queries each have three components related to the 10, 50, and 90 percent cases. The knowledge component is only based on the overall percent correct (Eq. 3).

The criteria I used to select how the scoring components were applied include the fields for Assessor ID, Group ID, Category, Date of Assessment, Date of True Value, and Details of the assessment. This set of criteria had to be applied to each of the components of the Brier score to provide a way to sort and filter the scoring rule. I set up the criteria for the queries so that they would be linked to a form, where the inputs for the criteria may be specified. Having the criteria applied in a form allows a clean interface for the user to specify the criteria they choose. I programed the criteria so that if a field is left blank in the form it would not be included for sorting. The form that I created to run the Brier score and its component queries is shown in Fig. 6. The criteria are entered in the text boxes of the form and scoring rule calculations are run by selecting the desired scoring rule component with the buttons on the bottom right of the form.

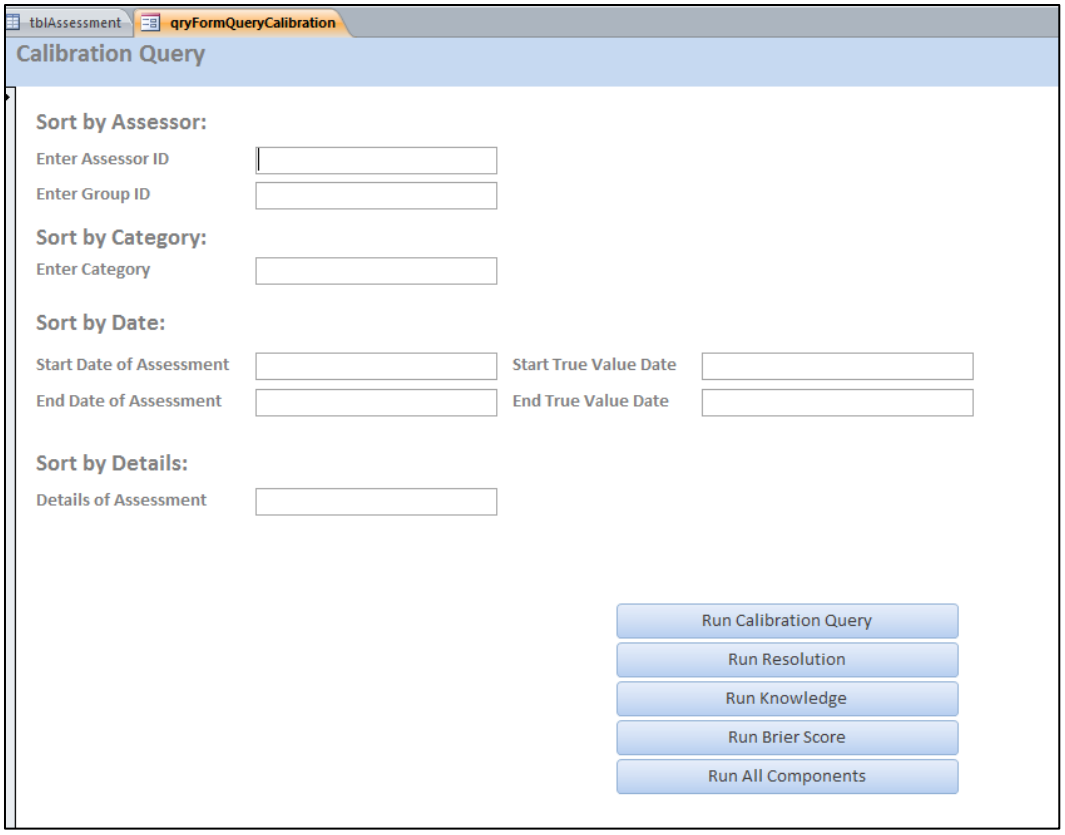

Fig. 6 Form for running the Brier score and its components

The form in Fig. 6 may be used to sort the different queries based on the available fields. If all of the fields are left blank when a query is selected the database will run the query for all the data available in the database. It is possible to run a query for an individual assessor by entering their Assessor ID. If an evaluation of a group is desired, the Assessor ID should be left blank and the Group ID may be used to filter the data. It is possible to combine the criteria fields so that an even narrower data set is used for the desired query. For example, Assessor ID, Category, and Date of Assessment may be used in combination to find assessments for a specific person, category, and date.

2.4 Importing and Exporting Data

After designing the structure of the database I imported student assessments as a test case. Probabilistic assessments from students were collected for the Fall of 2011 and 2012 during a senior level petroleum engineering course to test students' calibration and to track their improvement over time. There are several different ways that data may be imported into a Microsoft Access database. One way data may be imported is through using the external data tab (Fig. 7). When importing all of the student estimates I used this method to pull the data from Microsoft Excel into Access. In order to import data into an existing table in Access the field headers in Excel must match the corresponding column headers in Access. Changing the headers requires a small amount of formatting. I created a template Excel file with the appropriate names for the headers. Fig. 8 shows the required headers for importing into the Assessor table. Fig. 9 shows the headers for the Question table. Typically when I import data, I begin by adding the assessor(s) and then importing the questions, so that I know the Question ID and Assessor ID numbers to create in the Assessment table. The template for the headers of the Assessment table may be found in Fig. 10. With additional programming, for example a program in Visual Basic, it may be possible to pull the data from Excel into the respective tables all at once. It is important to note that when importing dates they should be saved with a general format in Excel in order to import correctly into Access.

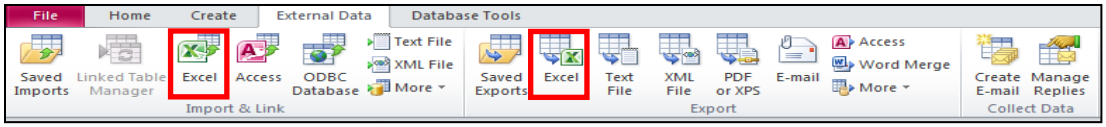

Fig. 7 External data tab in Microsoft Access

	A	B	C
1	Lastname	Firstname	GroupID
2			
3			

Fig. 8 Template headers for importing into assessor table

	A	B	C	D	E
1	AssessedQuantity	Truevalue	DateTrueValue	Category	Units
2					

Fig. 9 Template headers for importing into question table

	A	B	C	D	E	F
1	QuestionID	AssessorID	PAssigned	Vassigned	Dassigned	Details
2			10			

Fig. 10 Template headers for importing into the assessment table

Data may also be entered in the database manually, by opening a table and typing in the required fields. Another manual option for importing data is through the use of a form. I created an example form named Assessment to show how a form may be used to quickly make assessments for existing assessors and questions (Fig. 11). As seen in Fig. 11, the Probability Assigned, Value Assigned, Question ID, and Assessor ID may be quickly assigned.

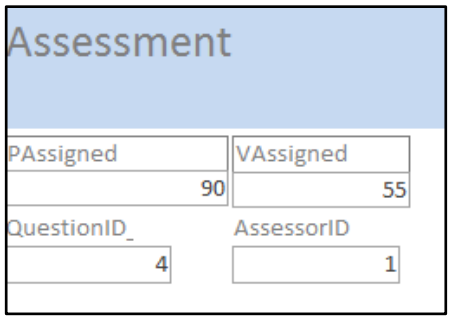

Fig. 11 Assessment form for manually inputting assessments

Just like importing data from Microsoft Excel, data may be quickly exported from Microsoft Access to Excel through a similar process. The external data tab in Access has the option to export data to Excel. Once a query is run or the data in a table are filtered as desired, they may be exported to an Excel spreadsheet, PDF, or even an email.

2.5 Plotting Data

In the database I developed it is possible to generate plots of the data. The plotting capabilities of Microsoft Access are very limited. It is difficult to include multiple series of graphs and to manipulate the format of the charts. I created a calibration plot in the database that plots the percent correct vs. percent assigned to be able to check the calibration of an assessor (Fig. 12).

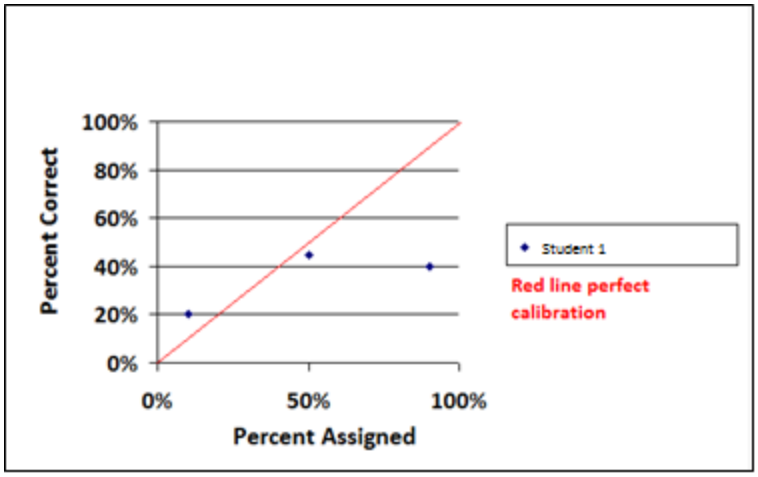

Fig. 12 Calibration plot in database

The plot in Fig. 12 pulls data from queries that are linked to the criteria available in the interface shown in Fig. 6. If criteria are not selected in the interface shown in Fig. 6, the criteria will be requested upon opening the plot. Since there is limited capability to manipulate plots in Microsoft Access, it is also possible to export data and results to Microsoft Excel where there is more plotting flexibility.

3. RESULTS

3.1 Introduction

In the following sections, three examples are presented to demonstrate the application of the database. The first two examples are from the petroleum engineering industry, and the third is from a general set of assessments, which show the general utility of the database. The first is a data set of drilling cost predictions for Barnett shale wells, the second is a set of externally calibrated reserves estimates for the Barnett shale, and the third is a data set of general knowledge assessments made by students.

3.2 Barnett Shale Drilling Cost Predictions

One of the data sets used to demonstrate the application of the database is a set of drilling cost predictions from the Barnett shale. Deterministic drilling cost predictions by engineers and the actual drilling costs were provided by an operating company. Valdes (2013) used this dataset to develop a model to improve deterministic drilling cost estimates. In the data set there are 158 wells drilled in 2011, 237 wells drilled in 2010, and 87 wells drilled in 2009.

The model that Valdes developed uses historical data to convert deterministic estimates into probabilistic estimates (Valdes, 2013). The model uses the actual cost divided by the estimated cost from historical data in order to obtain ratios for the deterministic estimates. Once ratios are obtained from historical data, distributions are fit to the ratios to obtain correction factors, and correlations are determined to account for

dependencies between the correction factors. Using simulation with Latin Hypercube sampling, deterministic estimates for future wells may be converted to probabilistic estimates using the correction factor distributions. The probabilistic cost estimates were demonstrated to be more accurate than the deterministic cost estimates. Treating the deterministic estimates as median (P50) values, the probabilistic cost estimates were also demonstrated to be better calibrated graphically, and quantitatively.

Valdes developed models for three different Barnett drilling-cost datasets. One model uses 2009 well costs to predict 2010 well costs, one uses both 2009 and 2010 costs to predict 2011 wells costs, and the last uses only 2010 costs to predict 2011 well costs. Each of the three cases has different correction factor distributions, due to the different historical data. I imported the engineers' deterministic estimates, Valdes' probabilistic estimates, and the true well costs into the database.

After importing the data into the database, the calibration and proper scoring of the engineers' estimates and the probabilistic estimates were evaluated by running the calibration query in the database. The results for both the engineers and the probabilistic estimates may be found in Table 1 which contains the percent correct values for the corresponding assigned probabilities of 10%, 50%, and 90%, as well as the calculated calibration scores. The deterministic estimates of the engineers were treated as P50 estimates.

Table 1 Drilling cost prediction results					
Assessor	**Model**	**Calibration**	**Actual Percent Correct Values**		
			10%	**50%**	**90%**
Valdes	2011 Well Cost (Using 2010 Historic Data)	.00005	.09	.50	.91
Valdes	2011 Well Cost (Using 2009 – 2010 Historic Data)	.00134	.11	.54	.94
Valdes	2010 Well Cost (Using 2009 Historic Data)	.00407	.06	.6	.93
Engineers	2011 Well Cost (Deterministic Estimates 50%)	.00901		.405	
Engineers	2010 Well Cost (Deterministic Estimates 50%)	.00748		.414	

The probabilistic estimates proved to be better calibrated than those of the engineers. The 2011 well cost estimates made by Valdes using the 2010 historical data have a calibration score two orders of magnitude smaller than the engineers' deterministic estimates for 2011. The model that Valdes used to predict the 2011 well costs using the 2009-2010 historic data was not as well-calibrated as the case that used just the 2010 history, but it is still better calibrated than the engineers' estimates for 2011. The 2010 well cost estimates made by Valdes using the 2009 historical data have a calibration score .003 less than the engineers' calibration for 2010 well costs.

To demonstrate some of the flexibility of the database, calibration scores for the drilling cost estimates were calculated by quarter using the Date of True Value criterion. The results of the drilling cost predictions for 2011 by quarter are shown in Table 2. The engineers' calibration improved for the second quarter, while Valdes's calibration was nearly the same for both quarters. Valdes's correction factors were not recalculated

for each quarter, which could explain why the calibration for the probabilistic estimates remained nearly constant for both quarters.

Table 2 Drilling cost predictions for 2011 sorted by quarter

Assessor	Model	Calibration	Actual Percent Correct Values			Quarter
			10%	50%	90%	
Valdes	2011 Well Cost (2010 Historic Data)	.00946	.01	.36	.88	Q1
Engineers	2011 Well Cost (Deterministic Estimates 50%)	.05216		.272		Q1
Valdes	2011 Well Cost (2010 Data)	.00942	.17	.65	.94	Q2
Engineers	2011 Well Cost (Deterministic Estimates 50%)	.00207		.545		Q2

In order to better explain the application of the relational database for this drilling prediction case, a scenario is provided to summarize the process. Historic estimates are stored over time in the database, both the estimates and true values. The historic estimates are used to correct the new estimates. The new estimates may also be stored in the database. All of the estimates, both the historic and new estimates, may be easily exported to Valdes's model in Microsoft Excel. In this case since the estimates are deterministic, Valdes's model is used to improve the estimates and convert them to probabilistic estimates using distributions of correction factors. The probabilistic estimates may then be imported back into the database to evaluate the calibration, once the true values are available. The deterministic estimates may also be used as historic

data for the next set of estimates once the true values for the new estimates are available. The database is an effective way to evaluate if an assessor is well calibrated. The results show that keeping track of estimates and calibration over time may be used to improve estimates.

3.3 Barnett Shale Reserves and External Calibration

An additional data set that is used to show the application of this database comes from a set of reserves estimates for Barnett wells. There are 197 Barnett wells that were used to hindcast reserves using half of the known history to predict the other half. The wells have roughly 80 months of production data. The hindcast for the wells were based on the first 40 months, to predict the last 40 months. The Arps decline curve analysis is applied with the constraints of $\boldsymbol{D_i} \geq 0$, and $\mathbf{0} \leq \boldsymbol{b} \leq 1$. The Jochen and Spivey method was the probabilistic method used to make the predictions (Jochen and Spivey, 1996). The Jochen and Spivey method is based on the bootstrap method in this case with 120 realizations. From Gong et al. (2011) it is known that the Jochen and Spivey method is not well calibrated. The data for the Barnett wells were imported into the database from a Microsoft Excel file. The data set includes P10, P50, and P90 estimates and the true value for the reserves. The database was used to evaluate the calibration of the wells and the percent correct values as shown in Table 3.

Table 3 Calibration of Barnett shale reserves estimates using Jochen and Spivey model					
Assessor	**Model**	**Calibration**	**Actual Percent Correct Values**		
			10%	**50%**	**90%**
Gonzalez, Raul	Jochen and Spivey	0.07981	.17	.29	.46

The calibration results in Table 3, may also be viewed graphically in a calibration plot generated with the database, and shown in Fig. 13. The calibration plot uses a query to plot the percent correct versus the percent assigned. It may be seen in Fig. 13 that the percentages assigned values of 50% and 90% are greater than the percent correct. When the percent assigned is greater than the percent correct it is a sign of overconfidence. A perfectly calibrated case would match the unit slope red line shown in Fig. 13.

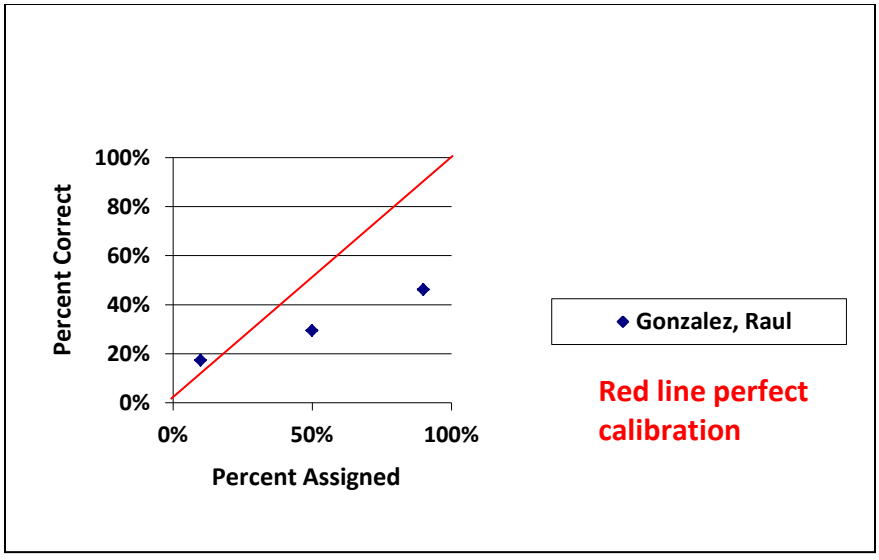

Fig. 13 Calibration plot of uncalibrated reserves estimates

In order to improve estimates, Raul Gonzalez developed a model to externally calibrate data. The model that Gonzalez developed uses knowledge of the calibration to improve estimates. This example of the Barnett shale will be used to describe the method of external calibration. For each well, estimates were made at P10, P50, and P90 probabilities. These first estimates will be referred to as the uncalibrated estimates, because external calibration has not been performed. At the field level, for the 197 wells, the percent correct values may be assessed, which will only match the percentages assigned if the wells are perfectly calibrated. Using the knowledge of the percent correct values a distribution may be defined for the wells. In order to perform the external calibration a lognormal distribution is chosen to obtain estimates that match the percentages assigned of 10%, 50%, and 90%. In order to define the lognormal distribution a mean and standard deviation are required. A guess for the mean and standard deviation are initially chosen to define the lognormal distribution for each well. The three percent correct values, obtained from the 197 wells, may be used to obtain reserves values from the lognormal distribution. The squared difference between the reserves values from the distribution and the uncalibrated reserves values are set as an objective function. The three squared differences are minimized to solve for the mean and variance of each well. Once the distribution is defined for each well the externally calibrated reserve values may be solved for using the assigned percentage values of 10%, 50%, and 90%. The external calibration uses the distribution to properly calibrate the estimates, for the appropriate response percentages.

The external calibration process was used to externally calibrate the estimates for the Barnett wells in this example. The results for the externally calibrated case are shown in Table 4.

Table 4 External calibration for Barnett reserves forecast using Jochen and Spivey model

Assessor	Model	Calibration	Actual Percent Correct Values		
			10%	50%	90%
Gonzalez, Raul	Jochen and Spivey	0.00067	.12	.54	.91

The results in Table 4 show a major improvement in the calibration and Brier score of the estimates. In order to visualize the calibration for the case in Table 4, a calibration plot for the externally calibrated estimates is shown in Fig. 14. The results in Fig. 14 are well calibrated and follow the unit slope almost perfectly.

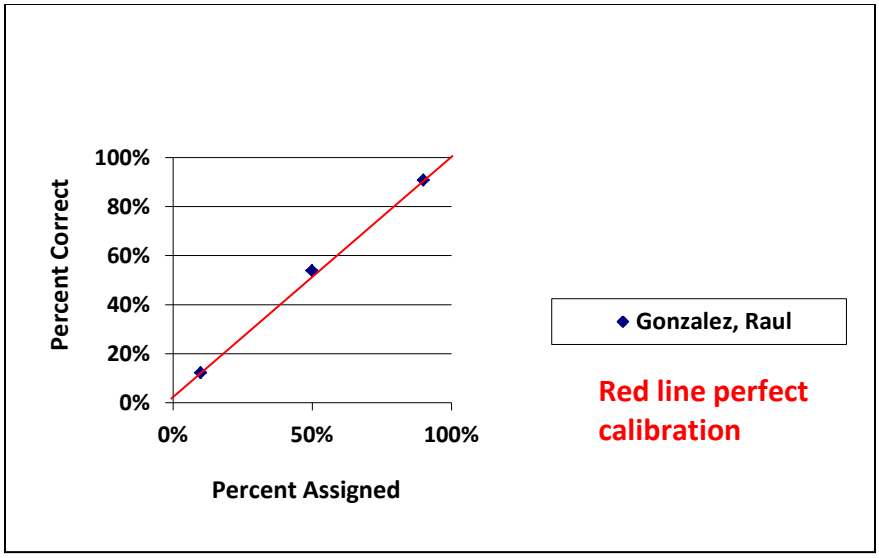

Fig. 14 Calibration plot of externally calibrated reserves estimates

In order to summarize how to apply this method of external calibration, an example for forecasting a longer period of time is provided. For this example a group of wells have an expected life of 7 years. In order to make a forecast a certain amount of known data is required. After collecting production data for 1.5 years a reserves forecast for 7 years is made and an additional reserves forecast for 3 years is also made. In order to improve our forecast for 7 years we want to check the calibration of the short term forecast. In order to check the calibration of our model we will perform a look-back. After 3 years we will check our forecast made at 1.5 years. Using the true values and estimates from the 1.5 year forecast, the percent correct values for the estimates may be obtained. The percent correct values should be close to the percentages assigned if the model is well calibrated. Using the percent correct values we may externally calibrate and improve the 7 year forecast that was made after 1.5 years. This process may be

repeated using 2 years of production history to predict 7 years and 4 years. The percent correct values and calibration will be evaluated again, and external calibration will be performed using the new percent correct values. The reason for repeating the process is to use the new production data to accurately capture the production decline. The number of look-backs may vary based on the amount of history available and expected life of the wells. It is expected that the externally calibrated forecasts should become more consistent as more production data is used in the forecast. This example demonstrates the importance of storing historic data in the database over time, for the purpose of look-backs and external calibration. As additional knowledge is acquired over time, additional look-backs and external calibration may be applied to improve future estimates.

External calibration may be applied by managers throughout a company or on an individual basis. Knowledge of a consistent error in the percent correct values of a company or an individual's estimates may be used to externally calibrate future estimates. Since external calibration is preformed after the initial estimates, external calibration may be performed by the estimator or an outside party such as a manger.

3.4 Student Estimates

Student estimates were collected for petroleum engineering and general knowledge questions. The estimates were collected throughout a semester and students were provided with feedback of their results. The goal of recording students' estimates was to make them aware of their overconfidence, and to keep track of their progress over time. At the onset of my research a Microsoft Excel spreadsheet was used to keep track of the student estimates, but it became apparent that there were limitations to using a

spreadsheet and that a database would be more suitable for keeping track of estimates. The student estimates were imported into the database for the Fall semesters of 2011, and 2012. The student estimates were used as a test case when developing the database.

The student data set provides an example of how the database is general enough to be applied in other industries. There were three question categories that students were tasked with estimating. The first category is a list of questions from a paper by Capen (1976), the second were Texas A&M football games, and the third petroleum engineering questions. The questions imported in the database included the Capen questions, and the football games. The petroleum engineering questions were not imported because they were in multiple choice format. The database currently is not designed to handle multiple choice assessments.

The calibration of the students' estimates for the football games showed interesting results. The calibration of the students improved over the course of the semester for the estimates of football games in 2012. I used the database to evaluate the calibration for the students for the first and second half of the football seasons. The calibration for the students for the first 6 games and last 6 games of the 2012 football season may be seen in Table 5. The results for the 2011 estimates were evaluated also, but no conclusions could be easily drawn between the first and second half of the semester, because there were an odd number of assessments.

Table 5 Calibration of 2012 Texas A&M football estimates				
1st half of 2012 season football games		Actual Percent Correct Values		
Assessor	**Calibration**	**10%**	**50%**	**90%**
501-2012	0.03944	.11	.38	.59
2nd half of the 2012 season football games		Actual Percent Correct Values		
Assessor	**Calibration**	**10%**	**50%**	**90%**
501-2012	0.01544	.04	.31	.81

The results of Table 5 show that the calibration for the estimates of the last six games improved from the first six games. The students showed an improvement of calibration for the second half of the semester from the first. The calibration plot for the results shown in Table 5 may be seen in Fig. 15.

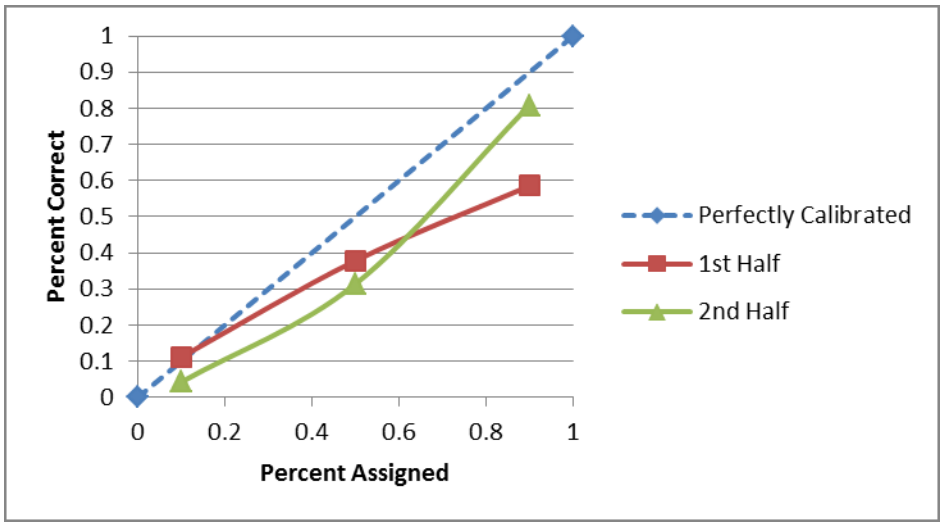

Fig. 15 Student estimates for the 2012 Texas A&M football season

The results of Fig. 15 show that the calibration for the 90th percentile estimates improved for the students in the second half of the semester. The results convey that the students became better calibrated the second half of the semester compared to the first. The reason for the improvement in calibration is related to the students applying look-backs for their estimates. Look-backs, provided students the opportunity to learn from their previous estimates and make better estimates in the future by scaling their estimates accordingly.

4. CONCLUSIONS

I created a relational database that facilitates the process of applying calibration and improving probabilistic estimation. The database stores probabilistic estimates and the true values over time, performs look-backs, and assesses proper scoring. The three application examples presented show how the database may be applied to improve probabilistic estimates. The first two examples were from the petroleum engineering industry; the third was from a general set of assessments and shows how the database may be applied in other industries. Each of the examples demonstrates the value of storing historical estimates and their actual values in the database over time, for the purpose of externally modifying and improving future estimates. The drilling cost prediction results show how historical drilling cost estimates may be used to improve future cost estimates. The reserves prediction example shows how reserves estimates may be improved using knowledge of previous reserves estimates. The student estimates example illustrates the broad application of the database, and demonstrates how look-backs contributed to better quantification of uncertainty and improvement of the students' estimates over time. Improving probabilistic estimation is expected to reduce disappointment. Continuous application tools and methods similar to those proposed herein should improve probabilistic forecasting, benefitting company and industry performance.

5. FUTURE WORK

There are areas for future work to improve the database that I have developed. Areas for improvement include extending the database functionality to allow assessors to assign any probability to assessments, functionality to accommodate multiple choice assessments, and providing methods of external calibration within the database. There are also additional areas that would add to the overall functionality of the database.

Currently the database allows assessors to make numerical assessments and assign probabilities of 10%, 50%, and 90% to their assessments. It could be beneficial to provide the functionality within the database for an assessor to assign any probability from a continuous distribution. Allowing assessors to assign any probability would impact the resolution score. The resolution is dependent on the probabilities assigned that have the largest difference from the overall percent correct. It may be that the resolution score would have more meaning if an assessor could assign any percentage from a continuous distribution.

The database currently allows users to make numerical assessments and assign a probability to their assessments which has, or will have, a true value. The database is not designed however to evaluate the scoring and calibration for multiple choice assessments. It would benefit the general nature of the database to add the capability for multiple choice assessments. There may be a situation where a questionnaire or survey is given in multiple-choice format. It might be beneficial to add functionality to the

database so that multiple-choice assessments may be included in evaluating an assessor's calibration.

In addition to the previous two improvements mentioned, it would be beneficial to add functionality within the database for external calibration. Currently it is possible to export the assessments to Microsoft Excel and perform external calibration using models developed in Excel. It would provide value to the database to have an external calibration method within the database. It could be difficult to provide a method of external calibration in the database since one of the limitations of Microsoft Access is the ability to easily manipulate data, or perform complex operations. It might be more practical to export the data to Excel, and perform the external calibration using models developed in Excel.

There are some additional alterations to the database that would improve functionality. One of which is the ability to run a loop of queries for a particular group of assessors so that the proper scoring for each person in the group is evaluated. Currently a user is required to run the Brier score components for each assessor individually or a group collectively. One is not able to automatically generate the individual results for each assessor in a group. In addition to adding a loop for the queries, there may be additional fields that the user may want to define to describe the quantities assessed or the assessments. The database was designed for general use, and it was not possible to foresee every scenario in which it would be used. The diverse applications for the database will drive the need for new descriptive fields, and its continued development.

REFERENCES

Balter, A. 2010. *Using Microsoft® Access® 2010*. Indianapolis, Indiana: Que.

Bickel, J.E. 2010. Scoring Rules and Decision Analysis Education. *Decision Analysis* **7** (4): 346–357.

Bickel, J.E. and Bratvold, R.B. 2007. Decision Making in the Oil and Gas Industry: From Blissful Ignorance to Uncertainty-Induced Confusion. Paper presented at the SPE Annual Technical Conference and Exhibition, Anaheim, California, U.S.A. Society of Petroleum Engineers SPE-109610-MS.

Bjorkman, M. 1992. Knowledge, Calibration, and Resolution: A Linear Model, Uppsala University, Uppsala, Sweden.

Brier, G.W. 1950. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review* **78** (1): 1-3.

Capen, E.C. 1976. The Difficulty of Assessing Uncertainty (Includes Associated Papers 6422 and 6423 and 6424 and 6425). *SPE Journal of Petroleum Technology* (08). 5579.

Gong, X., Gonzalez, R.A., McVay, D. et al. 2011. Bayesian Probabilistic Decline Curve Analysis Quantifies Shale Gas Reserves Uncertainty. Paper presented at the Canadian Unconventional Resources Conference, Alberta, Canada. Society of Petroleum Engineers SPE-147588-MS.

Jochen, V.A. and Spivey, J.P. 1996. Probabilistic Reserves Estimation Using Decline Curve Analysis with the Bootstrap Method. Paper presented at the SPE Annual Technical Conference and Exhibition, Denver, Colorado. 1996 Copyright 1996, Society of Petroleum Engineers, Inc. 00036633.

Lichtenstein, S. and Fischhoff, B. 1977. Do Those Who Know More Also Know More About How Much They Know? *Organizational Behavior & Human Performance* **20** (2): 159-183.

McVay, D.A. and Dossary, M. 2012. The Value of Assessing Uncertainty. Paper presented at the SPE Annual Technical Conference and Exhibition, San Antonio, Texas, USA. Society of Petroleum Engineers SPE-160189-MS.

Valdes, A. 2013. Uncertainty Quantification and Calibration of Well Construction Cost Estimates. Master of Science, Texas A&M, College Station, Texas (August 2013).

Wilks, D.S. 2011. Forecast Verification. In *Statistical Methods in the Atmospheric Sciences,* 3rd edition, ed. D. S. Wilks, Chap. 8, 301-394. Waltham, Massachusetts: Academic Press.

Wolff, M. 2010. Probabilistic Subsurface Forecasting - What Do We Really Know? *SPE Journal of Petroleum Technology* **62** (5): 86-92. SPE-118550-MS.