Casualty Actuarial Society
E-Forum, Summer 2013

The CAS *E-Forum,* Summer 2013

The Summer 2013 Edition of the CAS *E-Forum* is a cooperative effort between the CAS *E-Forum* Committee and various other CAS committees, task forces, or working parties.

This *E-Forum* includes three independent research papers and report 5 of the CAS Risk-Based Capital (RBC) Dependencies and Calibration Working Party (DCWP). Reports 1 and 2 are posted in *E-Forum* Winter 2012-Volume 1 and reports 3 and 4 are in *E-Forum* Fall 2012-Volume 2.

Risk-Based Capital Dependencies and Calibration Research Working Party

Allan M. Kaufman, *Chairperson*

Karen H. Adams	Robert Haimov	Douglas Robert Nation
Emmanuel Theodore Bardis	Jed Nathaniel Isaman	G. Chris Nyce
Jess B. Broussard	Shira L. Jacobson	Jeffrey J. Pfluger
Robert P. Butsic	Shiwen Jiang	Yi Pu
Pablo Castets	Alex Krutov	Ashley Arlene Reller
Joseph F. Cofield	Terry T. Kuruvilla	David A. Rosenzweig
Jose R. Couret	Apundeep Singh Lamba	David L. Ruhm
Orla Donnelly	Giuseppe F. LePera	Andrew Jon Staudt
Chris Dougherty	Zhe Robin Li	Timothy Delmar Sweetser
Nicole Elliott	Lily (Manjuan) Liang	Anna Marie Wetterhus
Brian A. Fannin	Thomas Toong-Chiang Loy	Jennifer X. Wu
Sholom Feldblum	Eduardo P. Marchena	Jianwei Xie
Kendra M. Felisky	Mark McCluskey	Ji Yao
Dennis A. Franciskovich	James P. McNichols	Linda Zhang
Spencer M. Gluck	Glenn G. Meyers	Christina Tieyan Zhou
James C. Guszcza	Daniel M. Murphy	Karen Sonnet, *Staff Liaison*

CAS *E-Forum,* Summer 2013

Table of Contents

Working Party Report

An Economic Basis for Property-Casualty Insurance Risk-Based Capital Measures

Independent Research

Kurtosis and Skewness Estimation for Non-Life Reserve Risk Distribution

The Impact of Different Forms of Decision-Aids on User Best Assessments

Weaving Actuarial Stories

E-Forum Committee

For information on submitting a paper to the *E-Forum*, visit
http://www.casact.org/pubs/forum/.

An Economic Basis for Property-Casualty Insurance Risk-Based Capital Measures

(RBC Research Working Parties Report 5)

Robert P. Butsic

Abstract: A solvency measure is needed to consistently and fairly determine the level of an insurer's capital, which is needed for protection against defaulting on policyholder claims. There are several competing measures in current use, including VaR and the expected policyholder deficit. However, there is no published analytical method for selecting or calibrating any of these measures to produce a level of capital consistent with economic principles.

  This paper develops an economic basis for selecting the solvency measure, and additionally determines how the measure can be calibrated to produce optimum capital. By maximizing policyholder welfare, a reasonable goal for regulation and corporate governance, I show that the optimal capital amount can be established by assessing the policyholders' perceived value of the expected default relative to the insurer's cost of holding capital. This optimality is achieved while allowing insurers a competitive rate of return.

  The result is that the proper solvency measure is *adjusted ruin probability*, where the probability distribution of losses or assets is modified to reflect policyholders' *risk preferences*. The optimal level of the adjusted ruin probability is uniquely determined by the frictional cost of holding capital. With this foundation, I also show that the *subadditivity* property of a coherent risk measure is an unnecessary criterion for evaluating insurance solvency.

  Under the policyholder welfare framework, the level of the adjusted ruin probability standard will vary by degree of policyholder risk aversion, interest rates, insurer income tax rates, amount of guaranty fund protection and other factors not considered in applying the above conventional solvency measures. I also discuss the relationship between the minimum regulatory level of capital and the insurer's optimal level.

Keywords: Solvency risk measures; policyholder welfare; optimal capital; adjusted probability distribution; certainty-equivalent losses; frictional capital costs; exponential utility; stochastic mean; subadditivity.

1. INTRODUCTION AND SUMMARY

The primary purpose of capital in an insurance organization is to protect policyholders, who in the event of insolvency, would not receive the full claim payment to which they are contractually entitled. Since there is an inverse relationship between the amount of an insurer's capital and the impact of insolvency on its policyholders, it is important to know (1) what kind of protection is desired, (2) how much protection is needed and (3) how much capital will provide the desired protection.

The first issue is addressed by selecting a *solvency measure*. The commonly used solvency measures are ruin probability, value-at-risk (VaR), expected policyholder deficit (EPD) and tail value-at-risk (TVaR). These solvency measures, which are discussed more thoroughly in section 5.4, use the probability distribution of losses and assets to characterize the harm to policyholders in an insolvency. The first two measures assess policyholder harm simply by whether or not an

insolvency occurs. The latter two measures incorporate both the likelihood of insolvency and its average value provided that it occurs. Given a particular solvency measure, the amount of protection is addressed by choosing the *level* of the solvency measure (for example, with VaR a specific confidence level must be assumed). Selecting the solvency level is called *calibrating* the solvency measure. After calibration, the required capital follows directly using actuarial and statistical techniques applied to the probability distributions for the relevant balance sheet items.

There is much debate over the proper choice of risk measure: some adherents tout technical features, such as subadditivity (e.g., Artzner [1999]) or practical ones, such as ease of explanation or common use in other financial service industries. However, to my knowledge, there is no literature that establishes a particular solvency measure based upon economic principles. Furthermore, despite the widespread use of solvency measures, there has been no analytic basis for setting the level of the risk measure — calibration has been arbitrary, using judgment. Although there is a vast literature on implementing risk measures, especially VaR, each author inevitably assumes that the calibration level (say, 99% VaR over one year) is given. There is no discussion regarding how to determine the specific level.[1] This is surprising, since it is well known that there is a trade-off between the cost of having too much capital and the downside of not having enough capital. For example, few would believe that a 99.999% annual VaR standard is appropriate, since this level implies too much capital, which would be extremely costly to carry. Conversely, a 60% VaR standard would indicate an intolerable risk of insurer insolvency. Therefore, some intermediate value of the VaR standard must be best.

In this paper, I have addressed both the solvency measure and the calibration concerns by establishing an analytical framework that directly applies the above cost-benefit relationship for capital. Given that the economic objective in setting capital standards is to *maximize policyholder welfare* while allowing a fair return to the insurer's owners, I show that this goal implies that there is an optimal capital amount for each insurer. That amount depends on three key inputs: the probability distribution of losses and assets, the insurer's cost of holding capital and the risk preferences of the policyholders. If the values of these underlying variables are known, then the optimal capital is uniquely determined. The theoretical optimal capital amount then forms the basis for regulatory capital standards, internal insurer risk management and

---

[1] The 99.5% VaR standard of Solvency II is based on mimicking the annual default probability underlying a Standard & Poors BBB rating. However, this approach dodges the question, since there is no objective reason why this particular default rate (0.5%) is superior to that of any other rating (e.g., AA).

pricing applications.

The key result of this paper is that the appropriate solvency risk measure is *adjusted ruin probability* (or a simple function of it), using a transformed distribution of outcomes for each component risk of the insurer. The adjusted distribution incorporates policyholder risk preferences and has a much fatter tail than the original distribution. This gives a suitable heavy weight to the extreme outcomes and prevents engineering the tail shape to manipulate capital requirements. Therefore, on an economic basis, the best solvency measure is *none* of the conventional measures. I show that using a conventional risk measure, such as VaR or expected policyholder deficit, will overstate the capital for low-risk losses (and assets) while understating the amount of capital for high-risk components. The latter effect is more serious.

The level of the adjusted ruin probability standard is unique and is a function of the *frictional cost of capital*. Thus, the calibration is not arbitrary. However, the adjusted ruin probability is equivalent to a conventional ruin probability standard that varies by the volatility of the insurers' component risks and by its policyholder risk preferences. So, even though the adjusted ruin probability standard may be fixed for all insurers, the corresponding unadjusted ruin probability will vary by line of business and by insurer.

Under the policyholder welfare framework, the level of the adjusted ruin probability standard will differ by degree of policyholder risk aversion, interest rates, insurer income tax rates and other factors not considered when applying the above conventional solvency measures. Thus, the adjusted ruin probability standard is not static and will vary over time. Another consequence of the policyholder welfare basis is that the amount of guaranty fund protection will also influence insurer capital. This result is important and (to my knowledge) has been ignored in the previous insurance literature.

Although the optimal level of capital may be appropriate as a standard for internal insurer governance and for pricing applications, I discuss how the regulatory level of capital should be lower than the insurer's optimal level.

1.1 Outline

The remainder of the paper is summarized below:

Section 2 provides some historical background for the development of solvency risk measures as applied to insurance and other financial firms.

Section 3 develops the notion of consumer surplus (relabeled as consumer value) and the certainty-equivalent value for insurance losses. The basic idea here is that consumers are risk-averse and will pay more for insurance than the expected value of their losses. I have used these concepts, which are the economic foundation for insurance, to value insurer default from the policyholder's perspective. This section also relates the certainty-equivalent loss concept to utility theory.

Section 4 develops a simple one-period model of an insurer with risky losses and riskless assets and specifies the cost of holding capital. This section also formulates the premium charged to policyholders, which includes the frictional capital cost.

Section 5 shows how the consumer value of the insurance transaction is maximized by minimizing the cost of holding capital plus the value to the policyholder of the insurer's default. This section shows that the optimum amount of capital is determined from the adjusted (for policyholder risk preferences) ruin probability. It compares results from the adjusted ruin probability to those from conventional solvency measures and shows that the coherent risk measure property of subadditivity is not necessary for an economically valid insurance solvency risk measure.

Section 6 discusses how the results of section 5 can be extended to include asset risk, guaranty funds and multiple-period assets and liabilities.

Section 7 examines implementation issues in applying the above capital-setting methodology, including its use in regulatory risk-based capital.

Section 8 provides a brief conclusion.

2. HISTORY OF SOLVENCY RISK MEASURES

European actuaries have applied risk measures for decades. Ruin theory, also called collective risk theory, is a branch of actuarial science that studies an insurer's vulnerability to insolvency based on mathematical modeling of the insurer's surplus (capital). The theoretical foundation of ruin theory, known as the classical compound-Poisson risk model in the literature, was introduced in 1903 by the Swedish actuary Filip Lundberg.[2] Usually, the main objective of the classical model and its extensions was to calculate the probability of an insurer's ultimate ruin.

The ruin probability measure has seen some use for internal insurer risk management, but has not yet been directly used for solvency regulation (although the closely related VaR has).

The VaR measure was introduced in 1945, as a means of measuring bond portfolio risk.[3] In the 1970s, as leverage became widespread, securities firms sought more effective ways to manage portfolio risk. They wanted a single risk metric that could be applied consistently across asset categories, including derivatives, which were becoming increasingly complex. Concurrently, computing power became cheap enough to analyze large portfolios. However, VaR was still viewed as a theoretical tool.

During the early 1990s, concerns about the proliferation of derivative instruments, some well-publicized massive trading losses and the 1987 stock market crash spurred the field of financial risk management. Through its RiskMetrics service, JP Morgan introduced VaR to professionals at many financial institutions. Ultimately, the value of proprietary VaR measures was recognized by the Basle Committee, which authorized their use by banks for performing regulatory capital calculations.

VaR became common in the banking and finance industry in the 1990s onward. It is used to control the risk of the positions in investment portfolios or bank divisions for managers of these units. Supporters of VaR-based risk management claim that a major benefit of VaR is the improvement in systems and modeling it forces on an institution (see Jorion [2006]). For insurance, it is the measure used in Europe for the capital standards of the Solvency II regime.

After the 2008 financial crisis, VaR came under severe criticism (see Nocero [2009] and Einhorn [2008]), primarily because of abuses in its implementation. It has been argued that the

---

[2] See Lundberg [1903].
[3] For a more detailed discussion on the history of VaR, see Holton [2003].

2008 financial crisis was exacerbated by bankers misusing VaR. In order to reduce apparent risk levels (and thereby regulatory capital) for mortgage-backed derivatives, the banks engaged in "tail-stuffing," wherein the securities were purposely designed to increase the amount of risk in the tail, while keeping VaR at a low level. These abuses highlighted a technical weakness of VaR, in which very large extreme events are treated equally with events just large enough to breach the VaR confidence level.

The expected policyholder deficit (EPD) measure first appeared in the insurance financial literature in Butsic [1994]. This work arose from participation in the American Academy of Actuaries Property-Casualty Risk-Based Capital group, which advised the NAIC in its development of the current RBC method in the early 1990s. The concept developed as a response to a perceived deficiency in using ruin probability (or its VaR equivalent) as a solvency standard in that it did not incorporate the *depth* of an insurer's insolvency.

TVaR had a similar genesis in banking and investment management as the EPD in insurance. It was also a response to the same deficiencies in applying VaR. The above tail-stuffing abuses would have been severely mitigated under a TVaR metric. The TVaR concept saw implementation and became common in the 2000s. It is presently used as the solvency measure in Swiss insurance capital regulation.

Within the last decade or so, a new class of risk measures called *spectral measures* have been developed (see Acerbi [2002]). They are based on TVaR and include a risk-aversion component; i.e., extreme tail events are given weights that correspond to the investors' desire to avoid them. If the weights are large, more capital is required. I have used the concept of risk-aversion in this paper, although under a different context (i.e., optimization).

3. CONSUMER VALUE AND CERTAINTY EQUIVALENT LOSSES

The important concept of this section is that individuals (and organizations as well) will pay more than expected value to insure losses. The implied value placed on losses by the policyholder is called the *certainty equivalent* value. The economic gain from buying insurance is called the consumer surplus, which I have renamed as *consumer value*. We can measure the consumer value by using a modified version of the underlying probability distribution of losses. The adjusted distribution provides the means for determining the expected loss of consumer value due to the possibility that the insurer becomes insolvent. The adjusted distribution can be

determined directly from a policyholder's utility function.

3.1 The Consumer Surplus Concept

The fundamental basis for insurance is that individuals are risk-averse: they are willing to pay more than the mathematical expected value of their potential loss in order to buy insurance.

As a simple example, suppose an individual is subject to loss on a home worth $100,000. There is a 1% chance of a total loss and a 99% probability of no loss. The mathematical expected value of the loss is $1,000 = 0.01 x 100,000. However, suppose that the homeowner is willing to pay up to $1,500 to an insurer to completely remove the risk of loss. Meanwhile, an insurer will charge only $1,100. The insurer is able to charge less than the policyholder (PH) is willing to pay because, through the law of large numbers, the risk to the insurer is reduced by pooling similar risks from other PHs.

The above three amounts ($1,000, $1,500, and $1,100) are important, and deserve a distinct nomenclature. The first is commonly called the *expected loss,* and in actuarial parlance the *pure premium.* The second is the *certainty equivalent expected loss* and the third the *premium.* In setting prices, actuaries normally include specific loadings for the insurer's expenses and for its bearing the risk, although the level of those loads is often limited by competition. In any event, the premium represents the insurer's price actually charged for bearing the risk.

The difference between the premium and the expected loss ($100 in the example above) is called the *provision for expenses and profit.* The difference between the certainty equivalent loss and the premium ($400) is called *consumer surplus* in the economics literature.[4] It is the difference between the total amount that consumers are willing and able to pay for a good or service and the total amount that they actually do pay (i.e., the market price for the product). If the consumer surplus is greater than zero, then the policyholder will buy insurance; otherwise the policyholder will self-insure.

The consumer surplus concept was introduced by Alfred Marshall in 1890,[5] and was designed to measure the welfare effects of economic policy. Standard microeconomics textbooks use the consumer surplus concept to equilibrate supply and demand. Consumer surplus applications are

---

[4] In finance, the equivalent concept is called the *risk premium* (See Panjer [1998]). I have not used this term here, since the term as used in insurance often represents the market price of risk (what the insurer charges for risk) and not what the policyholder is *willing* to pay.
[5] Marshall, Alfred (1947 [1890]): *Principles of Economics.* London, Macmillan.

common in welfare economics and government regulation, where cost-benefit analyses are needed.[6]

For insurance, consumer surplus represents the monetary benefit to the PH of having insurance. Since "surplus" commonly represents a different concept (capital) in insurance, for this paper I have renamed consumer surplus as *consumer value*.

I assume that in purchasing insurance, the PH will seek to maximize consumer value, which equals the difference between the certainty equivalent value of the potential loss, and the premium the PH must pay.

3.2 Mathematical Formulation of Certainty Equivalent Losses

For an *individual* PH having an exposure to an insurable risk, let *y* represent the size of a possible loss and $p(y)$ the probability that *y* occurs. The expected loss *L* is the summation of each possible loss times its probability: $L = \int_0^\infty y p(y)\, dy$.

Assume that for each possible loss amount *y*, there is a unique amount $\hat{y} = k(y)y$ representing the certainty equivalent (CE) loss. Call the term *k(y)* the *certainty equivalent function*. Therefore, the CE expected loss (CEL), denoted by $\hat{L}$, will be the expectation

$$\hat{L} = \int_0^\infty \hat{y} p(y)\, dy = \int_0^\infty k(y)y p(y)\, dy. \tag{3.21}$$

Since policyholders are risk-averse, we have $\hat{L} \geq L$. The value $k = \hat{L} / L$, or the average of the certainty equivalent function, is a useful parameter. Notice that $k \geq 1$.

If the premium equals the expected loss, then the consumer value of the insurance equals the difference between the CEL and the expected loss, or $\hat{L} - L$.

The contribution to the CE expectation from loss size *y* in equation 3.21 can also be expressed as $k(y)y p(y) = y\hat{p}(y)$, where $\hat{p}(y) = k(y)p(y)$. Here, $\hat{p}(y)$ is a transformed probability that will give greater weight to large loss values and less weight to small values than $p(y)$, thus producing an expected value greater than *L*. An alternative version of equation 3.21 is then

$$\hat{L} = \int_0^\infty y\hat{p}(y)\, dy = \int_0^\infty k(y)y p(y)\, dy. \tag{3.22}$$

---

[6] For example, see Einav, Finkelstein, and Culleny [2010].

Notice that the certainty-equivalent probabilities $\hat{p}(y)$ are conceptually similar to the risk-neutral measures that form the cornerstone of modern finance theory.[7] Transformed probability measures have also been used to value insurance losses in pricing models (see Wang [1996] and Butsic [1999]). The CE probabilities can be considered as subjective weights attached by the PH to the various loss sizes. Since these weights are equivalent to probabilities, they must sum to 1. Thus, an important restriction on the CE function $k(y)$ is that $\int_0^\infty k(y)p(y)\,dy = 1$. Appendix A3 discusses this restriction further. It also explains why $k(y)$ depends not only on the particular loss size y, but on all the other possible loss sizes as well.

3.3 A Numerical Example

A numerical example will help to illustrate these concepts: suppose a PH faces a loss which can have three values {100, 400, 1200} with respective probabilities {0.60, 0.30, 0.10}. The corresponding CE function values are {0.80, 1.10, 1.90}, showing an increasing risk aversion with loss size. Table 3.3 below shows details of the CE expected loss calculation.

Table 3.3
Certainty-Equivalent Loss Calculation for Numerical Example

					Total
Loss Amount	y	100	400	1200	
Probability	$p(y)$	0.60	0.30	0.10	1.00
Expected Value Component	$y \cdot p(y)$	60	12	120	300
CE Function	$k(y)$	0.80	1.10	1.90	1.36
Certainty Equivalent Loss	$y \cdot k(y)$	80	440	2280	
CE Exp. Loss Component	$y \cdot k(y) \cdot p(y)$	48	132	228	408
CE Probability	$k(y) \cdot p(y)$	0.48	0.33	0.19	1.00

The expected loss L is 300 and the CE expected loss is 408, giving an average CE function value of $k = 1.36 = 408/300$. The CE probabilities have shifted from their unadjusted counterparts: the subjective chance of the small (100) loss drops from 60% to 48% and the subjective likelihood of the larger losses increases, from 30% and 10% to 33% and 19%.

---

[7] The risk-neutral concept was first introduced by Arrow and Debreu [1954]. The Black-Scholes option pricing model can be derived using risk-neutral valuation, as shown in Hull [2008] (pages 307-309).

3.4 Utility Theory and Certainty Equivalent Losses

By using basic principles from utility theory, we can derive some general properties for the CE function. There is a direct connection between utility theory and the certainty equivalent. As shown in Appendix A1, the CE function can be determined from the utility function and the loss distribution. The certainty equivalent and the expected utility formulations are dual processes; each can be determined using the inverse of the other.[8]

The first property is that the CE value function $k(y)$ *increases* with loss size y. This occurs because utility increases with wealth, as Appendix A2 shows.

Second, because of risk aversion, $k(y)$ increases at a growing rate: its second derivative with respect to loss size is positive. Appendix A2 discusses this property in more detail.

Third, as discussed below, the certainty-equivalent expected loss (and thus each $k(y)$ value) depends on the *variance* of the loss distribution.

If the variance is non-zero, then we apply the basic utility theory assumption that PHs are risk-averse. For $U(x)$ representing the utility of (wealth given) a loss x, this implies a downward-sloping utility function, (i.e., $U'(x) < 0$) and that the function is concave (i.e., $U''(x) \leq 0$). The *absolute risk aversion* function[9] is defined as

$$R_A(x) = U''(x) / U'(x). \tag{3.41}$$

Denoting the variance by σ^2, it is straightforward to show[10] that the certainty-equivalent loss is approximated by

$$\hat{L} \cong L + \tfrac{1}{2}\sigma^2 R_A(W_0 - L). \tag{3.42}$$

Here, W_0 is the initial wealth of the PH. This important result shows that $\hat{L} - L$ is (approximately) directly proportional to the *variance* of the loss distribution and also proportional to a measure of risk aversion. If the loss distribution is *normal* and the utility

---

[8] This duality is established by Yaari [1987], who determines that the CE value of a risky prospect can be found using an adjusted probability distribution.

[9] See Pratt [1964]. As shown in Appendix A1, the sign of the absolute risk aversion function is negative when utility is a function of wealth, but positive when a function of loss size.

[10] See Panjer et al. [1998], page 137.

function is *exponential* with risk aversion parameter *a*, then equation 3.42 is exact,[11] and as shown in Appendix A4, the CEL becomes

$$\hat{L} = L + \tfrac{1}{2} a \sigma^2. \tag{3.43}$$

To summarize the above results, we see that the CEL is a function of both consumer risk preferences and the variance of the loss distribution. Additionally, the certainty equivalent value function increases with loss size at an increasing rate. These properties are essential to determining the optimal capital for an insurer, as I develop in section 5.

Although the optimal capital results can be developed directly from the underlying utility function, I prefer the certainty equivalent approach to valuing risk aversion, since it is more direct than the utility method and provides a tangible, monetized conversion of the expected loss. Also, there is a *unique* CE value for each possible loss size, while under expected utility, any linear transformation of the utility function will give valid results. Thus, the expected utility of a particular loss size has no meaning by itself.

3.5 The Certainty Equivalent Expected Default Value

In order to determine the optimal capital amount for an insurer (in section 5) it is necessary to find the consumer value of the insurance contract. This entails knowing the certainty equivalent value of the expected default. Here, for simplicity, I assume an insurer with a single policyholder. The expected default, also known in the actuarial literature as the *expected policyholder deficit*, is

$$D = \int_{A}^{\infty} (y - A) p(y) \, dy. \tag{3.51}$$

Here, *A* represents the insurer's assets, which are assumed to be fixed (non-stochastic) for this application, and *y* is the individual policyholder loss size. The CE value of the expected default, denoted by $\hat{D}$, and abbreviated to CED, can be determined by finding the CE value of the loss actually paid by the insurer (allowing the possibility of default) and subtracting it from the CE of

---

[11] The exponential utility function can be expressed as $U(x) = -\exp[-a(W_0 - x)]$, where *a* is the risk-aversion parameter, W_0 is initial wealth and *x* is the loss value. Since $\exp(aW_0)$ is a constant and utility functions are invariant to scale transformations, the utility reduces simply to $U(x) = -\exp(ax)$. Thus the utility at any loss size *x* is independent of initial wealth.

the loss $\hat{L}$ without possibility of default. The CE value for losses limited to the amount *A* is
$\hat{L}(A) = \int_0^A y\hat{p}(y)dy + A\int_A^\infty \hat{p}(y)\,dy$. Notice that, for losses larger than *A*, the amount of loss paid
by the insurer is simply *A*. Here, to be consistent with the valuation of losses below the amount
A, we must also use the subjective CE probability $\hat{p}(y)$ of the loss being greater than *A*. We
cannot use the unadjusted probability *p(y)*. To get the CED, we have $\hat{D} = \hat{L} - \hat{L}(A)$, or

$$\hat{D} = \int_0^\infty y\hat{p}(y)\,dy - \int_0^A y\hat{p}(y)\,dy - A\int_A^\infty \hat{p}(y)\,dy = \int_A^\infty (y-A)\hat{p}(y)\,dy. \qquad (3.52)$$

Here, the only difference from the equation 3.51 expected default calculation is the
substitution of the CE probability $\hat{p}(y)$ or its equivalent *k(y)p(y)* for the unadjusted probability
p(y). Notice that if *A* = 0, the CE default value equals the CE expected loss.

To illustrate the certainty equivalent default concept, assume a simple case where there is a
98% chance of a $0 loss and a 2% chance of a $1000 loss. The expected value of the loss is $20
= 0.02(1000). Also assume that the PH has risk aversion governed by exponential utility:
$U(x) = -e^{ax}$, where *x* is the loss size. The risk aversion parameter is *a* = 0.002. The utility of the
0 loss is –1 and the utility of the $1000 loss is
 –7.389 = –exp[0.002(1000)]. The expected utility is –1.128 = 0.98(–1) + 0.02(–7.389). The
certainty-equivalent loss is the loss size for which the actual utility equals the expected utility;
thus the CEL is $60.13:   –exp[0.002(60.13)] = –1.128.

Now suppose that the loss to the PH is limited to $900. The utility of this amount is
 –6.050 = –exp[0.002(900)]. The expected utility of the PH's retained loss is therefore
–1.101 = 0.98(–1) + 0.02(–6.050). The certainty equivalent value for the retained loss is $48.11.
Therefore the CE value of the uppermost $100 of protection is the difference between the CE of
the entire loss and the CE of the retained loss:
$12.02 = 60.13 – 48.11. Note that the expected value of the $100 coverage is only $2.00.

The $12.02 also represents the *CE value of default* when only the first $900 of loss is actually
covered by the insurer. It is the difference between the CE expected value of the entire loss,
minus the CE value of the coverage actually provided.

It is interesting to compare the result of covering the last $100 of the loss (as above) with

covering the first $100 of loss.[12] If the insurer covers the amount *above* $100 (i.e., the deductible), the utility of the retained loss is –1.221, with an expected utility of –1.004 and a CE of $2.21. If the deductible is $200, then the CE of the retained loss is 4.89, giving a $2.69 = 4.89 – 2.21 CE value for the layer from $100 to $200. The CE value of each layer progressively increases as we move up to higher layers.

In this example, the ratio of $\hat{D}$ to D is 6.01 for assets of $900, which is greater than 3.01, the ratio of the $60.13 CEL to the $20 expected loss. In general, the ratio of $\hat{D}$ to D will increase with the asset (and thus capital) amount.

This *tail leverage* is a consequence of PH risk aversion (which creates a fatter tail than for an unadjusted distribution), combined with the volatility of the loss distribution. To analyze the tail leverage in more detail, consider a normal distribution of losses for a PH with an exponential utility function. Appendix A4 provides a general method for determining the CE default value for a given loss distribution paired with a specific utility function. It also derives explicit results for the normal-exponential model.

Even with a small variance, the ratio of $\hat{D}$ to D can be large: assume a normal distribution with mean loss of 1000 and a 100 standard deviation. Suppose we have assets of 1100, which is 1 standard deviation above the mean, and that the risk aversion parameter is $a = 0.02$. From equation 3.43, the overall CEL is $1100 = 1000 + (.01)(200)^2$, giving k (the average CE function across all loss sizes) of 1.10. The straight expected default D is 8.33, but its certainty equivalent $\hat{D}$ is 57.39, a ratio of 6.89 to 1. For 2 standard deviations above the mean ($A = 1200$), we have $D = 0.85$ and $\hat{D} = 20.17$, for a ratio of 23.75. Table 3.5a shows results for these and other asset values:

---

[12] Notice that, given that the loss has *already occurred*, it doesn't matter whether the $100 amount comes from retaining a $100 deductible or sustaining a net $100 loss on a $1000 loss where the insurer pays the first $900. The losses are completely equivalent to the PH. In fact, the certainty-equivalent concept makes no sense here, because *ex post*, the losses are certain. However, *ex ante*, the CE and expected utility concepts transform risky outcomes to fixed values that differ from the actual values that may occur.

Table 3.5a
Tail Leverage for Numerical Example;
Normal Loss Distribution, Exponential Utility (a = 0.02)

Assets	D	CED	Ratio CED/D	Ruin Probability (RP)	Adjusted Ruin Prob. (ARP)	Ratio ARP/RP
1100	8.33	57.39	6.9	15.866%	50.000%	3.2
1200	0.85	20.17	23.7	2.275%	25.161%	11.1
1300	0.04	4.44	116.2	0.135%	8.054%	59.7
1400	0.001	0.50	701.4	0.003%	1.291%	407.5

The adjusted ruin probability is the probability of default with the CE probability distribution used instead of its unadjusted counterpart. Figure A4 in Appendix A4 shows the probability densities for the normal distribution and its CE transformation. Here, it is clear that the tail of the adjusted distribution is much fatter than that of the parent normal distribution.

Table 3.5b gives results for a higher risk aversion (*a* = 0.04):

Table 3.5b
Tail Leverage for Numerical Example;
Normal Loss Distribution, Exponential Utility (a = 0.04)

Assets	D	CED	Ratio CED/D	Ruin Probability (RP)	Adjusted Ruin Prob. (ARP)	Ratio ARP/RP
1100	8.33	136.49	16.4	15.866%	68.281%	4.3
1200	0.85	77.25	91.0	2.275%	50.000%	22.0
1300	0.04	36.49	954.8	0.135%	31.719%	235.0
1400	0.001	13.01	18200.4	0.003%	15.883%	5015.0

Notice that even though the unadjusted probability and default values at the extreme tail are rather small, their CE equivalents may be meaningful. However, it is important to recognize that for a practical loss distribution and CE function (with extremely large losses truncated by policy limits and with CE factors restricted by wealth effects[13]) the CED values would not be as large as shown in this illustrative example.

---

[13] For example, bankruptcy laws limit the harm of an uninsured large third-party loss to an individual. An individual with $100,000 of net worth is likely to value a $1 million loss about the same as a $2 million loss.

3.6 Default Values with Multiple Policies

The previous sections have analyzed results for a single policyholder under the assumption that an insurer covers only that PH. Here, I extend the analysis to an insurer with multiple PHs. However, the analytical perspective remains that of the *individual* PH. But now if a default occurs, the default amount is *shared* among the individual policyholders in proportion to their respective loss amounts. Assume that the PHs are homogeneous, with the same loss distribution and risk preferences.

To illustrate CE valuation of multiple-policy default, I use the binary loss[14] example with exponential utility introduced in section 3.5. Appendix A5 develops the CE values for the binary loss model, including the CED and the adjusted ruin probability. Assume that we have two PHs with independent losses defined by the section 3.5 example: a loss of $1,000 with a 2% probability and zero otherwise. The CE values are determined from exponential utility with a risk aversion parameter $a = 0.002$.

To illustrate the expected default calculation, suppose that an insurer covers these two PHs with an amount of assets equal to $200 per PH, or $400 in total. The default amount for a single PH (say, PH 1) depends on whether a loss occurs for PH 1 and whether a loss occurs for PH 2. If PH 1 doesn't have a loss, then there cannot be a default amount for PH 1 no matter what happens to PH 2. If PH 1 has a loss and PH 2 does not have a loss (with 0.0196 = 0.02 x 0.98 probability) then the default amount for PH 1 is $600 = 1000 – 400: all $400 of the insurer's assets cover the loss. If both PHs have a loss (with probability 0.0004 = 0.02 x 0.02) the total default amount is $1,600 = 2000 – 400, but it is shared equally, so PH 1 has a default amount of $800. Therefore, the expected default for PH 1 is $12.08 = 0.0196(600) + 0.0004(800). Notice that in the case where the insurer has only one PH (with $200 of assets), the expected default is larger: $16.00 = 0.02(1000 – 200).

To calculate the CE expected default value for PH 1, we use equation A5.5 from Appendix A5, getting $22.99. This compares to $38.05 for the single-risk insurer. Figure 3.6 compares, by asset value per PH, the CED per PH for the single-risk insurer to that of a two-risk insurer.

---

[14] This simple loss model with a general utility function is used in the influential paper on insurance market equilibrium by Rothschild and Stiglitz [1976]. Note that a one-year term life insurance contract has a binary loss distribution.

Figure 3.6
CE Default Values per PH by Asset Amount for a Single Risk and Two Risks
Binary Loss with Exponential Utility
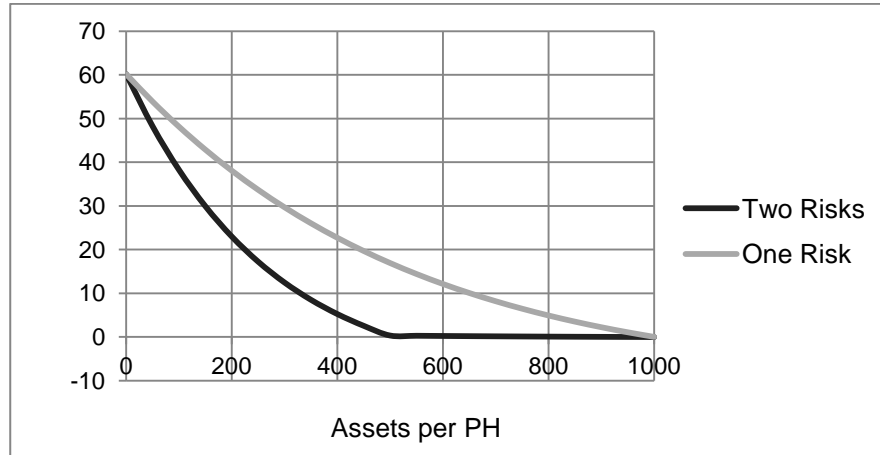
Here, the CE expected default value per PH is lower when the risks are combined. This effect is a consequence of diversification, where the variance of losses per policy is reduced by adding risks to the insurer's portfolio.

In general, we can determine the per-PH CED for multiple risks by finding (or approximating) their joint CE probability distribution. This process is analogous to that of finding the unadjusted expected default for a portfolio of losses. In the case where the sum of the CE losses has the same distribution as a component CE loss (as in the normal distribution), the expected default calculation is straightforward. Otherwise, we must resort to approximation methods.[15]

The assumption of statistical independence will drive the expected default value to an extremely low level if the number of policies is large. However, in reality, insurance losses are correlated. They are subject to common factors such as inflation, regulation, the legal system and multi-loss events like catastrophes. Further, the mean of losses for a given line of business (or other subdivision of an insurer's risk portfolio) is not known; it must be determined empirically. This effect is an important case of parameter risk,[16] which adds to the uncertainty of standard

---

[15] One method is to assume that the distribution for the sum of the CE losses has the same distribution as the unadjusted losses, but with a different mean and/or variance. Manipulating these two parameters will generate a range of corresponding utility functions.
[16] For an example, see Meyers and Schenker [1983].

insurance risk models.

Appendix B develops a model of losses based upon a *stochastic mean*, where the expected value of the loss per policy is itself a random variable. For example, suppose the unconditional mean loss per policy is 1000, with a standard deviation of 300. This mean is considered a random variable: the 1000 amount is multiplied by a separate random variable with a mean of 1 and a 0.10 standard deviation. As the number of policies becomes large, the average policy will still have an expected loss of 1000, but will take on the risk of the stochastic mean variable, so it will have a standard deviation of 100 = (0.10)(1000). The influence of the original per-policy standard deviation of 300 vanishes. Thus, beyond a certain point the size of an insurer has little influence on the risk characteristics that determine the value of default.

4. A ONE-PERIOD PREMIUM MODEL

This section develops a basic insurer model and determines the premium, which includes the cost of holding capital. Extensions to the model are discussed in sections 6 and 7.

4.1 A Basic Insurer Model

To understand how optimal capital values can be determined, in this section I establish a simple, bare-bones model of an insurance company containing only a few necessary components.

I have assumed that the insurer's policyholders have the same individual loss distributions and the same risk aversion. Using the section 3.6 framework for multiple policies, this homogeneity implies that we can analyze portfolios of risks (even entire insurance companies) as if they were insurers having only a single policyholder.

In this model, I eliminate extraneous variables such as expenses, income taxes and investment returns by assuming that they are zero. In section 5, I include these components.

With no investment return, all assets are cash. I assume that the insurer is operating efficiently and thus the insurer's costs of holding capital can be passed on to policyholders as long as it improves their welfare. In fact, the *policyholders* determine the amount of capital and then pay for its associated costs through their premium, denoted by π. Thus, the owners of the insurer are indifferent to the amount of capital actually held by the insurer, since they are fairly compensated for its use.

The model is one-period: the premium and capital are determined at the beginning of the

period and the actual[17] loss is determined at the end of the period. Further assume that there is no secondary insurer default protection for policyholders, such as a guaranty fund.

Its owners capitalize the company with an initial capital *C*. The initial assets of the insurer equal the premium from the policyholders plus the capital. Prior to the end of the period, the insurer's cost of holding its capital is expended, so that amount is not available to pay the PH's claims. The ending asset amount, denoted by *A*, thus equals the initial capital plus the premium, minus the capital cost. The loss amount is recognized at the end of the period and the default amount (if any) is determined accordingly.

If there is no cost to the insurer for holding the capital in the company, then the insurer will hold enough capital to exceed the largest possible loss, and thus there is no possibility of insurer default. In this case, the insureds simply pay a premium $\pi = L$, the expected value of the losses. With no default, their consumer value is maximized at an amount $\hat{L} - L$, the CEL minus the expected loss.

4.2 The Cost of Holding Capital

There is a cost to an insurer for holding capital to mitigate default risk. This cost is separate and distinctly different than the "cost of capital," which is the return expected by the capital suppliers (e.g., equity holders or bondholders) and is commensurate with the risk borne by these investors. To avoid the confusion created by the similar terminology, a useful name for the cost of holding capital is the *frictional cost of capital*, as defined by Hancock *et al* [2001]. The frictional capital cost (FCC) is the opportunity cost that accrues to the use of capital in an insurance firm, and which the investor would not incur if investing directly in financial markets. These costs include double taxation, financial distress, agency and regulatory restriction costs.

The primary component of the FCC for U.S. insurers is double taxation.[18] Of the above FCC components, it is also the easiest to determine empirically. To illustrate, assume that an investor provides $100 of equity capital to an insurer, whose corporate income tax rate is 30%. The insurer invests the $100 in assets *A* with an expected return of 6%. At the end of one year the expected return on the assets, after taxes, is $4.20 = 100(0.06)(1 – 0.30). This amount is returned to the investor as a capital distribution, giving a net return on the capital of 4.2%. On the other

---

[17] For regulatory and other external party uses, a multiple-period model must address the fact that the insurer might not use an unbiased estimate of losses. Accordingly, the risk of under-reserving must also be assessed, with additional capital required beyond what is needed for the pure loss-variation risk addressed in this paper.

[18] See Harrington [1997].

hand, if the investor had invested directly in the assets *A*, rather than through the insurer, he/she would have received an expected return of $6.00.

The $1.80 difference must be made up by charging policyholders through additional premium. But the extra premium itself is taxed at the 30% corporate tax rate, so it must be grossed up to $2.57 = 1.80/(1 – 0.30). So, the double taxation component of FCC, as it applies to premium, equals $[rt/(1 – t)]C$, where *r* is the insurer's investment return,[19] *t* is its income tax rate and *C* is the capital as defined in section 4.1. Notice that if the investment return is zero, then the double-taxation component of the FCC is also zero.

I assume that the FCC is at least equal to the above double taxation cost, and that all FCC components are proportional to the capital amount. Let *z* denote the FCC per unit of capital. Since the FCC must be borne by the policyholders in order to provide a fair market return[20] (the cost of capital) to investors, the premium must include an amount *zC* in addition to expected losses and other insurance expenses. Therefore, the *basic premium model* is

$$\pi = L + zC.$$

(4.21)

4.3 Fair Premium With Default

Since the frictional costs of capital must be passed on to PHs, they will not want the insurer to carry unlimited capital. Therefore, the insurer will have a non-zero probability of becoming insolvent. Then, in order to be actuarially fair, the basic premium (*L* + *zC*) must be reduced by the expected value of default *D*:

$$\pi = L - D + zC.$$

(4.31)

---

[19] In a competitive market, the investment return for the FCC will tend to equal the risk-free interest rate, despite the insurer's own expected return on investments. A higher return corresponds to greater risk and therefore requires a greater return to shareholders. Similarly, a high FCC cannot be passed on to policyholders if other insurers have lower investment returns and charge a smaller premium amount to cover double taxation costs.

[20] The expected return to investors, or the traditional "cost of capital" is built into the profit margin, another component of the fair premium. For simplicity, I have ignored it in the premium model. The profit margin can be directly embedded into the loss value by taking its present value at a risk-adjusted interest rate. The fair premium (with no default or expenses) will then equal the risk-adjusted PV of the expected loss plus the PV of the frictional capital costs. The cost of capital doesn't directly enter into the premium calculation.

So now the premium has three components: the base amount L is increased by the FCC and is reduced by D. Although, as I will discuss in section 5.1, the fair premium is approximated in practice by the basic premium. However, the basic model is not a competitive equilibrium[21] model, where the premium is a market-clearing price. Since I have assumed a zero interest rate and that market risk is captured by the loss value L, the market expected return on capital is zero. For equilibrium to occur, the premium must provide investors a zero expected return. Under the basic premium model, the expected return is D, since the policyholders' loss is the investors' gain. In a competitive market, this gain is reduced to zero by decreasing the basic premium by D. Thus, the fair premium satisfies both policyholders and investors, representing an equilibrium result.

---

[21] See Varian [1992], page 219.

5. DETERMINING OPTIMAL CAPITAL

Sections 3 and 4 have provided the ingredients to determine an insurer's optimal capital: we have a model for the value of policyholders' default as well as a specification for the insurer's cost of carrying capital. Since the capital amount governs the default value, we can balance these two factors to maximize consumer value.

5.1 General Model

For simplicity, I initially assume that the insurer does not deduct[22] the expected default from the premium, so we have the basic model $\pi = L + zC$. Since I have assumed that the interest rate is zero, the frictional capital cost rate z does not contain an income tax component (as discussed in section 4.2).

The value to the policyholders of their insurance is the certainty equivalent of the insured losses minus the cost of the insurance (which is a certain amount). Because insolvency is possible, the CE value of the actual coverage is the certainty-equivalent expected loss minus the certainty equivalent of the expected default. The consumer value (denoted by V) of the insurance transaction,[23] therefore, equals the certainty equivalent of the covered losses minus the premium:

$$V = \hat{L} - \hat{D} - \pi. \tag{5.11}$$

As we increase the amount of assets (by adding capital), the CE value of expected default decreases, while the premium (through the capital holding cost zC) increases with capital. This situation is a classic economics optimization problem, which can be solved by taking the derivative of V with respect to assets and setting it to zero.

Since $\hat{L}$ is constant with respect to a change in assets, taking the derivative of V with respect to A in 5.11 and setting the result to zero gives $-\partial \hat{D} / \partial A = \partial \pi / \partial A$. From Appendix C, which applies the derivative of an integral, we have $\partial \hat{D} / \partial A = -\hat{Q}(A) = -\int_A^\infty \hat{p}(x)\,dx$, so the general

---

[22] This is certainly the case in practice for the U.S. with regard to an *explicit* premium component for default. However, it can be argued that weaker insurers (with higher expected default amounts) will charge a lower premium to remain competitive.

[23] If the CE loss distribution is derived from some utility functions (such as the square root model), the CE value of a constant plus a random variable is not equal to the constant plus the CE of the random variable. In this case, equation 5.11 is an approximation. However, it is exact for CE values derived from exponential utility.

condition for optimum assets is

$$\hat{Q}(A) = \partial\pi / \partial A. \tag{5.12}$$

Here, $\hat{Q}(A)$ is the *adjusted ruin probability* (ARP), or the chance that losses exceed assets, under the transformed density $\hat{p}(x)$. The corresponding *unadjusted* ruin probability under $p(x)$ is denoted by $Q(A)$.

For the *basic premium* model, where the premium excludes the expected default, we have $\pi = L + zC$. Since L is constant with respect to assets, the derivative becomes $\partial\pi / \partial A = z(\partial C / \partial A)$. The assets available to pay claims equals the initial capital C plus the premium, minus the frictional capital cost, so we have $A = C + (L - zC) - zC = C + L$. Thus, $\partial C / \partial A = 1$, with the result

$$\hat{Q}(A) = z. \tag{5.13}$$

This rather simple result establishes that the optimal level of assets (and thereby capital) is determined by a risk measure that is an *adjusted ruin probability*. The ARP is a function of the probability distribution of losses and the policyholder risk aversion, as incorporated into the transformed density function. The risk measure is calibrated to the frictional capital cost rate z.

Given the optimal asset level from equation 5.13, the optimal capital is readily found by using the above relationship $A = L + C$.

For a *fair premium*, where the insurer deducts the expected default from the premium, we have $\pi = L - D + zC$ and $A = L + C - D$. Taking derivatives of these two expressions, and noting that $\partial D / \partial A = -Q(A)$, we get $\partial\pi / \partial A = Q(A) + z(\partial C / \partial A)$ and $\partial C / \partial A = 1 - Q(A)$. Thus, $\partial\pi / \partial A = z[1 - Q(A)] + Q(A)$ and equation 5.12 gives

$$\theta(A) = \frac{\hat{Q}(A) - Q(A)}{1 - Q(A)} = z \tag{5.14}$$

as the condition for optimal capital.

From section 3.5, tables 3.5a and 3.5b show that the transformed ruin probability $\hat{Q}(A)$ is

much larger than the unadjusted ruin probability $Q(A)$, particularly at lower ruin probabilities (which correspond to the high safety levels that would be required in practice). Thus, $Q(A)$ can be set to zero and equation 5.13 may be considered as an approximation[24] to the optimal capital condition for a true fair premium.

Notice that the level of insurer expenses doesn't affect the optimal capital, as long as the expenses are a function of the expected losses and not capital. Let the premium be $\pi = L - D + zC + e_0 + e_1 L$, where e_0 and e_1 are constants that determine expenses. The derivative of V in equation 5.11 will be the same with or without expenses, since the derivatives of L, e_0 and e_1 with respect to capital are all equal to zero.

Equation 5.12 is general in scope and can be used for alternative premium formulations. For example, if the frictional capital cost zC is not consumed prior to default, then the optimal capital is determined from $\hat{Q}(A) = z / (1 + z)$, which approximates equation 5.13.

If an insurer's policyholders have *heterogeneous risk preferences*, (but with the same loss distribution) the optimal capital can still be calculated using equation 5.11. However, the premium for each PH will differ. For a given capital amount, each PH will have a specific CED value (based on their risk aversion) along with a share of the joint capital cost. The share is allocated to each PH via their willingness to pay (determined from their respective consumer values). This gives a higher premium for the more risk-averse PHs: in effect, they pay the low risk-aversion PHs in order to use a high capital amount. However, this result is theoretical, since normally an insurer does not charge different premiums for policyholders with identical loss characteristics. Accordingly, in practice the optimal capital for a group of insureds must be based on an *average* (weighted in some fashion) of their risk preferences. In this case, low and high-risk aversion PHs will have consumer values that are less than the theoretical optimum.

Consequently, they might improve their consumer values by moving to an insurer whose other PHs have similar risk preferences to their own; i.e., to the extent that insurers incorporate capital costs into their premium, PHs are best served by choosing an insurer whose capital strength suits their needs. In finance, this grouping behavior is called the *clientele effect*. For insurance, it has implications for pricing (section 7.1) and regulation (section 7.2). However, further analysis of this topic is beyond the scope of this paper.

---

[24] Using $z = 2\%$, and the normal-exponential model from section 3.5 (with risk aversion of 0.02), the optimal capital under equation 5.13 is 379.73. Under equation 5.14 it is 379.56, a difference of only 0.045%.

5.2 The Effect of Income Taxes

Another variation to the premium model includes the effect of income taxes, which, as discussed in section 4.2, can be the major component of the frictional cost of holding capital. In order to pursue practical applications, the impact of income taxes on optimal capital must be addressed. Appendix D develops the result for optimal capital in this case. As in the general model with a constant frictional cost of capital, the optimal asset amount is found by setting the CE ruin probability equal to a constant value:

$$\hat{Q}(A) = \frac{rt}{1 + r - t}. \tag{5.21}$$

Here, *A* is the end-of-period assets (before subtracting the loss and income taxes), *t* is the income tax rate and *r* is the riskless investment rate of return.

Equation 5.21 is important because it establishes a benchmark for practical applications. The effective corporate income tax rate for insurers will be less than the current nominal 35% highest marginal rate, due to the ability to defer taxes on capital gains, shelter income using municipal bonds and other measures. Assume that the effective rate is 30%. As discussed in section 4.2, the appropriate investment benchmark is the Treasury rate (which should be matched to the average liability duration: about 3 years for U.S. property-casualty insurers). The 3-year rate has varied from about 1.5% to 6% over the past 10 years. Consequently, the optimal adjusted ruin probability has been in the range of about 0.4% to 1.8% over this period.

The corresponding optimal *unadjusted* ruin probabilities can be smaller than the 0.4% to 1.8% range, as indicated in the section 3.5 examples (tables 3.5a and 3.5b). However, several factors (e.g., using a more realistic loss distribution,[25] incorporating the regulatory constraints in section 7 and including the effects of guaranty funds) will increase the unadjusted ruin probabilities. Consequently, the above range of frictional capital costs appears to be broadly consistent with more subjective solvency measures such as the Solvency II 99.5% VaR standard, which translates to a 0.5% unadjusted ruin probability. In other words, the overall required capital for

---

[25] The normal and lognormal distributions that I've used for illustration have unlimited losses. In practice, insurance coverage imposes policy limits. Consequently, policyholders must absorb the high end of their losses, which carry the greatest CE values. This effect reduces the CE default value and increases both the ARP and the unadjusted ruin probability corresponding to the optimal capital.

an average insurer under the ARP risk model is not inconsistent with current practice.

5.3 Numerical Examples

Here, I've used the normal distribution with exponential utility from section 3.5. For the same parameters (1000 mean, 100 standard deviation, 0.02 risk aversion) and with a capital cost rate z = 0.05, we get the optimal ARP of 5% and optimal capital of 330.66 using equation 5.13 and equation A4.9 in Appendix A4. The CE expected loss is 1100 and CED value is 2.46 from equation A4.8. The premium is 1016.53 = 1000 + 0.05(330.66), and so the consumer value of the insurance contract is 81.00 = 1100 – 1016.53 – 2.46. Figure 5.3 below shows how the consumer value of the insurance varies by amount of capital.

Figure 5.3
Consumer Value of Insurance by Capital Amount in Numerical Example
Normal Loss/Exponential Utility Model
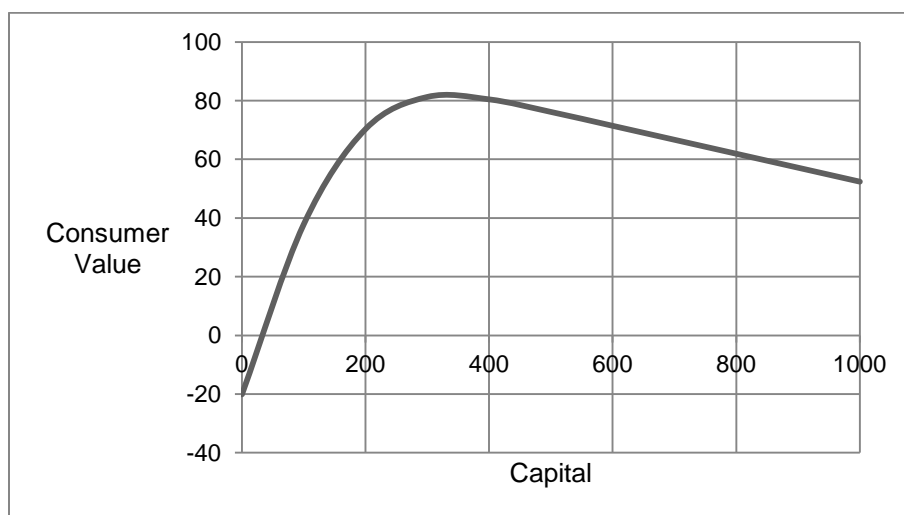
Notice how the shape of the CV curve is steep at low levels of capital and flattens with higher capital. This behavior indicates that, beyond the optimal capital level, adding more capital has only a slight impact on PH welfare. Thus, the relative insensitivity of the optimal capital might be exploited in practical applications, where it could be necessary to use approximate values for some of the underlying variables.

Table 5.3 shows how the optimal capital varies by standard deviation (SD) and risk aversion:

Table 5.3
Optimal Capital by Standard Deviation and Risk Aversion in Numerical Example

Risk Aversion	SD = 25	SD = 50	SD = 100	SD = 200
0.005	44	92	205	493
0.010	46	103	247	661
0.020	51	123	331	1007
0.040	62	165	504	1720
0.080	83	252	860	3186

As we would expect, the optimal capital increases with both risk aversion and the volatility of the losses.

5.4 Comparison to Other Risk Measures

The conventional solvency risk measures can be considered as equal to or as simple functions of the *tail moments* of the loss distribution. Here I define the *n*th tail moment as

$$MT(n) = \int_A^\infty (x - A)^n p(x) dx, \qquad (5.41)$$

where $p(x)$ is the density function and A is the assets, as defined earlier. Notice that if assets are zero, the tail moment equals the regular moment of the entire distribution.

Observe that the ruin probability is the 0th tail moment and the expected default (policyholder deficit) is the *first* tail moment. Define the *valuation level* as the predetermined numerical value of the tail moment, such as 1% or 5%, that produces the desired level of assets. In other words, if the risk measure is ruin probability (RP) and the valuation level is 1%, then $MT(0) = 0.01$ and we solve equation 5.41 for A.

VaR is the amount of assets such that $\alpha = 1 - MT(0)$ or 1 minus the RP, where α is the VaR *confidence level*. Tail value-at-risk, or TVaR,[26] is the amount of assets equal to VaR + $MT(1)/MT(0)$ at the α confidence level. Thus, the conventional risk measures are simple functions of the tail moments with n equal to 0 or 1.[27] In the following discussion, I use the ruin probability and the expected default ratio to loss (*D/L*) to characterize the tail-moment based

---

[26] Notice that TVaR and the EPD are not equivalent and will not necessarily produce the same capital amount. EPD includes only the amount of loss *exceeding* the asset threshold A, while TVaR also includes the portion of the loss *below A* for losses exceeding A.
[27] The value of *n* need not be an integer. For example, with $n = 0.5$, the weight of the tail losses will be somewhere between that of a ruin probability and an expected default measure.

solvency measures.

It is noteworthy to compare the adjusted ruin probability (ARP) measure to the common tail-moment risk measures. Assume that both a straight (unadjusted) ruin probability (RP = 1 – VaR) and an unadjusted EPD measure are also used to determine capital. Further assume that all three measures (ARP, RP and EPD) are calibrated to give the same capital for a typical insurer. We observe the valuation level for each risk measure implied by the capital and keep it fixed as we change the variance of the loss distribution.

For this exercise, let the typical insurer have the characteristics of the section 5.3 example, with a 5% ARP providing optimum capital of 330.66. This capital amount implies that RP = 0.047% and the EPD/Loss ratio = 0.0012%. We fix all three measures and consider two other insurers, also having normally distributed losses. One insurer has low-risk policyholders with a standard deviation (SD) of 50; the other has high-risk PHs with a 200 SD. We apply each risk measure to these insurers, and compare to the typical insurer:

Table 5.4a
Comparison of Risk Measures Calibrated to Normal SD = 100
Adjusted Ruin Probability = 5%

SD	Optimal Capital	Implied RP Using Optimal Capital	Implied EPD Ratio Using Optimal Capital	A Capital: RP Standard	B Capital: EPD Standard	A % Diff.	B % Diff.
50	123.3	0.685%	0.0112%	165.3	156.0	34.1%	26.6%
100	330.7	0.047%	0.0012%	330.7	330.7	0.0%	0.0%
200	1007.0	0.00002%	0.000001%	661.3	697.0	-34.3%	-30.8%

Notice that the low-risk insurer has more capital under RP and EPD, with the high-risk insurer having less capital, compared to the optimal capital under ARP. Because the resulting capital for the low and high-risk insurers is not optimal, using a conventional risk measure reduces *consumer value*. For the high-risk insurer, using RP (or its VaR equivalent) lowers the CV from 344.07 to 311.79, a decrease equaling 3.2% of the expected loss. For the low-risk insurer, the CV drops from 17.70 to 16.60, a reduction of only 0.1% of the expected loss. Similarly, applying the EPD ratio reduces CV by 2.5% of expected loss for the high-risk insurer and 0.1% for the low-risk insurer. Note that this disparity between the high and low-risk insurers is due to the fact that,

for the same level of risk-aversion, high-risk policyholders gain more consumer value from insurance than low-risk PHs.[28]

   We get similar results with a lognormal[29] distribution paired with exponential utility. Here the typical insurer again has a mean loss of 1000, standard deviation of 100 and a risk aversion of 0.02. The initial calibration gives RP = 0.037% and the EPD ratio = 0.0013%. Varying the SD and keeping the risk measures constant, we get similar results to the normal case. Figure 5.4 shows the loss in CV/ expected loss for the lognormal case compared to the optimal ARP capital. For comparison, it also provides the results for the normal distribution.

Figure 5.4
Loss of Consumer Value by Using RP and EPD Measures
Compared to Optimal ARP Result; Percentage of Expected Loss
Normal and Lognormal Distributions
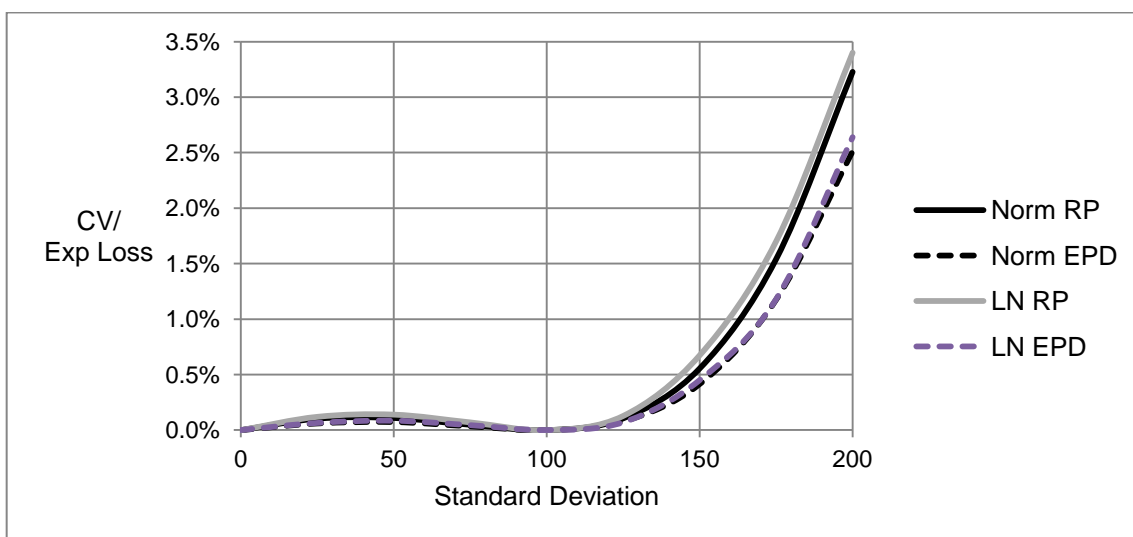
   For both distributions, the CV loss is slightly less (for the high-risk PHs) using the EPD, compared to the RP measure. To summarize the above results, we see that, compared to the ARP

---

[28] The CE value of the loss (being related to the variance) is lower with respect to its expected value as variance decreases. At the extreme, with a zero variance, there is no difference between the CEL and the expected loss and so the difference in CV from using any risk measure must be zero.

[29] Here I've approximated the lognormal distribution using the binomial option pricing method as described by Panjer [1998], page 246. The CE ruin probabilities and default values are directly determined using the method of Appendix A4. Also, I have adjusted the risk aversion parameter to produce, for each standard deviation value, the same CE loss as with the normal distribution.

standard, the conventional risk measures overstate the optimal capital for low risk lines and *understate the capital for above-average risk lines*. The latter effect is more serious, because the under-capitalization can produce a meaningful loss in consumer value. For the low-risk lines, the loss in CV appears to be negligible.

To gain further insight regarding the deficiencies of the conventional measures, assume that we have calibrated all three risk measures to produce the same capital for an insurer, as in the above example (e.g., Table 5.4 with a 100 standard deviation). Now we increase the size of a particular possible loss $x > A$ in the tail by an amount Δ.

If RP or VaR is the standard, there is no change in capital since the loss increment does not affect the RP. Yet policyholders are worse off. This effect was exploited in a perverse way during the 2008 financial crisis, when financial firms "stuffed the tails" to keep their apparent risk low (see section 2.1). To manipulate the tail probabilities, the firms designed securities to have a low probability of loss, but with an extreme loss size when the loss occurred.

Under an EPD or TVaR standard, the effect of a loss increment is more subtle. Suppose that we take two "slices" of the tail, one with smaller losses and the other with larger losses. The widths of the slices (i.e., the probability that the losses in the intervals will occur) are selected so that the probability of losses being in each interval are equal. Let x_1 be the average loss in the lower interval and x_2 be the average loss in the upper interval. If we simultaneously adjust losses so that x_2 increases by Δ and x_1 decreases by Δ, the expected default amount remains the same (as does the default probability). Note, however, that this operation increases the variance of the tail losses. Since the certainty equivalent value function $k(x)$ is concave upward, we have $k(x_2) > k(x_1)$, and thus this adjustment will increase the CE value of the default. Therefore, more capital is needed, even though the EPD stays constant.

The preceding discussion shows that, given a particular loss distribution, under the ARP method it is not possible to "engineer" the tail to produce an artificially low capital amount. If the tail is altered by reinsurance or other financial techniques, the CE value function will automatically produce the proper capital as long as the firm uses the correct loss distribution (adjusted for PH risk preferences).

A practical disadvantage of using the ARP measure is that it does not translate to any fixed conventional standard. For example, to get the correct optimal capital under the 5% ARP standard, the appropriate unadjusted RP in table 5.4 ranges from 0.00002% to 0.685%. The EPD

ratio has a similar large range. Although this ARP feature presents no difficulty in calculating capital, it may create problems in comparing results to conventional solvency measures.

5.5 Subadditivity

To clarify risk definition in financial economics, theoreticians have described several properties for a risk measure. A *coherent risk measure*[30] is a function that satisfies monotonicity, subadditivity, homogeneity, and translational invariance. *Subadditivity* has become the most important of these properties when applied to risk measures used in practice.

The subadditivity (SA) property requires that the value of the risk measure for the combination of two risks is less than or equal to the sum of the risk measure values taken separately. For insurance capital requirements, this means that when two risks (or risk portfolios) are combined, the required assets derived under the risk measure must be less than or equal to the sum of the assets derived from applying the risk measure to the risks individually. To be consistent with the individual PH focus of our analysis, the subadditivity requirement can be restated: if two risks are combined, the assets *per risk* under the risk measure cannot be greater than assets for either risk taken separately under the risk measure.

Assume that two PHs have identically distributed losses. Let $RM_1(A)$ represent a risk measure that is a function of assets per policyholder A for a PH of an insurer with one risk and $RM_2(A)$ for a PH of an insurer with two risks combined. For two different asset amounts $A_2 > A_1$, we have $RM_1(A_2) \leq RM_1(A_1)$ and $RM_2(A_2) \leq RM_2(A_1)$; i.e., increasing assets decreases the value of the risk measure (this is the monotonicity property of a coherent risk measure). A *subadditivity violation* will occur when

$$RM_2(A) > v > RM_1(A), \tag{5.51}$$

where v is the valuation level of the risk measure and A is any asset amount. Under this inequality, a SA violation occurs because, for both measures to equal v, assets in the combined-risk insurer must *increase* and assets in the single-risk insurer must *decrease* (this effect is shown graphically in figure 5.51). Therefore, assets per PH in the combined-risk insurer will be greater than for the single-risk insurer.

---

[30] See Artzner [1999] for a discussion of coherent risk measures with insurance applications.

A classic bond risk example,[31] which has a *binary loss distribution*, is used to illustrate SA violation. I have modified this investment illustration to represent insurance by using the section 3.6 example with two independent binary risks each with a 2% probability of a $1000 loss. Suppose that the VaR measure is set at 97%. This means that a single risk must have at least a 3% chance of loss in order to require assets (and thereby capital). Otherwise no assets are required to back the loss. Thus, with a 2% chance of loss, no assets or capital are required. However, if two independent risks are combined, the probability of a loss is 3.72% = 2(0.02)(0.98) + 0.02(0.02) and therefore $1,000 of total assets (for both PHs) is required. This reduces the default probability to 0.04% and satisfies the 97% VaR valuation level. Subadditivity is violated here since more assets are required per PH for the combined risks ($500 each) than for either of the separate risks (zero).

Using the ruin probability counterpart to VaR as the risk measure, we have $v = 0.03$. Let $Q_1(A)$ denote the RP for the single risk and $Q_2(A)$ the RP for the combined risks. Applying equation 5.51, we see that for A (assets per PH) from 0 to $500, $[Q_2(A) = 0.036] > 0.03 > [Q_1(A) = 0.02]$. Thus, subadditivity is violated. For assets above $500, we have $0.03 > [Q_1(A) = 0.02] > [Q_2(A) = 0.004]$, so there is no SA violation. Also, for a valuation level $v > 3.68\%$ (required assets are zero for both the single or combined risks) or $v < 2\%$ (required assets are $1000 per PH for the single risk and $500 for the combined risks), there is no SA violation.

For the *adjusted ruin probability* measures $\hat{Q}(A)$ or $\theta(A)$ in equations 5.13 and 5.14, there also is SA violation for a range of valuation levels. Applying equations A5.4 and A5.6 in Appendix A5 to the parameters for this example, we get Figure 5.51, which graphically shows the pairs $\hat{Q}_2(A)$, $\hat{Q}(A)$ and $\theta_2(A)$, $\theta(A)$. These are labeled respectively as Qhat 2, Qhat, Theta 2 and Theta (for simplicity, I have dropped the subscript 1 denoting a single risk).

---

[31] This example is from Albanese [1997] and has been used by Artzner and others.

Figure 5.51
Adjusted Risk Measures for Section 3.6 Example
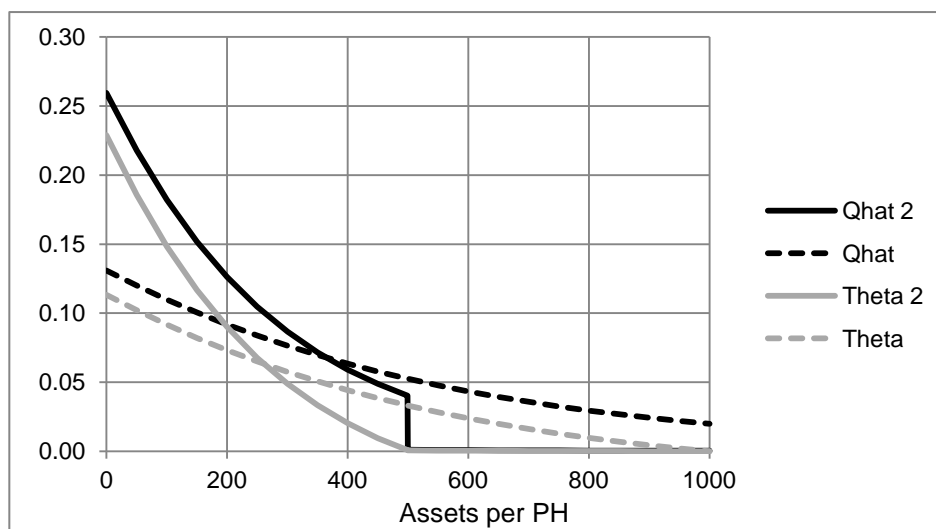
Notice that a risk measure based on PH risk preferences can be a *continuous* function of the asset amount, even if the underlying loss distribution is discrete. This occurs because the values of the CE expected default and its derivatives are continuous with respect to the amount of assets if the underlying CE function $k(y)$ or the equivalent utility function is continuous.

The graph shows that $\hat{Q}_2(A) > \hat{Q}(A)$ for $\hat{Q}(A) \geq 6.78\%$, where assets per PH are less than 364.44. Subadditivity can be violated in this region. Similarly, $\theta_2(A) > \theta(A)$ when $\theta(A) > 6.34\%$, corresponding to $A < 260.80$.

Does the SA violation create any negative effects for the policyholders in this example? To answer this, suppose that the frictional cost of capital is $z = 7\%$, which is the valuation level for the adjusted risk measures. It exceeds the above critical values of 6.78% and 6.34%, so there will be a SA violation for each measure. From section 5.1, we have the premium and from section 3.6 the CED values by asset amount. Thus, the consumer value for any asset value can be readily found.

For a *fair premium*, when assets are zero, the expected default equals the expected loss of $20. The premium and capital are also zero. Since the CED equals the $60.13 CE loss, the consumer value of the insurance is zero. When assets equal the $1,000 loss value, the expected default is zero, but the premium equals $L + zC$ and capital equals assets minus the premium (the capital cost zC is not an available asset to pay losses). Thus, the capital is $980 and the premium is

$88.60 = 20 + 0.07(980)$. The consumer value is negative: $-\$28.47 = 60.13 - 88.60$. Between these asset value extremes, the CV will have an optimal value. Figure 5.52 shows the per-PH CV by assets for a single risk and for two combined risks.

Figure 5.52
Consumer Value per PH by Asset Amount for Single Risk and Two Risks
Fair Premium Model; z = 7%
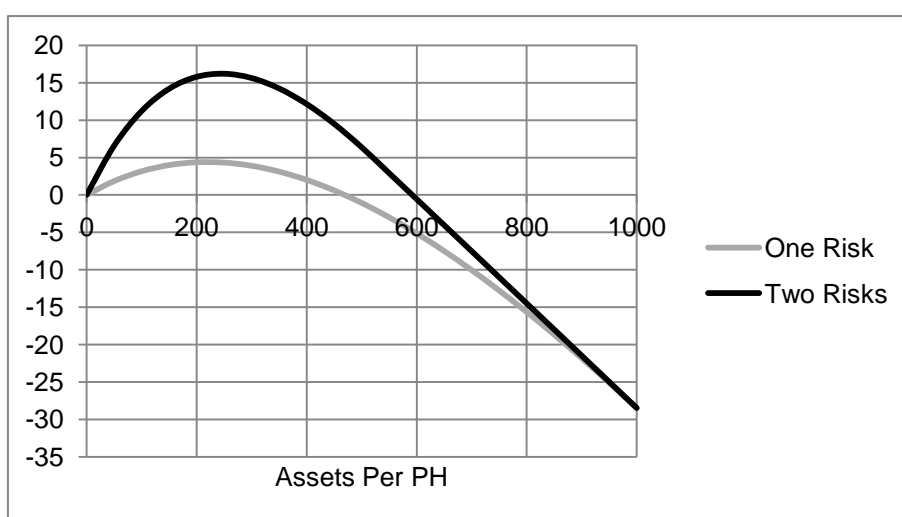
Here, for both cases, the optimal CV is positive. This is achieved with a per-PH asset value of $219.51 for one risk and $244.59 for the two combined risks. The corresponding respective optimal capital amounts are $215.12 and $234.90. These optimal amounts are derived directly by solving for $\theta(A) = 0.07$ in equation 5.14, using the Appendix A5 relationships.

This example clearly shows that the subadditivity criterion is violated, since *more* assets (or capital) are required per policyholder for the combination of two risks than for the single risk. It is also clear that the PHs are *better off* with the SA violation under the ARP risk measure, since their optimal consumer value is higher when the risks are combined.

For the *basic* (non-fair) *premium* case, where the risk measure is $\hat{Q}(A)$, we get similar results, with the optimal assets for a single risk being $347.43, which is less than the $356.15 optimal per-PH assets for the combined risks. However, since the premium is not actuarially fair, the CV is lower than for the fair premium case, for both the single risk and the combined risks:

Figure 5.53
Consumer Value per PH by Asset Amount for Single Risk and Two Risks
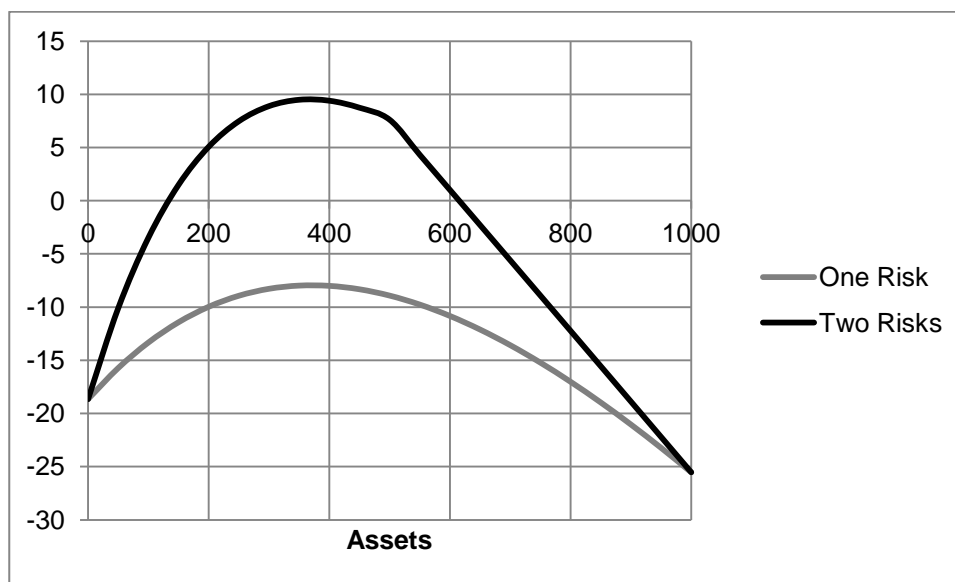*Basic (Non-Fair Premium) Model; z = 7%*

Notice that the CV for the single risk here is negative for all asset values, indicating that the risk is not insurable — the PH is better off without insurance. However, when the two risks are combined, they become insurable.

With lower values of z in this binary loss example, subadditivity is not violated. For instance, with fair premium and $z = 2\%$, optimal per-PH assets for the single-risk insurer are \$648.35, compared to only \$401.65 for the two-risk insurer.

In the financial economics literature an economic justification for the subadditivity constraint is that "if a firm were forced to meet a requirement of extra capital that did not satisfy this property, the firm might be motivated to break up into two separately incorporated affiliates, a matter of concern for the regulator."[32] But the above examples show that PHs are clearly better off being combined — with the consequent subadditivity violation requiring extra capital — than being insured separately. In fact, for the basic model, violating subadditivity turns uninsurable risks into insurable ones.

---

[32] See Artzner [1999], page 14.

Nevertheless, given that a major purpose of the subadditivity constraint is to promote aggregation of risks, the underlying economic basis of the adjusted risk measures used here will always indicate that PHs are better off[33] (or no worse off, if the risks co-vary) when risks are combined. Therefore, these measures promote the *spirit* of the SA constraint.

To summarize this section, I have shown that risk measures based on PH risk-preference can violate subadditivity, but when they do, the result makes perfect sense economically. Further, policyholders are never worse off when risks are combined — a fact that does not depend on the risk measure used to determine capital. Therefore, we must conclude that *subadditivity is an unnecessary criterion for an insurance solvency risk measure.*

6. EXTENSIONS OF RESULTS

The analysis in the preceding sections is based on a simplified model of an insurer, and concentrates on estimating optimal capital for insurance losses only. I have omitted some important elements that must be addressed before implementing the concepts for regulation, internal insurer risk management, pricing or other applications.

This section discusses some of these important missing pieces. The scope of this paper does not permit a full development of the topics, so for each of them I have stated the issue and outlined the general direction of the analysis. Although these areas present some difficulties, they can be attacked using the major idea of section 5: optimal capital can be determined by trading off the cost of holding the capital and the value to policyholders of having the capital.

6.1 Asset Risk

The treatment of asset risk adds another dimension to the section 3 formulation of default risk for losses, where I assumed that assets were riskless, with a zero return. Now assume that the insurer has a portion of its investments in risky securities. For initial assets of A_0, the ending asset value will be random, with an expected value of

$A \geq A_0$ (the reward for bearing market risk is an expected return exceeding the risk-free rate). Further assume that the insurer keeps the total asset and loss risk constant, so that it varies its asset (and capital) level by changing the amount of riskless assets.

---

[33] For the basic premium model, at any per-PH asset level A, the difference between the combined-risk CV and the single-risk CV equals the difference between the single-risk CED and the combined-risk CED. This is so because the premiums for the two cases are identical. Thus the result will be non-negative. For a fair premium, the difference is the basic premium differential *minus* the difference between the unadjusted expected default amounts. This too will always be non-negative.

The certainty equivalent value of a risky asset is the converse of that for a risky loss: it is *less* than the expected value of its possible payoffs. Therefore, with a zero risk-free investment return, the beginning of period certainty-equivalent value of the ending asset amount equals the expected value of the ending amount, where the expectation is taken over an adjusted probability distribution. For an average investor, the expected value must equal the market value of the assets. Thus, this calculation *removes the expected return* from the ending asset values. The adjusted, or risk-neutral distribution (see Hull [2008]), reflects the concept that the *ex ante* perceived value of a particular asset outcome will depend on the *economic scenario* that generated the asset value.[34] For example, a low asset value may correspond to an unfavorable economy where a dollar is worth more; in that case the investor will weigh the result more heavily than the symmetrical high asset value.

Assume that the policyholder has the same risk preferences as the typical investor. For each ending asset value A_i, the CE expected default is $\hat{D}_i = \hat{L} - \hat{L}(A_i)$, where from section 3, $\hat{L}(A_i)$ is the CE amount of loss limited to A_i. Therefore, the unconditional CE of default over all asset values is the sum of the conditional values $\hat{D}_i$ weighted by the risk-neutral probabilities $p_A(A_i)$ for the asset values occurring:

$$\hat{D} = \hat{L} - \int_0^\infty p_A(x)\hat{L}(x)dx. \qquad (6.11)$$

Note that, if losses and assets are correlated, each $\hat{L}(A_i)$ will derive from a different expected loss corresponding to each A_i. For most applications using continuous distributions, the above integral can be evaluated with numerical techniques.[35] The result gives the CED for a particular initial asset amount A_0. The optimal capital occurs when $-\partial\hat{D}/\partial A_0 = z$.

To illustrate this calculation, I return to the normal-exponential example from section 5.3. Here, the expected loss is 1000 with a standard deviation of 100, the risk aversion is 0.02 and the capital cost rate is 5%. Assume that the insurer has 400 of risky assets, also with a normal distribution, with a 5% expected return and a SD of 80 (the *volatility*, or SD per unit of risky assets is 20%). The remaining assets are riskless, with a zero return. Also, asset risk is

---

[34] Note that a risky asset whose return is uncorrelated with market returns will generally not command a positive expected return above the risk-free rate.

[35] For the calculations here, I have used a discrete binomial approximation to the normally distributed ending assets. Each asset value determines the CE value of the loss limited to the asset amount using the conditional bivariate normal loss distribution (where the mean loss is a linear function of the asset/loss correlation). The overall CED was determined by inverting the expected utility of the conditional CE values using the risk-neutral asset probabilities.

independent of loss risk.

Assuming that the risk-neutral distribution is also normal, its expected value equals the initial asset amount.[36] The optimal capital with the risky assets becomes 356.89, which is 26.23 greater than the 330.66 for the case with riskless assets. The optimal CV at 78.88, is 2.13 lower than with the riskless assets.

Generally, under the policyholder welfare framework, the capital is always higher and the CV is lower when an insurer has a risky investment portfolio. Since on average, insurers' investment managers cannot beat the market, financial theory indicates that there is no benefit to PHs for holding the risky assets. In other words, the optimal investment portfolio has *only riskless assets*. Then why do insurers in practice have risky assets? To resolve this puzzle, there are several hypotheses, including: the insurers may believe that their investment managers can individually beat the market (although collectively they cannot); management compensation schemes reward positive income without penalizing negative income; and insurers may build an above risk-free investment return into their pricing models.

However, even though risky assets may lower the consumer value for PHs, the reduction may be small enough so that it is not material. For example, in the above calculation, the loss in CV is only about 0.2% of the expected loss. So the risky asset conundrum is theoretically interesting but generally may not be a practical issue. Regulators, rating agencies and insurance management recognize that a large amount of risky assets is imprudent. Especially large risky investment portfolios require additional capital whose costs cannot be passed on to PHs in a competitive market.

An important point to make here is that the risk-neutral probability distribution removes the positive expected excess market return from the CE default calculation. If an insurer increases its asset risk through securities whose return is uncorrelated with the market, the expected default will rise and more capital is required. Thus it is essential for an insurer to maintain a diversified investment portfolio.

Also worth observing is that both insurance losses and investment returns are not considered to be normally distributed; often a lognormal model or some other skewed distribution is used to approximate these variables. It may not be possible to represent the joint distribution in a tractable form. A more realistic application of the CE approach for total asset and loss risk will require a more elaborate method, such as a simulation model.

---

[36] If one does not reduce the expectation to the beginning asset level, the result can be an optimal capital amount that is *less* than that for the riskless assets. In this case the additional capital required for risky assets is *negative*.

6.2 Guaranty Funds

This topic deserves a full treatment in a separate paper. There is much academic literature[37] on the economic basis and design of guaranty funds, but I have found none that analyzes the effect of the funds on insurers' capital requirements.

Under the policyholder welfare concept, guaranty funds will substantially reduce the optimal capital for an insurer. To see why this is so, consider an economy in which all policyholders of all insurers are completely covered by a single guaranty fund (GF). Clearly, no policyholder will suffer an uncovered loss unless the entire industry defaults. Thus, the aggregate capital for *all insurers* is used to protect any individual policyholder. Contrast this situation with the opposite extreme, where no policyholder has GF protection. Here, the policyholder has access only to the capital of his/her own insurer. In this case, the insurer needs much more capital than in the full GF situation.

Under a GF within the U.S., essentially all of the capital for each insurer in a particular state is pooled to provide default protection for policyholders. The coverage is limited (usually $300,000 per policyholder for most lines of business), but some lines, such as workers compensation, have unlimited protection and others, such as surety, have no protection. For lines protected by the $300,000 limit, the GF coverage can vary significantly. For example, assuming a lognormal distribution with a 5.0 coefficient of variation, policyholders in a line with an average loss per policy of $1,000 (e.g., personal insurance) will have 99.53% of their expected losses covered by the GF, with only 0.47% exposure to the insurer's default. However, those policyholders in a line with an average loss per policy of $5,000 (say, commercial insurance) will only have 95.96% covered, with a 4.04% exposure. Relative to their expected loss, the ratio of non-covered losses for the two lines is 8.6 to 1. So, for this example the presence of GF protection is a major factor in assessing the optimal capital for the two lines.

Also, the GFs themselves can become exhausted[38] in extreme events, since there is an annual limit to the amount they can assess the solvent insurers. Thus, in order to estimate optimal capital for a particular insurer, the *risk of GF exhaustion* must be analyzed. When this threat is considered, a much higher portion of default risk becomes attributed to *extreme events*.

---

[37] Cummins [1988] is one of the most often cited references. He argues that a pre-funded GF system is superior to the predominant post-failure assessment model in current use. However, based on the analysis here, a properly constructed RBC implementation might produce equivalent results.

[38] The term guaranty *fund* is somewhat of a misnomer. The vast majority of the state GFs merely assess other solvent insurers; they have no "fund" to pay claims. Thus, the GFs themselves cannot become insolvent.

Consequently, modeling these becomes paramount. The extreme events can be national or world-wide in scope (e.g., a financial crisis or a deep pricing down-cycle) or regional (such as a natural catastrophe).

An important implication for analysis is that, with GF protection the optimal capital depends not only on the risk of a policyholder's own insurer's default, but also the *default risk of the other insurers* covered by the fund. Therefore, in analyzing the effect, say, of catastrophes on capital, one must also estimate the effect of the same catastrophes on the other insurers. This modeling might be simplified by using a *default correlation* parameter (DCP) for the insurer, where the parameter measures the correlation between the insurer's default and that of the remainder of the insurers in a particular state. A value of zero for the DCP would mean the insurer' capital can be modeled as a stand-alone entity. At the other extreme, a value of 1 would mean that whenever the insurer defaults, the GF is exhausted due to the simultaneous defaults of other insurers.

Although the protection afforded by a GF is considerable (the expected loss above a $300,000 threshold is a small fraction of the total expected loss), the certainty equivalent value of the above-threshold amount is large relative to its expectation. Consequently, the value of the GF protection for the policyholder is reduced somewhat in comparison to its straight expected value.

Other considerations in modeling the effect of GFs on optimal capital are that there may be a degradation of service (e.g., a delay in settlement) when, upon insolvency, a policyholder's claim is transferred to another claims management firm or that there may be market disruption from the insolvency of a large insurer. These effects can be incorporated into the model by modifying equation 5.11 to include a coefficient greater than 1 for the CE of the default.

The analysis of optimal capital under a GF should feature an additional term in the premium calculation: the expected GF assessment for the failure of other insurers. This is an unavoidable cost to the policyholder that is paid *ex post*, so its value is stochastic at the time of the policy is purchased. Note that, for a specific insurer, the expected GF assessment depends on the capital levels of the other insurers, so the optimal capital level for that insurer is influenced by both the GF assessment and the ability of the other insurers to provide GF protection for the insurer.

The presence of guaranty funds adds another element to the regulator's role of solvency protection. By monitoring capital for a particular insurer, the regulator must not only protect the interests of that insurer's policyholders, but also the interests of the *policyholders of the other*

insurers who would be assessed in the event of the particular insurer's demise.

Summarizing this section, GF protection adds two important variables to incorporate into optimal capital determination. The first is the degree of GF coverage, which varies by line of business. More capital is required for lines with less GF coverage. Second, the optimum capital for an insurer depends on the default risk of other insurers covered by the GF. Thus, the correlation of the default risk with other insurers will affect capital: the higher the correlation, the higher the capital amount. Including these variables requires analysis of an insurer's data by state: for example, to properly determine catastrophe risk capital, the effect of GF exhaustion must be estimated for each state where there is material exposure.

6.3 Multiple Periods

This topic is the subject of much debate in the actuarial and insurance finance literature.[39] There are two camps: one side advocates using an annual time horizon, wherein capital is only needed to offset default risk based on market values over the *upcoming year*. The other side argues that capital is needed to offset the risk that cash flow will not be sufficient to pay claims over the *entire duration* (the runoff horizon) required to settle the liability. This topic also deserves a separate paper, so here I have only outlined a procedure that will establish optimal multi-period capital.

The single period model, with the above and other extensions, should suffice to determine optimal capital for lines of business, such as property, whose claims are paid over a short duration. For liability insurance, workers' compensation and life insurance, we need to expand the model to encompass long-duration claims. Although the long-duration contracts can be modeled in continuous time, it makes sense to use a discrete, multi-period time frame. This is because accounting time frames determine the valuation of insurer assets and liabilities and hence capital. The annual time period is especially important, so for practical purposes, we need to examine long time-horizon asset and liability risk over one-year time increments. For shorter time periods (e.g., quarterly), a similar analysis will apply.

With long-horizon risks, we can use the same fundamental assumptions that drive optimal capital for a single period. The main point is that the optimal capital over several periods still depends on the balance between capital costs and the certainty equivalent value of default.

A key component of the analysis is that the value of a long-horizon risk element (e.g., losses

---

[39] For a good discussion of this topic, see Lowe et al. [2011].

or assets) is stochastic (i.e., random) at the end of every period. Assume that we know the probability distribution for the evolution[40] of the risk element value and there are *n* periods. Suppose the risk element is a loss with expected present value *L*. So, if the insurer begins the first year with an optimal capital level C_1 and the loss value at the end of the year happens to be $L_1 \neq L$, then the risk of default (either in the next period or ultimately) will change if the original capital remains the same. Thus, the capital must be changed accordingly to regain the optimal position.

This process leads to a sequence of capital amounts $\{C_1, C_2, \cdots C_n\}$ corresponding to the sequence of loss values $\{L, L_1, \ldots L_{n-1}\}$. It will also produce a series of CE expected default amounts at the end of each year driven by the loss values: $\{\hat{D}_1, \hat{D}_2, \ldots \hat{D}_n\}$.

For each of these sequences of value realizations, we can determine the present value of the consumer value. However, we need a rule or strategy to determine the capital amount at each period C_t (based on L_t) that optimizes the expected present value over *all possible realizations* of the $\{L, L_1, \ldots L_{n-1}\}$ sequence. If we find a single strategy that does this, then we have settled the issue of setting capital for a long-horizon risk element.

This type of problem can be solved by a process called *discrete time stochastic dynamic programming*.[41] One of the techniques used in this method is backward induction, where one starts at time *n* – 1, finds the optimal decision rule, then steps backward to time *n* – 2, finds the optimal decision rule at that stage, and so forth, all the way back to the beginning of the first period. If the stochastic process is regular (such a random walk with a constant drift), then the decision rule at each stage will likely be the same.

7. APPLICATIONS

The preceding sections have developed a theoretical framework for determining optimal capital for insurers. This section discusses several issues involved in applying the theory in a practical setting.

There will be some applications of these results that are not directly related to setting the level

---

[40] This process is analogous to a discrete model of interest rate evolution, where the value at any period will generate multiple possibilities for the next period. Graphically, the structure will look like a tree, with each successive period having more branches.
[41] See Birge and Louveaux [1997].

of an insurer's capital, such as capital allocation,[42] but I will leave those topics for further research. Also, since capital is an essential ingredient in pricing models, the optimal capital results will be relevant to that application;[43] however, this work is outside the scope of this paper.

7.1 Implementation of Results

The results here are new and somewhat contrary to current practice. In my view, there are three major obstacles to implementing them.

The first is that there is little empirical work, especially for insurance, in quantifying policyholders' risk preferences. All we know for certain is that insurance consumers are risk-averse, and will pay more than expected value for their coverage. In the absence of empirical evidence the best we can do is to assume a functional form for the risk aversion process, such as the exponential utility used in sections 3 through 5. (I do not necessarily advocate this model; I have used it because it is familiar and provides mathematically tractable results). This can be calibrated to a presumed certainty equivalent factor (*k*) for an individual PH based on judgment.

Second, the adjusted ruin probability risk measure is not as easy to understand as the conventional ruin probability measure. That fact that a constant ARP translates into different conventional ruin probability standards for different risk elements (e.g., lines of business) may be difficult for some to comprehend, and may undermine acceptance of the results.

Third, the analysis has unearthed several currently unrecognized variables (e.g., the frictional cost of capital, which reflects interest rates and income tax rates; also, the level of guaranty fund protection) that should be considered in setting capital. Incorporating them will require considerably more data-gathering and analysis than is presently done. Based on the analysis of sections 3 through 6, Table 7.1 shows the key variables that should be considered in establishing optimal capital levels:

---

[42] The Myers and Read [2001] capital allocation method uses an expected default that is not adjusted for policyholder risk preferences. Incorporating this element will allocate relatively more capital to lines with more risk-averse policyholders.

[43] Section 5.1 has shown that, to the extent that PHs tend to select insurers having capital levels based on their risk preferences, then the insurer's *actual capital* level (rather than an industry standard) will be relevant to setting prices.

Table 7.1
Summary of Key Variables for Setting Capital Requirements
N = None; L = Low; M = Moderate; H = High

	Inputs				
	Risk-free investment rate	Effective tax rate	Risk-aversion parameter	Fraction of default not covered by GF	Insurer-GF default correlation
Variation by:					
Time	M	L	N	L	L
Insurer	L	L/M	L/M	L/M	M
Line of Business	L	L/M	M	H	H
State	N	N	N	L/M	H
Econ. Scenario/ Extreme Event	M	L	M	M	H
Difficulty of Estimation:	L	L	H	L/M	M/H

Here I have provided my subjective estimates of the importance of each relevant variable introduced in this paper, and how much each variable varies by time, insurer, line of business and other factors. I have also indicated the difficulty of estimating the parameters (in modeling the capital requirements) for each variable. I have assumed that incorporating the risk aversion component is done simply, perhaps with a single parameter. Similarly, the correlation between insurer default and guaranty fund exhaustion is modeled with a single parameter. Notice that other variables, such as the loss distribution, are quite important but are currently considered when assessing risk-based capital.

Although it may appear that the conventional risk measures are better than the ARP because they are simpler (needing fewer variables to evaluate), this is not the case if one accepts the policyholder risk-preference basis of this paper: these variables were always important, but simply were not recognized by the conventional risk-based capital methodology.

7.2 Regulatory Role in Capital Standards

The preceding sections have addressed finding the optimal capital for an insurer. In an efficient market, insurers will gravitate toward these optimal levels without regulatory intervention. However, the market is far from efficient from the perspective of maximizing policyholder welfare, and the involvement of regulators is often necessary. An important role of the regulator is to mimic the *outcome* of an efficient market, or at least to mitigate the effects of the market imperfections.

Consequently, this means attempting to maximize policyholder welfare while maintaining a competitive market for insurance. The goal of this process is to approximate the optimal capital generated in an efficient market. Using risk-based capital standards, the regulator has the authority to force an insurer to maintain a minimum level of capital. If the insurer fails to achieve the desired capital, the regulator can impose various restrictions on the insurer's operations,[44] including shutting down the insurer. So, the regulator will want to set the intervention thresholds at levels that will tend to produce optimal capital levels.

This means that the more severe (i.e., shut-down) thresholds will be *lower* than the optimal capital amount for an insurer. For example, if the regulator sets the shut-down level at the optimum level, the insurer's management will need to carry *more* than that amount of capital. There are several reasons why the stringent thresholds should be lower than either the optimal capital level:

(1) An insurer operating above, but near the shut-down level would have a strong chance of being forced out of business if the business does not perform well over the following year. Therefore, its management will try to maintain a sufficient clearance above the threshold to minimize this possibility.

(2) With a high threshold, there is strong possibility of misidentifying companies that are actually strong as weak. Harrington (in H. Scott, ed., 2005) discusses this problem. Regulators will tend to value this type of error (Type 2) more than the converse (Type 1 error) where weak companies are incorrectly identified as being strong. This type of forbearance will lower the stringent threshold levels.

(3) An insurer cannot operate near a stringent threshold without the market knowing about it. Operating near the threshold will signal that the insurer is weak, resulting in loss of business. Consequently, the insurer will become weaker, and the result will tend to be a self-fulfilling prophecy.

Additionally, the regulator cannot be certain that an insurer with a low capital level is truly undercapitalized or the insurer's policyholders have low risk aversion. In the latter instance, the low capital amount could be appropriate for those PHs. This possibility requires a lower stringent

---

[44] Under the current U.S. risk-based capital framework designed by the National Association of Insurance Commissioners (NAIC), there are five control levels (thresholds), ranging from no regulatory action when capital exceeds 200% of the base RBC amount, to mandatory takeover of the insurer at 70% of the base RBC.

threshold than if all PHs had the same risk preferences.

Since the insurers will tend to carry more capital than the stringent intervention thresholds, these must be set low enough to induce insurers to generally carry the optimum level of capital. It will be difficult to quantify the relationship between the optimal capital and the regulatory thresholds that will produce the optimal capital. So, as is currently done, it may be necessary to use expert judgment to establish the threshold levels,[45] even if the optimal capital itself can be estimated reasonably well.

I observe here that it is quite possible that insurers will in practice carry more capital than the optimal amount required to benefit policyholders. This will imply a wide gap between the stringent-threshold regulatory amount and the amount typically held by insurers. Because of incentive conflicts, the interests of insurance management, shareholders, regulators and rating agencies may differ from those of policyholders. For example, shareholders may be interested in protecting the franchise value of the insurer and may shortchange the interests of current policyholders to obtain future profits. Insurance management may desire capital sufficient to protect their private interests (e.g., future employment prospects) and may care more about the *chance* of insolvency than the *expected amount* of the default. Regulators and rating agencies have a vested interest in limiting the frequency of insurer insolvencies, since the failures can be viewed as a breakdown of supervision or of the rating system.

The above incentive conflicts are exacerbated by the presence of guaranty funds, since the GFs allow for a somewhat painless insolvency experience from the policyholder perspective, but not painless to the other parties such as regulators or insurance management.

8. CONCLUSION

Based on maximizing policyholder welfare, it is possible to determine the optimal capital that an insurer should carry. To accomplish this, the appropriate solvency risk measure is ruin probability, using an adjusted probability distribution that reflects policyholders' risk aversion. The level of the adjusted ruin probability standard depends only on the insurer's frictional cost of holding capital. The assessment of underlying probability distributions of losses and assets, however difficult, is a standard actuarial problem. Determining the frictional cost of capital is a straightforward financial economics problem. On the other hand, estimates for the policyholder

---

[45] Setting the proper threshold level is conceptually another optimization problem: find the level that will create the best overall policyholder welfare, recognizing the above market-disrupting effects.

risk preferences are presently not available. This presents a ripe new area for empirical research.

The results of the analysis here establish that a number of variables, which are not considered in conventional risk measures, are important to properly establish an insurer's optimal capital. These features are absent when applying conventional solvency risk measures such as VaR or expected policyholder deficit. Incorporating these new factors is also a rich opportunity for further study.

Finally, although I have focused on property-casualty insurers in particular, the underlying principles will apply to other financial institutions as well. These entities have primary stakeholders such policyholders, depositors and investors. As with property-casualty insurance, the welfare of these parties is governed by the same general relationship between consumer value and the cost of carrying capital.

APPENDIX A: UTILITY THEORY AND CERTAINTY EQUIVALENT LOSSES

In this appendix I show the relationship between utility theory under risk and the certainty equivalent valuation of insurance losses. The utility of a wealth amount W is designated by $u(W)$ and the initial wealth of the PH by W_0. Accordingly, the utility of wealth given a loss y is $u(W_0 - y)$. However, since we are concerned with insurance losses here, it is convenient to redefine the utility to be a function of the loss amount:

$U(y) = u(W_0 - y)$.

Since the utility theory axioms have $u'(W) > 0$ and $u''(W) \leq 0$, with the derivatives taken with respect to wealth, when we convert to the utility of loss basis, we get

$U'(y) < 0$ and $U''(y) \leq 0$. Here, the derivative is taken with respect to the loss size y. These results are developed in Appendix A2. On the utility of loss basis, the relative risk aversion function $R_A(W) = -u''(W) / u'(W)$ becomes $R_A(y) = U''(y) / U'(y)$.

A1: Finding CE Values From a Utility Function

The *expected utility* (of wealth) is

$$EU = \int_0^\infty U(y)p(y)\,dy. \qquad (A1.1)$$

The certainty-equivalent wealth is the amount of wealth that gives as actual utility, the same amount as the expected utility. Thus, $CEW = U^{-1}(EU)$, where U^{-1} is the inverse of the utility function. The certainty-equivalent wealth, in turn, equals the actual wealth minus the certainty equivalent of the expected loss, or

$$CEW = U^{-1}(EU) = W_0 - \hat{L}. \qquad (A1.2)$$

Suppose that a policyholder faces a loss of size y with probability p, or no loss with probability $1 - p$. We want to determine the *certainty equivalent* amount corresponding to y, or $k(y)y$, where $k(y)$ is the CE function defined in section 3.2. The CE loss is

$$\hat{L} = k(0)(1 - p) \cdot 0 + k(y)p \cdot y = k(y)py. \qquad (A1.3)$$

From equations A1.2 and A1.3, we can determine the CE function value:

$$k(y) = \frac{W_0 - U^{-1}(EU)}{py} \quad .$$
(A1.4)

Thus, it is possible to determine the CE value of an individual loss amount directly from a utility function and the initial wealth.

To illustrate, consider the utility function $U(y) = \sqrt{W_0 - y}$. The initial wealth is 1600 and a loss of 1200 has a 10% probability. The expected wealth is 1600 – 0.1(1200) = 1480. The utility of the initial wealth is $40 = \sqrt{1600}$. The utility if the loss occurs is $20 = \sqrt{1600 - 1200}$, so the expected utility is 38 = 0.9(40) + 0.1(20).

The certain wealth corresponding to the expected utility of 38 is $1444 = U^{-1}(38) = 38^2$. Therefore, the certainty equivalent value of the expected loss is 156 = 1600 – 1444. From equation A1.4, the CE function value for the 1200 loss amount is $k(1200) = [1600 - 1444]/[0.1(1200)] = 1.3$. Notice that this value corresponds to a CE probability of 13% for the 1200 loss amount.

A2: The Shape of the CE Value Function

Because utility theory axioms impose constraints on the shape of the utility function, these restrictions will be reflected in the shape of the corresponding certainty-equivalent function. From Appendix A1, the CE function is related to the *inverse* of the utility function, so the properties of inverse functions will govern the translation from utility to certainty equivalence.

The first utility axiom is that utility increases with wealth: the derivative of utility with respect to wealth is $u'(W) > 0$. This means that utility declines as the loss size *y* becomes larger (i.e., $U'(y) < 0$) and thus the certainty-equivalent function value *increases* with *y*: $k'(y) > 0$.

The second utility property is that, because individuals are assumed to be risk-averse, the second derivative of the utility function with respect to wealth is negative:. $u''(W) \leq 0$. This means that utility declines as the loss size *y* increases, but at an *increasing* rate. This property translates to a CE function that increases at an increasing rate: $k''(y) \geq 0$.

Returning to the Appendix A1 example, we can vary the loss size from 0 to 1600, keeping the other parameters the same (e.g., *p* = 0.1). Thus the CE function value is

$$k(y) = 1.8 \frac{1600 - 40\sqrt{1600 - y}}{y} - 0.1. \qquad \text{(A2.1)}$$

Values for this function are shown graphically in Figure A2:

Figure A2
CE Function for Appendix A2 Numerical Example
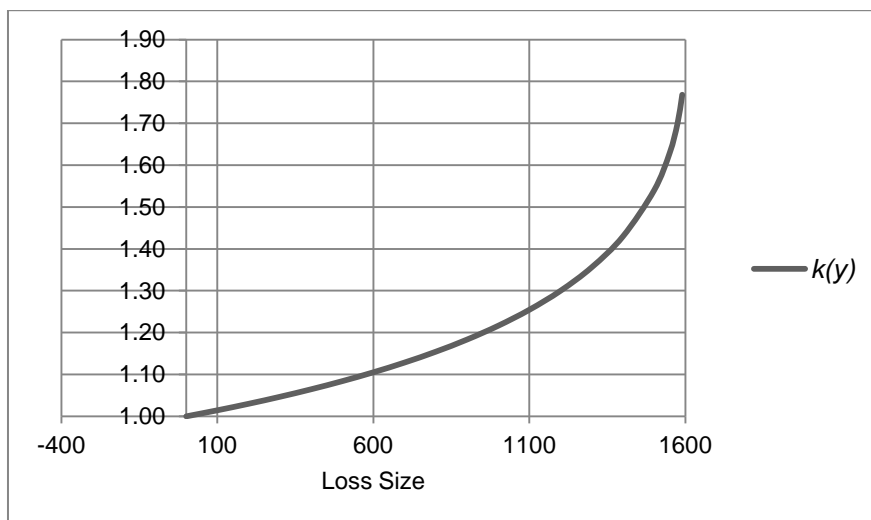
A3: The CE Loss Distribution

An important restriction on the CE function $k(y)$ is that $\int_0^\infty k(y)p(y)dy = 1$. This constraint means that the *average* value of the CE function equals 1, and thus will *be less than* 1 for losses that are small. This result seems anomalous, since the PH will be averse to the risk of small losses as well as for larger ones (albeit less risk-averse for the small ones), since the small losses are also random. However, it makes sense when we consider the entire loss distribution: since losses are mutually exclusive, two different loss values are negatively correlated. The negative co-variation will reduce the CE function value if the loss amounts are simultaneously considered, as compared to a situation where some loss sizes are considered selectively.

To illustrate the effect of negative covariance, consider a PH facing a loss of $y - \varepsilon$ with probability ½ or another loss of $y + \varepsilon$ also with probability ½. The amount ε is very small. One

of the two amounts will occur, but not both. If the PH insures against the first event and not the second, the CE expected loss will be greater than its expected value. This is because there is apparent risk: the loss will be zero or $y - \varepsilon$ with equal probability. The same is true if the second event is insured but not the first. However, if both events are insured (i.e., the entire loss distribution), then the CE expected loss will equal the expected value *y*, since essentially, the entire distribution is a single point *y* and there is no variance. The value *y* (plus or minus ε) is certain to occur. Thus the CE expectation over the entire distribution will be less than the sum of the CE values of the individual loss sizes taken in isolation.

The effect of the negative loss co-variation is that the *k*(*y*) values (such as in Appendix A2) will be reduced somewhat when other loss values are considered simultaneously. Extending the Appendix A2 example, suppose that (in addition to a loss amount of 1200 with a 10% probability) another loss of 1500 can occur, also with a 10% probability. Either may occur, but not both. Thus, there is an 80% chance that no loss will occur.

The utility if the 1200 loss happens is $20 = \sqrt{1600 - 1200}$, and the utility if the 1500 loss happens is $10 = \sqrt{1600 - 1500}$. So, the expected utility is 35 = 0.8(40) + 0.1(20) + 0.1(10). This gives a CE wealth of 1225 = 35^2 and the CE of the expected losses is 375 = 1600 – 1225. Thus the *joint* CE of the two loss amounts is 3750 = 375/0.1. However, taken separately, the CE of the 1200 loss is 1560 and the CE of the 1500 amount is 2310 (determined as in the Appendix A1 example). This gives a total CE for the separate losses of 3870, which is 130 more than their joint CE.

In this example, adding more possible loss values to fill out the entire probability distribution will reduce the CE values for all of the loss amounts even further.

To summarize, the particular value *k*(*y*) of the CE factor for a loss size *y* depends not only on *y* but also on *all other loss values* in the loss distribution, and their respective probabilities.

A4: Finding CE Default and Ruin Probability from a Utility Function

If we know the probability distribution of losses and the utility function, the certainty equivalent loss can be determined by inverting the expected utility, as shown in section A1. The expected utility for losses limited to an amount of assets *A* is

$$EUL(A) = \int_0^A U(y)p(y)\,dy + U(A)Q(A)\,, \qquad (A4.1)$$

where $U(y)$ is the utility if loss size y occurs and $Q(A)$ is the ruin probability, or chance that the loss exceeds assets. The expected utility for the entire loss distribution equals equation A4.1 with A set to infinity. The CE value of the limited loss is determined from the inverse utility function: $\hat{L}(A) = U^{-1}[EUL(A)]$. The CE of default is the difference between the CE of the entire loss and the loss limited to assets, just as the expected default is the difference between the expected loss and the limited expected value. Thus $\hat{D} = \hat{L} - \hat{L}(A)$. Since (from appendix C) the CE ruin probability $\hat{Q}(A)$ equals $-\partial \hat{D} / \partial A$, we have $\hat{Q}(A) = \partial[\hat{L}(A)] / \partial A$. Note that the derivative of $\hat{L}$ is zero, since it is not a function of A. Equation A4.1 thus provides a method for determining $\hat{D}$ and $\hat{Q}(A)$ given $U(y)$.

This method for getting the CE values for default and ruin probability can be illustrated using a general loss distribution with *exponential utility*. Here, I define the utility of wealth for loss size y as $U(y) = -e^{ay}$, with a being the risk aversion parameter. The expected utility of the limited loss is

$$EUL(A) = \int_0^A -e^{ay} p(y)\, dy - e^{aA} Q(A). \tag{A4.2}$$

To find $\hat{Q}(A)$, we first take the derivative of equation A4.1 with respect to A, getting

$$\frac{\partial EU[L(A)]}{\partial A} = [U(A)p(A) - U(0)p(0)] + U(A)[-p(A)] + Q(A)\frac{\partial U(A)}{\partial A}$$
$$= Q(A)\frac{\partial U(A)}{\partial A}. \tag{A4.3}$$

The terms involving $p(A)$ and $p(0)$ vanish since $p(0) = 0$ and the derivative of $Q(A)$ is $-p(A)$. The inverse of the exponential utility function for a utility value X is $U^{-1}(X) = \ln(-X) / a$. Next, we take the derivative of $\hat{L}(A)$. Note that since $U(A) = -e^{aA}$, its derivative equals $aU(A)$:

$$\hat{Q}(A) = \frac{\partial \hat{L}(A)}{\partial A} = \frac{\partial}{\partial A} \frac{\ln[-EUL(A)]}{a} = \frac{-1}{aEUL(A)} \frac{\partial[-EUL(A)]}{\partial A}$$
$$= \frac{Q(A)}{aEUL(A)} \frac{\partial U(A)}{\partial A} = \frac{Q(A)U(A)}{EUL(A)}. \tag{A4.4}$$

or $\hat{Q}(A) / Q(A) = U(A) / EUL(A)$. With a normal loss distribution, we can directly determine *EUL(A)*. The normal density is

$$p(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(y-L)^2}{2\sigma^2}\right]. \tag{A4.5}$$

Since the density in equation A4.1 is multiplied by $-\exp(ay)$, the product $U(y)p(y)$ for the normal $p(y)$ becomes another normal density $p_s(y)$ with a variate (shifted from the mean *L*) of $z_s = (y - L - a\sigma) / \sigma$, multiplied by the constant $-\exp[aL + \frac{1}{2}a^2\sigma^2]$. Thus, the expected utility equals this constant and its CE is the inverse, equal to $L + \frac{1}{2}a\sigma^2 = \hat{L}$. This derives the result in equation 3.43. For the normal distribution, equation A4.1 becomes

$$EUL(A) = -e^{a\hat{L}} \int_0^A p_s(y)\,dy - e^{aA}Q(A) = -e^{a\hat{L}}P_s(A) - e^{aA}Q(A). \tag{A4.6}$$

Here P_s is the cumulative normal probability with the shifted variate z_s. Converting EUL(A) to a certainty equivalent, we get

$$\hat{L}(A) = \frac{\ln[e^{a\hat{L}}P_s(A) + e^{aA}Q(A)]}{a}. \tag{A4.7}$$

Since $\hat{D} = \hat{L} - \hat{L}(A)$, we finally have

$$\hat{D} = \frac{-\ln[P_s(A) + e^{a(A-\hat{L})}Q(A)]}{a}. \tag{A4.8}$$

To determine the adjusted ruin probability $\hat{Q}(A)$ for the normal distribution, we use equations

A4.4 and A4.6, getting:

$$\hat{Q}(A) = \frac{Q(A)}{Q(A) + e^{-a(A-\hat{L})}P_s(A)} \ . \tag{A4.9}$$

Notice that the distribution for $P_s(A)$ has a mean of $L + a\sigma^2$, while for $Q(A)$ the mean is L. The mean underlying the $\hat{Q}(A)$ distribution is $\hat{L} = L + \frac{1}{2}a\sigma^2$. Letting $A = \hat{L} + x$, we see that $P_s(\hat{L} + x) = Q(\hat{L} - x)$. From equation A4.9 it is straightforward to show that $\hat{Q}(\hat{L} + x) + \hat{Q}(\hat{L} - x) = 1$ and therefore the CE distribution is symmetric around $\hat{L}$. However, as shown below, the CE distribution is not normal.

To show the non-normality, I use the section 3.5 example with $L = 1000$, $\sigma = 100$ and $a = 0.02$. With assets of 1300 (3 standard deviations above the mean) the unadjusted expected default D is 0.038 and the ruin probability is $Q(1300) = 0.135\%$. The CE expected loss is $1100 = 1000 + 0.5(0.02)(100)^2$. The shifted variate is $z_s = 1.00 = [1300 - 100 - (0.02)(100)]/100$, which is one standard deviation above the mean. Thus, the shifted cumulative probability is $P_s(1300) = 0.8413$. The factor $\exp[a(A - \hat{L})]$ equals $\exp[0.02(1300 - 1100)] = 54.598$, so we get $\hat{D} = 4.439 = \{-\ln[0.8413 + (54.598)(0.00135)]\}/0.02$. The CE ruin probability is $\hat{Q}(1300) = 8.05\% = 0.00135/[0.00135 + 0.8413/54.598]$. Notice that the denominator of equation A4.9 contains an exponential factor that is the reciprocal of the one in equation A4.8.

If the CE distribution were normal, then its ruin probability of 8.05% would imply a standard deviation of 142.71. Then, if we change the assets to 1200, we would get a CE ruin probability of 24.17%. However, following the above calculation, the true $\hat{Q}(1200)$ is 25.16%. Consequently, the CE distribution for the normal-exponential model is not normal.

It is interesting to compare the CE density $\hat{p}(y)$ with that of the underlying normal distribution $p(y)$. The approximate $\hat{p}(y)$ values can be calculated taking the difference of successive $\hat{Q}(A)$ values. Figure A4 below shows the two densities.

Figure A4
Probability Densities for Normal-Exponential Model
Illustrative Example
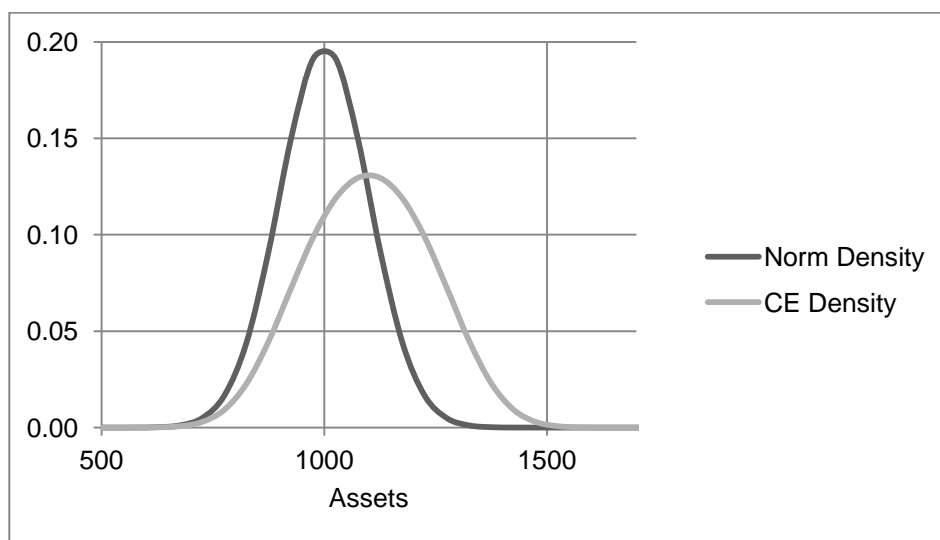
Notice that the CE density is symmetric, centered at the CE loss value of 1100. It also has a greater variance than its normal parent distribution. Also observe that, for a given asset amount (above the mean), the tail area of the adjusted distribution is much greater than that of the unadjusted distribution.

A5: CE Values for Binary Loss Model with Exponential Utility

Assume that an individual faces a loss of amount B with probability $p > 0$, and amount 0 with probability $q = 1 - p$. This is called a binary model, since there are two possible loss values: B or zero. The individual has risk preferences defined by exponential utility with risk aversion parameter a and has initial wealth of W_0 before considering the loss prospect. The utility of the initial wealth is $-e^{-aW_0}$ and the expected utility of wealth considering the loss is

$$EU(0) = -qe^{-a(W_0-0)} - pe^{-a(W_0-B)} = -e^{-aW_0}[q + pe^{aB}]. \qquad (A5.1)$$

Letting $\hat{W}_0$ denote the CE value of the wealth considering the loss, we have $\hat{W}_0 = W_0 - \hat{L}$, where $\hat{L}$ is the CE expected loss. Since $-e^{-a\hat{W}_0} = -e^{-aW_0}[q + pe^{aB}]$, we get

$$\hat{L} = \ln[q + pe^{aB}] / a. \tag{A5.2}$$

If the individual buys insurance for a premium π and the insurer has assets A, then the amount of loss absorbed by the individual is $B - A$ and the expected utility is $EU(A) = -qe^{-a(W_0-\pi)} - pe^{-a(W_0-B-A-\pi)}$. The certainty equivalent wealth after the insurance purchase is the initial wealth minus the premium minus the CE expected default, or $W_0 - \pi - \hat{D}$. Since $-e^{-a(W_0-\pi-\hat{D})} = EU(A)$, we solve for $\hat{D}$:

$$\hat{D} = \ln[q + pe^{a(B-A)}] / a. \tag{A5.3}$$

Notice that if $A = 0$, then $\hat{D} = \hat{L}$ and if $A = B$, then $\hat{D} = 0$. The CE ruin probability $\hat{Q}(A)$ equals the negative derivative of the CED, so we get

$$\hat{Q}(A) = \frac{pe^{a(B-A)}}{q + pe^{a(B-A)}}. \tag{A5.4}$$

If $A = 0$, we have $\hat{Q}(0) = pe^{aB} / (q + pe^{aB})$ and at $A = B$, $\hat{Q}(B) = p$.

For *two combined* independent binary risks, the development is similar (the values of variables for the combined risks are denoted with a subscript 2). Here we set the initial wealth and premium to zero, since they do not influence the CED and hence the CE ruin probability. Following section 3.6, assets are A per PH, for a total of $2A$. For $A < B/2$, the expected utility per PH is $EU(A) = -q - pqe^{a(B-2A)} - p^2 e^{a(B-A)}$. For $B/2 \le A < B$, $EU(A) = -q - pq - p^2 e^{a(B-A)}$. Then the respective CED values are

$$\hat{D}_2 = \ln(Z_1) / a \qquad \text{for } A < B/2 \tag{A5.5}$$
$$\hat{D}_2 = \ln(Z_2) / a \qquad \text{for } B/2 \le A < B,$$

where $Z_1 = q + pqe^{a(B-2A)} + p^2 e^{a(B-A)}$ and $Z_2 = q + pq + p^2 e^{a(B-A)}$. Here, if $A = 0$, we again get $\hat{D}_2 = \hat{L}$ and if $A = B$, then $\hat{D}_2 = 0$. Taking the derivative of the CED with respect to A, we get the CE ruin probabilities:

$$\hat{Q}_2(A) = [2pqe^{a(B-2A)} + p^2 e^{a(B-A)}] / Z_1 \quad \text{for } A < B/2 \qquad \text{(A5.6)}$$

$$\hat{Q}_2(A) = p^2 e^{a(B-A)} / Z_2 \qquad \text{for } B/2 \le A < B.$$

For $A = 0$, we get $\hat{Q}_2(0) = (1 - q^2)e^{aB} / (q + pe^{aB})$ and for $A = B$, $\hat{Q}_2(B) = p^2$. Notice that $\hat{Q}_2(0) / \hat{Q}(0) = 2 - p > 1$ and that $\hat{Q}_2(B) / \hat{Q}(B) = p < 1$. Thus, the CE ruin probability for the combined risks has a higher maximum value (at $A = 0$) and a lower minimum value (at $A = B$) than for a single risk. Thus, based on equation 5.51, there is a region where a subadditivity (SA) violation may exist and another region where a SA violation cannot happen.

For a fair premium, equation 5.14 defines the risk measure as $\theta(A) = [\hat{Q}(A) - Q(A)] / [1 - Q(A)]$. For a single binary risk, the unadjusted ruin probability is $Q(A) = p$, for $A \le B$. Consequently,

$$\theta(A) = [\hat{Q}(A) - p] / q, \qquad \text{(A5.7)}$$

so the fair premium risk measure $\theta(A)$ is a linear function of the basic premium risk measure $\hat{Q}(A)$.

For a combination of *two independent binary risks*, and for $A < B/2$, the expected default for a single PH is $D_2 = pq(B - 2A) + p^2(B - A)$. By taking the negative of the derivative of D_2 with respect to A we get $Q_2(A) = 1 - q^2$. For $B/2 \le A < B$, we have $Q_2(A) = p^2$. Therefore, equation 5.14 gives

$$\theta_2(A) = [\hat{Q}_2(A) + q^2 - 1] / q^2 \qquad \text{for } A < B/2 \qquad \text{(A5.8)}$$

$$\theta_2(A) = [\hat{Q}_2(A) - p^2] / [1 - p^2] \qquad \text{for } B/2 \le A < B.$$

For $A = 0$, with some manipulation, we get $\theta_2(0) / \theta(0) = (2 - p) / q > 1$ and for $A = B$ (as a limit), $\theta_2(B) / \theta(B) = p < 1$. Just as in the basic premium case, there is a region where a subadditivity violation may exist and another region where it cannot occur.

APPENDIX B: STOCHASTIC MEAN LOSS MODEL

The classic aggregate loss model from risk theory (see Lundberg [1903]) is the compound

Poisson process, where the number of losses is Poisson and each individual loss has the same distribution with prescribed parameters and thus a stable mean. The individual losses are independent. However, in practice, the losses are not independent (they are subject to common factors such as inflation, regulation and the court system). Further, the mean of losses for a given line of business (or other subdivision of an insurer's risk portfolio) is not known; it must be determined empirically.

Relaxing the independence assumption, we introduce another random variable that governs the mean from which all the individual losses are drawn (in this section, random variables are indicated with a tilde).

Let $\tilde{I}$ represent the stochastic mean variable, which itself has a mean of 1 and a variance γ^2. We assume that there are N policyholders. Each PH i has losses (we allow for multiple claims in the one-period model) denoted by $\tilde{x}_i$. The $\tilde{x}_i$ are measured before applying the mean variable $\tilde{I}$.

Let $\tilde{S}$ denote the sum of aggregate losses before applying the stochastic mean. The unconditional aggregate losses are

$$\tilde{X} = \tilde{I} \cdot \tilde{S} = \tilde{I}[\tilde{x}_1 + \tilde{x}_2 + \ldots + \tilde{x}_N]. \tag{B.1}$$

Here the $\{\tilde{x}_i\}$ are the individual losses and N is the number of losses. $\tilde{I}$ and $\tilde{S}$ are independent, so the covariance between losses is due to the stochastic mean. Let M be the mean of the individual loss $\tilde{x}_i$ and σ^2 its variance. Let ρ be the correlation between the individual losses. The variance of $\tilde{S}$ is

$$\mathrm{Var}(\tilde{S}) = \sigma^2[N + \rho(N^2 - N)] \tag{B.2}$$

and its mean is *NM*.

The variance of the product of two independent random variables $\tilde{I}$ and $\tilde{S}$ is

$$\mathrm{Var}(\tilde{I} \cdot \tilde{S}) = \mathrm{Var}(\tilde{I}) \cdot \mathrm{Var}(\tilde{S}) + \mathrm{E}(\tilde{I})^2 \cdot \mathrm{Var}(\tilde{S}) + \mathrm{E}(\tilde{S})^2 \cdot \mathrm{Var}(\tilde{I}). \tag{B.3}$$

Here, E( ) denotes the expectation. Thus the variance of the aggregate losses $\tilde{X}$ is

$$\mathrm{Var}(\tilde{X}) = \left[N(1-\rho) + \rho N^2\right]\sigma^2\left[1 + \gamma^2\right] + \gamma^2 M^2 N^2. \tag{B.4}$$

Let $\tilde{Y} = \tilde{X} / N$ denote the share of aggregate losses for an individual PH. Then $\mathrm{Var}(\tilde{Y}) = \mathrm{Var}(\tilde{X}) / N^2$. Thus, we have

$$\mathrm{Var}(\tilde{Y}) = \gamma^2 M^2 + \rho\sigma^2(1+\gamma^2) + (1-\rho)\sigma^2(1+\gamma^2) / N. \tag{B.5}$$

As *N* becomes large, the variance of the individual PH losses tends toward

$$\mathrm{Var}(\tilde{Y}) = \gamma^2 M^2 + \rho\sigma^2(1+\gamma^2). \tag{B.6}$$

So for large *N*, the variance of the individual PH losses tends to a constant limiting value. Also, if $\gamma = 0$ and $\rho > 0$, the limit is $\rho\sigma^2$. If $\gamma > 0$ and $\rho = 0$, then the limit is $\gamma^2 M^2$. Consequently, if either the losses are subject to a stochastic mean or they are correlated, then with a large number of policyholders, the variance of individual PH losses will reach a limit.

An example will illustrate the convergence. Assume that $M = 1{,}000$, $\sigma = 300$, $\gamma = 0.1$ and $\rho = 0.2$. For $N = 1{,}000$, equation B.5 gives an individual PH variance of 28,253, compared to the asymptotic value of 28,180, a difference of only 0.26%. Increasing *N* to 10,000 policies moves the variance to 28,187, cutting the difference to 0.03% from the asymptotic value. Notice that the asymptotic standard deviation is 167.87, which compares to 300.00 from the individual loss distribution.

APPENDIX C: DERIVATIVE OF THE EXPECTED DEFAULT

To determine the derivative of the expected default, we use the general method for the derivative of an integral, with the upper limit a constant *b* and the lower limit a function of the variable whose derivative is taken:

$$\frac{\partial}{\partial y}\int_{g(y)}^{b} F(x,y)dx = \int_{g(y)}^{b} \frac{\partial}{\partial y}F(x,y)dx - F[g(y),y]\frac{d[g(y)]}{dy}. \tag{C.1}$$

Thus,

$$\frac{\partial D}{\partial A} = \int_{A}^{\infty} \frac{\partial}{\partial A}[x - A]p(x)\,dx - (A - A)p(A)\frac{\partial A}{\partial A}. \tag{C.2}$$

The right-hand term in equation C.2 equals zero and the derivative becomes

$$\frac{\partial D}{\partial A} = \int_{A}^{\infty} -p(x)\,dx = -Q(A), \tag{C.3}$$

where $Q(A) = \int_{A}^{\infty} p(x)\,dx$ is the tail, or ruin, probability.

For an adjusted probability density $\hat{p}(x)$, we have in parallel fashion,

$$\frac{\partial \hat{D}}{\partial A} = \int_{A}^{\infty} -\hat{p}(x)\,dx = -\hat{Q}(A). \tag{C.4}$$

APPENDIX D: OPTIMAL CAPITAL WITH INCOME TAXES

Expanding the basic model to include income taxes, we also need to introduce an investment component. Assume that all cash is invested in riskless investments at a one-period rate r, and that all income is taxed at the end of the period at a rate t. Premium is collected at the beginning of the period and losses are paid at the end of the period. We further assume that the losses contain no market risk, so that the expected return to investors[46] in the insurer is also r. Alternatively, we can assume that the value of the loss is adjusted to include the market risk.

In order to attract capital from the insurer's investors, the expected rate of return *after taxes* must also equal r.

If the premium equals the present value of the expected loss, then initial assets are $A_0 = C + L / (1 + r)$. The expected value of the ending assets, prior to income tax, is $A_0(1 + r) - L = C + rC$. The investment income rC is taxed, leaving $C + rC - trC$. However, for a fair return to investors, the ending assets must be $C + rC$. The amount trC must be made up

---

[46] In this formulation, the expected default is not subtracted from premium, so the result approximates a true equilibrium optimum, which is a more complex version of equation 5.14.

by charging an extra premium amount *zC* at some rate *z* proportional to capital, so the premium is $\pi = L / (1 + r) + zC$. The extra premium is itself taxed as underwriting profit, so the amount *zC* will grow to $zC(1 + r) - tzC - trzC$ after taxes. Notice that the investment income *rzC* on *zC* is also taxed. Equating the ending after-tax value of the additional premium with the double-taxation burden *trC*, we solve for *z*:

$$z = \frac{tr}{(1 + r)(1 - t)} \ . \tag{D.1}$$

Let *A* represent the amount of assets prior to payment of the loss and income taxes. If the loss is larger than *A*, the insurer will default and no tax is paid. I assume here that a negative income tax liability arising from a large loss does *not* increase the assets available to pay the loss. Thus we have

$$A = [C + \frac{L}{1 + r} + zC](1 + r) = L + C(1 + z)(1 + r). \tag{D.2}$$

In parallel fashion to equation 5.11, the consumer value *V* of the insurance equals the present value of the CE of the covered losses, minus the premium:

$$V = \frac{\hat{L}}{1 + r} - \frac{\hat{D}}{1 + r} - \frac{L}{1 + r} - zC. \tag{D.3}$$

Taking derivatives and equating to zero, we have

$$z\frac{\partial C}{\partial A} = \frac{1}{1 + r}\frac{-\partial \hat{D}}{\partial A} = \frac{1}{1 + r}\hat{Q}(A). \tag{D.4}$$

From equation D.2, we get $\partial C / \partial A = 1 / [(1 + z)(1 + r)]$. From equation D.4 we get

$$\hat{Q}(A) = \frac{z}{1 + z} \ . \tag{D.5}$$

Using the value of z in equation D.1, we get the optimal CE ruin probability in terms of the interest rate and the income tax rate:

$$\hat{Q}(A) = \frac{rt}{1 + r - t} \ .$$

(D.6)

ACKNOWLEDGMENTS

The ideas behind this paper arose as a result of my work on the American Academy of Actuaries Property-Casualty Risk-Based Capital Committee, headed by Alex Krutov. After the 2008 financial crisis, I began to model extreme events for the property-casualty industry and developed the policyholder welfare approach to optimal capital. I am grateful to Alex for his encouragement as I attempted to advance these concepts.

In 2011, the American Academy of Actuaries committee sought help from the Casualty Actuarial Society (CAS) in preparing risk-based capital proposals for the National Association of Insurance Commissioners. I joined the CAS RBC Dependency and Correlation Working Party, led by Allan Kaufman. As my contribution to this effort, I began a project to determine the best solvency risk measure for property-casualty insurers. The assignment greatly expanded my earlier work, and this paper is the result. In fact, it serves as the report for my subcommittee on the working party. I am deeply thankful for Allan's stewardship in guiding me along and keeping a clear focus throughout the project. His innumerable astute comments, sharp critique and editorial suggestions were invaluable; they forced me to explain results more clearly — the paper is much better for his involvement.

I also thank Glenn Myers, a fellow member of the working party, for his helpful comments and insight. In particular, he suggested a comparison of utility theory and the certainty equivalent formulation. Additionally, I thank Alex Krutov and Tom Struppeck for spurring the discussion of subadditivity, a topic I had not previously considered.

REFERENCES

[1]   Acerbi, C. (2002) "Spectral measures of risk: A coherent representation of subjective risk aversion," Journal of Banking and Finance, 26 (7), 1487–1503.

[2]   Albanese, C. (1997), Credit Exposure, Diversification Risk and Coherent VaR, Working Paper, Department of Mathematics, University of Toronto, September.

[3]   Arrow, K. J. and G. Debreu. 1954. Existence of an equilibrium for a competitive economy. Econometrica 22:265–290.

[4]   Artzner, P. (1999). "Application of coherent capital requirements", North American Actuarial Journal, vol. 3(2), 11–25.

[5]   Birge, John R. and François V. Louveaux (1997). Introduction to Stochastic Programming. Springer Verlag, New York.

[6]   Butsic, Robert P., Solvency Measurement for Property-Liability Risk-Based Capital Applications, Journal of Risk and Insurance, 1994 Vol. 61, (4), 656-690.

[7]   Butsic, Robert P., CAS Forum 1999, Capital Allocation for Property-Liability Insurers: A Catastrophe Reinsurance Application, 1-70.

[8]   Cummins, J. David, 1988, Risk-Based Premiums for Insurance Guaranty Funds, Journal of Finance, 43: 823-839.

[9]   Einav, Liran, Amy Finkelstein, and Mark R. Culleny. 2010. Estimating welfare in insurance markets using variation in prices. Q. J. Econ. 125**.**

[10] Einhorn, David (June/July 2008), *Private Profits and Socialized Risk*, GARP Risk Review.

[11] Hancock, John, Paul Huber, and Pablo Koch (2001), "The Economics of Insurance: How Insurers Create Value for Shareholders," Zurich: Swiss Re Technical Publishing.

[12] Harrington, S., "Insurance Derivatives, Tax Policy, and the Future of the Insurance Industry," *Journal of Risk and Insurance,* 1997, 64:719-725.

[13] Harrington, Scott E., Edited version published in H. Scott, ed., Capital Adequacy Beyond Basel: Banking, Securities, and Insurance, Oxford University Press, 2005 Capital Adequacy in Insurance and Reinsurance.

[14] Holton, Glyn (2003). *Value-at-Risk: Theory and Practice*. Academic Press.

[15] Hull, John C. (2008). Options, Futures and Other Derivatives (7 ed.). Prentice Hall. ISBN 0-13-505283-1

[16] Jorion, Philippe (2006). Value at Risk*: The New Benchmark for Managing Financial Risk* (3rd ed.). McGraw-Hill. ISBN 978-0-07-146495-6.

[17] Lowe, Stephen, François Morin, and Dean Swallow, *Risk Horizon and the Measurement of Economic Capital for General Insurers*, Towers Watson, 2011.

[18] Lundberg F. (1903) Approximerad Framställning av Sannolikehetsfunktionen, Återförsäkering av Kollektivrisker, Almqvist & Wiksell, Uppsala.

[19] Marshall, Alfred (1947 [1890]): *Principles of Economics*. London, Macmillan.

[20] Meyers, G.G., Schenker, N., "Parameter Uncertainty in the Collective Risk Model", Proceedings of the Casualty Actuarial Society, LXX, 1983, 111-143.

[21] Myers, S. C., and J. A. Read, Jr. 2001. Capital Allocation for Insurance Companies. *Journal of Risk and Insurance* 68(4): 545–80.

[22] Nocero, Joe (January 4, 2009), *Risk Mismanagement*, The New York Times Magazine.

[23] Panjer, Harry H., ed., (1998) Financial Economics With Applications to Investments, Insurance, and Pensions, Society of Actuaries.

[24] Pratt, J. W., "Risk aversion in the small and in the large," Econometrica 32, January–April 1964, 122–136.

[25] Rothschild, M. and J. Stiglitz. Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. The Quarterly Journal of Economics, pages 629–649, 1976.

[26] Varian, Hal R., Microeconomic Analysis, Third edition. Norton, New York 1992.

[27] Wang, S. S. 1996. Premium Calculation by Transforming the Layer Premium Density. *ASTIN Bulletin* 26(1): 71–92.

[28] Yaari, M.E. (1987). "The dual theory of choice under risk", Econometrica 55, 95–115.

GLOSSARY OF ABBREVIATIONS AND NOTATIONS

Abbreviation	Meaning	Section Where Defined
ARP	Adjusted ruin probability	5.1
CE	Certainty equivalent	3.2
CEL	Certainty equivalent expected loss	3.2
CED	Certainty equivalent expected default	3.5
CV	Consumer value	3.6
DCP	Default Correlation Parameter	6.2
EPD	Expected policyholder deficit	1
FCC	Frictional capital cost	4.2
GF	Guaranty fund	6.2
NAIC	National Association of Insurance Commissioners	7.2
PV	Present value	4.3
PH	Policyholder	3.1
RBC	Risk-based capital	2
RP	Ruin probability	5.4
SA	Subadditivity	5.5
SD	Standard deviation	5.3
TVaR	Tail value-at-risk	1
VaR	Value-at-risk	1

Notation	Meaning	Section Where Defined
a	Exponential utility risk aversion parameter	3.4
A	Assets	3.5
A_0	Initial assets	App. D
b	Upper integration limit	App. C
B	Binary loss size	5.5
C	Capital	4.1
CEW	Certainty-equivalent wealth	App. A1
D	Expected default	3.5
$\hat{D}$	Certainty-equivalent expected default	3.5
e_0, e_1	Expense coefficients	5.1
$E(\ )$	Expectation operator	App. B
EU	Expected utility	App. A1
$EUL(\ )$	Limited expected utility	App. A4
$f(\ )$	General function	3.6
$F(\ )$	General function	App. C
$g(\ )$	General function	App. C
$\tilde{I}$	Stochastic mean variable	App. B
$k(\ )$	Certainty-equivalent function	3.2
k	Average value of the CE function	3.2
K	Variable in normal-exponential model	App. A4
L	Expected loss	3.2
$\hat{L}$	Certainty-equivalent loss	3.2
$\hat{L}(\ )$	Limited CE expected loss	3.5
M	Mean of individual loss	App. B
$MT(\ )$	Tail moment	5.4
n	Degree of the tail moment	5.4
N	Number of policies	App. B
p	Binary loss probability	App. A1
$p(\ )$	Probability density	3.2
$\hat{p}(\ )$	Probability density, adjusted for risk aversion	3.2
$p_A(\ )$	Risk-neutral probability density for asset size	6.1
$p_s(\ )$	Probability density with shifted mean	App. A4
$P_s(\ )$	Cumulative probability with shifted mean	App. A4
q	Probability of zero loss	App. A5
$Q(\ )$	Ruin probability	5.1
$\hat{Q}(\ )$	Ruin probability, adjusted for risk aversion	5.1
r	Investment return	4.2
$R_A(\ )$	Absolute risk aversion function	3.4
$RM_1(\ ), RM_2(\ )$	Risk Measure	5.5
$\tilde{S}$	Sum of aggregate losses, without stochastic mean	App. B

Notation	Meaning	Section Where Defined
t	Income tax rate	4.2
$u(\ )$	Utility function, based on wealth	App. A1
$U(\ )$	Utility function, based on loss size	3.4
v	Valuation level of risk measure	5.5
V	Consumer value of insurance contract	5.1
Var$(\ )$	Variance operator	App. B
W	Wealth	App. A1
W_0	Initial wealth	3.4
$\hat{W}_0$	CE wealth given a potential loss	App. A5
x	Loss size; also a general variable	3.4
$\tilde{x}_i$	Random individual loss size	App. B
$\tilde{X}$	Unconditional aggregate losses	App. B
y	Individual policy loss size	3.2
$\hat{y}$	Certainty equivalent loss size	3.2
$\tilde{Y}$	Policyholder share of aggregate losses	App. B
z	Frictional cost of capital rate	4.2
z_s	Shifted normal variate	App. A4
Z_1, Z_2	Intermediate variables	App. A5
α	VaR confidence level	5.4
γ^2	Variance of stochastic mean	App. B
Δ	Change in loss size	5.4
ε	Small change in loss size	App. A3
π	Premium	4.1
$\theta(\ )$	Fair premium risk measure	5.1
ρ	Correlation between losses	App. B
σ^2	Variance of loss	3.4

BIOGRAPHY OF THE AUTHOR

Robert P. Butsic is a retired actuary currently residing in San Francisco. He is a member of the American Academy of Actuaries Property-Casualty Risk Based Capital Committee and the Casualty Actuarial Society's Risk-Based Capital Dependency and Calibration Working Group. He previously worked for Fireman's Fund Insurance and CNA Insurance. He is an Associate in the Society of Actuaries, has a B.A. in mathematics and an MBA in finance, both from the University of Chicago. He has won the Casualty Actuarial Society's Michelbacher Award (for best Discussion Paper) five times. In the early 1990s he was a member of the American Academy of Actuaries working group that advised the NAIC in developing the current property-casualty RBC methodology. Since the 2008 financial crisis he has enjoyed reading economics blogs, which have stimulated the development of this paper.

Kurtosis and skewness estimation for non-life reserve risk distribution

Eric Dal Moro, Fellow of the French Actuarial Association

Abstract: In the daily tasks of a non-life actuary, the reserve risk distribution plays a central role. For example, the estimation of the cost of capital used in commutation pricing relies heavily on the assumption retained for the shape of the non-life reserve risk distribution.

Even though some distributions are widely used in the actuarial community (e.g. Lognormal distribution), it is interesting to note that very little is known on the determinants of the shape of the non-life reserve risk distribution. In general, the mean is usually defined as the Best Estimate and the standard deviation can be estimated using different methods (e.g. Mack 1993a). In terms of higher moments, Generalized Linear Models (GLM) and bootstrap techniques offer different possibilities of quantifying moments and quantiles (see Wüthrich-Merz 2008 and England and Verrall 2002). However, these models require the specification of some explicit parametric distribution (e.g. for the residuals) in order to be applied.

Following a first introduction of skewness estimation of non-life reserve risk distribution (see Dal Moro 2012), this article investigates the possibility to estimate the kurtosis of the non-life reserve risk distribution. In addition, the robustness of the skewness and kurtosis estimation based on the proposed formulas is tested on eight different triangles.

Keywords. Skewness; Kurtosis; Platykurtic; Variance; Chain-Ladder; Reserve risk distribution; Correlation; Gaussian copula; Generalized Pareto Distribution; Johnson distribution.

1. INTRODUCTION

1.1 Background

Today, whenever a non-life actuary has the necessity to use and determine the full reserve risk distribution, he has several options:

- The most usual option consists in assuming that the full reserve risk distribution follows a lognormal distribution which parameters are fitted to the reserve Best Estimate and to the reserve standard deviation. The reserve standard deviation can be determined using the Mack method in the Chain-Ladder framework (see Mack 1993a) or in the Bornhuetter-Ferguson framework (see Mack 2008) or using the Hybrid Chain-Ladder method (see Arbenz 2010).

- Another option consists of using bootstrapping techniques to determine the full empirical distribution (see England and Verall 2006).

- A final option consists of using GLM techniques (see Merz and Wüthrich 2008).

The GLM and Lognormal assumptions essentially rely on the first and second moment of the

distribution and do not attempt to reflect any further knowledge of higher moments. In the same way, the bootstrapping techniques are based on resampling techniques which do not reflect any knowledge of higher moments.

This article tries, in a first step, to determine some knowledge on higher moments of the reserve risk distribution, in particular skewness and kurtosis. With this knowledge, it tries to draw general conclusions on the appropriateness of the Lognormal assumption on the basis of test cases.

In all the following, we will limit ourselves to the Chain-Ladder framework.

1.2 Outline

This paper is divided into the following sections:

- Section 2 provides basic definitions of chain-ladder coefficients

- Section 3 establishes the estimator of the kurtosis of the incurred claim amount for each development year of each accident year.

- Section 4 provides a description of the simulations done to estimate the skewness and the kurtosis for non-life reserves for each accident year of a triangle.

- Section 5 provides numerical examples.

2. DEFINITIONS

2.1 Definitions of Chain-Ladder elements

The following brief review of the Chain-Ladder method is based on Mack (1993a).

Let $C_{i,k}$ denote the cumulative incurred claims amount of accident year i after k years of development, $1 \leq i, k \leq I$, where I denotes the most recent accident year. Then $C_{I-i+1,i}$ denotes the currently known claims amount of accident year I-i+1. The basic chain-ladder assumption is that there exists development factors $f_1, \ldots, f_{I-1}$ such that:

$$E(C_{i,k+1} \mid C_{i,1}, \ldots, C_{i,k}) = f_k C_{i,k}, 1 \leq i \leq I, \ 1 \leq k \leq I-1 \tag{1}$$

where the link ratios (age-to-age factors) can be estimated as follows:

$$\hat{f}_k = \frac{\sum_{j=1}^{I-k} C_{j,k+1}}{\sum_{j=1}^{I-k} C_{j,k}}, 1 \leq k \leq \text{I-1} \tag{2}$$

under the assumption that $\{C_{i,1}, \ldots, C_{i,I}\}, \{C_{j,1}, \ldots, C_{j,I}\}, i \neq j$ are independent.

In this paper, $\hat{f}$ will denote the estimator of the random variable f.

Note: The $\hat{f}_k$ are unbiased and uncorrelated (see Mack 1993a).

2.2 Variance of $C_{i,k}$

In the framework of distribution-free calculation of the standard error of the reserve estimates, several variance models exist. For the purpose of this article, we will focus on the Mack standard error.

As for the variance of $C_{i,k+1}$, T. Mack (1993a) induced that $Var(C_{i,k+1} \mid C_{i,1}, \ldots, C_{i,k})$ (where $Var(A \mid B)$ denotes the conditional variance of A knowing B) should be proportional to $C_{i,k}$, i.e.:

$$Var(C_{i,k+1} \mid C_{i,1}, \ldots, C_{i,k}) = C_{i,k}\sigma_k^2, 1 \leq i \leq I, \ 1 \leq k \leq \text{I-1} \tag{3}$$

where $\hat{\sigma}_k^2 = \dfrac{1}{I-k-1}\sum_{i=1}^{I-k} C_{i,k}\left(\dfrac{C_{i,k+1}}{C_{i,k}} - \hat{f}_k\right)^2 \quad for \ 1 \leq k \leq \text{I-2}$ (4)

It can be shown that the estimator $\hat{\sigma}_k^2$ is unbiased (see Mack 1993a).

2.3 Skewness of $C_{i,k}$

Let's denote $SK(A \mid B)$, the conditional skewness of A knowing B:

$$SK(A/B) = E\big((A - E(A/B))^3 / B\big)$$

Following the result in Dal Moro 2012, we assume that $SK(C_{i,k+1} \mid C_{i,1}, ..., C_{i,k})$ is proportional to $C_{i,k}^{3/2}$, i.e. :

$$SK(C_{i,k+1} \mid C_{i,1}, ..., C_{i,k}) = C_{i,k}^{3/2} Sk_k^3 , 1 \le i \le I, \ 1 \le k \le I\text{-}2 \tag{5}$$

with:

$$\hat{Sk}_k^3 = \frac{1}{\left(I - k - \dfrac{\left(\sum\limits_{i=1}^{I-k} C_{i,k}^{3/2}\right)^2}{\left(\sum\limits_{i=1}^{I-k} C_{i,k}\right)^3} \right)} \sum_{i=1}^{I-k} C_{i,k}^{3/2} \left(\frac{C_{i,k+1}}{C_{i,k}} - \hat{f}_k \right)^3 \quad \text{for } 1 \le k \le I\text{-}3 \tag{6}$$

For $k = I\text{-}2$ and $k = I\text{-}1$, as data is not available to estimate $\hat{Sk}_k^3$, it will be assumed that $\hat{Sk}_k^3 = 0$ for this article.

3. KURTOSIS ESTIMATOR

In probability theory and statistics, kurtosis is any measure of the "peakedness" of the probability distribution of a real-valued random variable. One common measure of kurtosis, originating with Karl Pearson, is based on a scaled version of the fourth moment of the data or population. Mathematically, it is described by the formula below for a random variable A:

$$\beta = \frac{E\left((A - E(A))^4\right)}{E\left((A - E(A))^2\right)^2}$$

For this measure, higher kurtosis means more of the variance is the result of infrequent extreme deviations, as opposed to frequent modestly sized deviations. If a distribution's kurtosis is greater than 3, it is said to be leptokurtic. If its kurtosis is less than 3, it is said to be platykurtic. Leptokurtosis is associated with distributions that are simultaneously "peaked" and have "fat tails." Platykurtosis is associated with distributions that are simultaneously less peaked and have thinner tails.

With this definition of kurtosis, the aim of this section is to find an expression for the kurtosis of $C_{i,k}$ within the Chain-Ladder framework. Let's denote $KT(A \mid B)$, the conditional fourth centered moment of A knowing B:

$$KT(A \mid B) = E\left((A - E(A \mid B))^4 \mid B\right)$$

For the purpose of this article, it will be assumed that:

$$\frac{KT\left(C_{i,k+1} \mid C_{i,1},...,C_{i,k}\right)}{\left[Var\left(C_{i,k+1} \mid C_{i,1},...,C_{i,k}\right)\right]^2} \text{ depends on k but not on i.}$$

This assumption implies that, for one development year within a triangle, the kurtosis is the same for any accident year. This assumption can be applied in the context of a reserving portfolio for which risks would have the same characteristics for every accident year.

With such an assumption, we have:

$$\exists \gamma_k \text{ such that } \gamma_k = \frac{KT\left(C_{i,k+1} \mid C_{i,1},...,C_{i,k}\right)}{\left[Var\left(C_{i,k+1} \mid C_{i,1},...,C_{i,k}\right)\right]^2} = \frac{KT\left(C_{i,k+1} \mid C_{i,1},...,C_{i,k}\right)}{\left[\sigma_k^2 C_{i,k}\right]^2}$$

$$\Rightarrow KT\left(C_{i,k+1} \mid C_{i,1},...,C_{i,k}\right) = \gamma_k \left[\sigma_k^2 C_{i,k}\right]^2$$

Hence, we have the result that $KT\left(C_{i,k+1} \mid C_{i,1},...,C_{i,k}\right)$ is proportional to $C_{i,k}^2$, i.e. :

$$KT\left(C_{i,k+1} \mid C_{i,1},...,C_{i,k}\right) = C_{i,k}^2 Kt_k^4 , 1 \leq i \leq I, \ 1 \leq k \leq I\text{-}3 \tag{7}$$

Based on the definition above, the theorem below will give an unbiased estimator for $\hat{K}t_k^4$.

Theorem:

Similar to the variance expression in equation (4), we show that the estimator below is unbiased.

$$\hat{Kt}_k^4 = \frac{\left[\sum\limits_{i=1}^{I-k} C_{i,k}^2 \left(\frac{C_{i,k+1}}{C_{i,k}} - \hat{f}_k\right)^4 - 3\left(\hat{\sigma}_k^2\right)^2 \left(2 - 6\frac{\sum\limits_{i=1}^{I-k} C_{i,k}^2}{\left(\sum\limits_{i=1}^{I-k} C_{i,k}\right)^2} + 4\frac{\sum\limits_{i=1}^{I-k} C_{i,k}^3}{\left(\sum\limits_{i=1}^{I-k} C_{i,k}\right)^3}\right)\right]}{\left(\sum\limits_{i=1}^{I-k}\left(1 - \frac{C_{i,k}}{\sum\limits_{i=1}^{I-k} C_{i,k}}\right)^4 + \frac{\left(\sum\limits_{i=1}^{I-k} C_{i,k}^2\right)^2 - \sum\limits_{i=1}^{I-k} C_{i,k}^4}{\left(\sum\limits_{i=1}^{I-k} C_{i,k}\right)^4}\right)} \quad \text{for } 1 \le k \le I\text{-4} \quad (8)$$

For $k = I\text{-}3, k = I\text{-}2$ and $k = I\text{-}1$, as data is not available to estimate $\hat{Kt}_k^4$, it will be assumed that $\frac{\hat{Kt}_k^4}{\left(\hat{\sigma}_k^2\right)^2} = 3$ for this article. This last assumption implies that the characteristic of a normal distribution is retained on the last development years.

The proof of this theorem is given in Appendix A.

Remark:

The formula for $\hat{Kt}_k^4$ has the standard shape of a kurtosis estimator: One element which corresponds to the weighted average of $\left(\frac{C_{i,k+1}}{C_{i,k}} - \hat{f}_k\right)^4$ and another element which has a component $3\left(\hat{\sigma}_k^2\right)^2$ (see formula in Cramer 1946).

4. SIMULATIONS

After finding the expressions for the kurtosis and the skewness of $C_{i,k}$, the next step would consist in finding closed formulas for the skewness and kurtosis of non-life reserves by accident year. This should be done by using the law of total cumulants. For example, in the case of skewness, the law of total skewness gives:

$$SK(A) = E[SK(A \mid B)] + SK[E(A \mid B)] + 3Cov[E(A \mid B), Var(A \mid B)]$$

However, when using such laws, there are difficulties which appear rapidly, in particular the need to estimate the element: $E\left(C_{i,I-1}^{3/2} \mid C_{i,1},...,C_{i,I-2} \right)$.

Considering the above difficulty, in order to estimate the skewness or the kurtosis of non-life reserves by accident year, it is therefore necessary to create a stochastic model. This model will have the characteristics described in this paragraph so that the Best Estimate and standard deviations match the moments coming from the Mack model (1993a). For the triangle below, we will fit the $\hat{C}_{i,k}$ to a Generalized Pareto Distribution (hereinafter "GPD") with parameters $(\mu_{i,k}; s_{i,k}; \zeta_k)$ with $\mu_{i,k} \in (-\infty;\infty)$, the location parameter, $s_{i,k} \in (0;\infty)$, the scale parameter and the shape parameter $\zeta_k \in (-\infty;\infty)$. As we will see below, the shape parameter can be related to either the asymmetry of this distribution or to the kurtosis of the distribution, which according to paragraphs 2.3 and 3, depends only on k.

The stochastic model is iterative: At step k (corresponding to the column k of the below triangle), all the parameters are known. On the basis of these parameters known at step k, the parameters of step k+1 are estimated using the formulas shown in this section.

$$\hat{C}_{i,k} \propto GPD(\mu_{i,k}; s_{i,k}; \zeta_k)$$
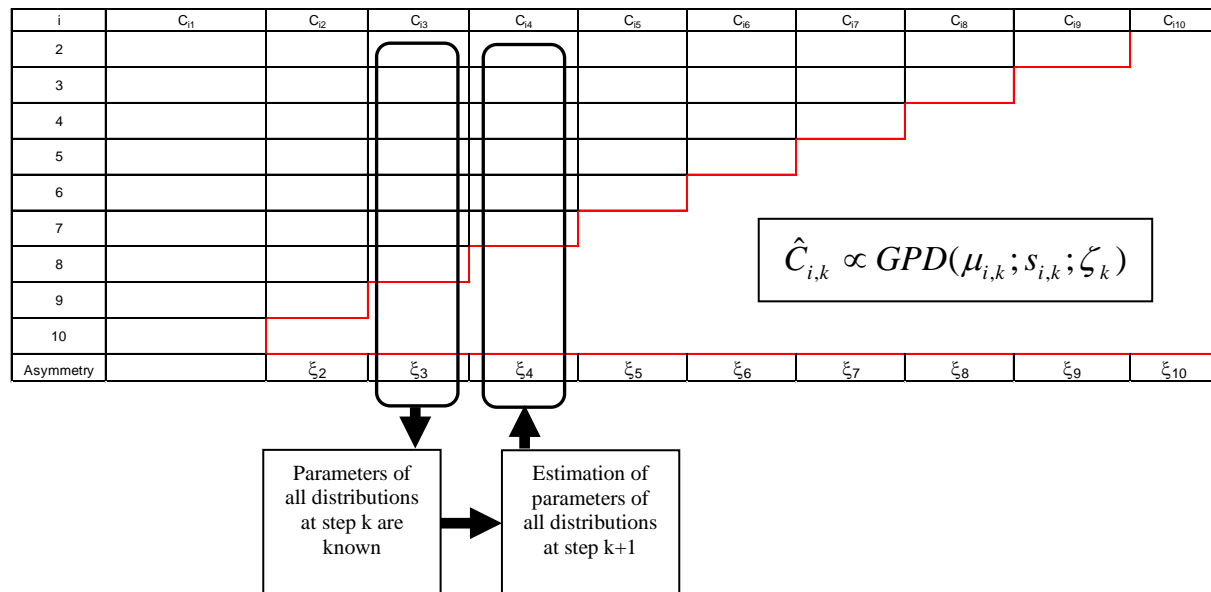
Note: The chosen distribution (GPD) is merely a medium to use the known skewness $\hat{Sk}_k^3$ or the known kurtosis $\hat{Kt}_k^4$ of a development year so as to estimate an accident year skewness / kurtosis.

With the above parameters, we have the following results:

- $\hat{Mean}_{i,k} = \mu_{i,k} + \dfrac{s_{i,k}}{1-\zeta_k} = E\!\left(\hat{C}_{i,k}\right)$

- $\hat{Variance}_{i,k} = \dfrac{s_{i,k}^2}{(1-\zeta_k)^2(1-2\zeta_k)} = Var\!\left(\hat{C}_{i,k}\right)\ \left(\text{with } \zeta_k < 1/2\right)$

- $Skewness\!\left(\hat{C}_{i,k}\right) = \dfrac{SK\!\left(\hat{C}_{i,k}\right)}{Var\!\left(\hat{C}_{i,k}\right)^{3/2}} = \dfrac{2(1+\zeta_k)\sqrt{1-2\zeta_k}}{1-3\zeta_k}$

- $Kurtosis\!\left(\hat{C}_{i,k}\right) = \dfrac{KT\!\left(\hat{C}_{i,k}\right)}{Var\!\left(\hat{C}_{i,k}\right)^2} = \dfrac{3\left(3-5\zeta_k-4\zeta_k^3\right)}{\left(1-7\zeta_k+12\zeta_k^2\right)}$

At step k of the stochastic run (which corresponds to a development year in the triangle above), assuming that we know the 3 parameters $(\mu_{i,k}; s_{i,k}; \zeta_k)$, we can generate the random variable $C_{i,k}$ with the following method:

- Draw a random variable $U_{i,k}$ uniformly distributed on $(0,1]$. We will see later in this section that there is a dependency structure between the $U_{i,k}$ which follows a Gaussian copula that needs to be specified to match the assumptions of Mack (1993a).

- Then: $\hat{C}_{i,k} = \mu_{i,k} + \dfrac{s_{i,k}\left(U_{i,k}^{-\zeta_k}-1\right)}{\zeta_k} \propto GPD(\mu_{i,k}; s_{i,k}; \zeta_k)$

At step k+1, we have the following formulas:

- In the case of skewness evaluation, we have: $\dfrac{\hat{Sk}_{k+1}^3}{\hat{\sigma}_{k+1}^3} = \dfrac{2(1+\zeta_{k+1})\sqrt{1-2\zeta_{k+1}}}{1-3\zeta_{k+1}}$ which allows to estimate the parameter ζ_{k+1}.

  In the case of kurtosis evaluation, we have: $\dfrac{\hat{Kt}_{k+1}^4}{\left(\hat{\sigma}_{k+1}^2\right)^2} = \dfrac{3\left(3-5\zeta_{k+1}-4\zeta_{k+1}^3\right)}{\left(1-7\zeta_{k+1}+12\zeta_{k+1}^2\right)}$

- $\hat{Variance}_{i,k+1} = \hat{C}_{i,k}\hat{\sigma}_k^2\left(1 + \dfrac{\hat{C}_{i,k}}{\sum\limits_{j=1}^{I-k} C_{j,k}}\right)$. The use of this definition allows us to match the overall

  mean squared error (hereinafter "mse") resulting from this simulation with the Mack mse (see also Mack 1999):

$$\hat{mse}\left(\hat{R}_i\right) = \hat{C}_{i,I}\sum_{k=I+1-i}^{I-1}\frac{\hat{\sigma}_k^2}{\hat{f}_k^2}\left(\frac{1}{\hat{C}_{i,k}} + \frac{1}{\sum\limits_{j=1}^{I-k} C_{j,k}}\right) \text{ where } \hat{R}_i = \hat{C}_{i,I} - C_{i,I+1-i}$$

  The knowledge of $\hat{Variance}_{i,k+1}$ and the parameter ζ_{k+1} allows the estimation of the parameter $s_{i,k+1}$.

- Finally, we have: $\hat{C}_{i,k+1} = \hat{f}_k\hat{C}_{i,k} = \mu_{i,k+1} + \dfrac{s_{i,k+1}}{1 - \zeta_{k+1}}$ which allows the estimation of the parameter $\mu_{i,k+1}$.

Gaussian copula:

According to Mack (1993a), the mse of the overall reserve estimate $\hat{R} = \hat{R}_2 + ... + \hat{R}_I$ is equal to:

$$\hat{mse}(\hat{R}) = \sum_{i=2}^{I}\left\{\hat{mse}\left(\hat{R}_i\right) + \hat{C}_{i,I}\left(\sum_{j=i+1}^{I}\hat{C}_{j,I}\right)\sum_{k=I+1-i}^{I-1}\frac{2\hat{\sigma}_k^2/\hat{f}_k^2}{\sum\limits_{n=1}^{I-k} C_{n,k}}\right\}$$

This equation includes an embedded assumption of correlation between the reserve estimates $\hat{R}_i$. A natural way to simulate such correlations is to use a Gaussian copula.

Re-writing the above equation in the usual way, we have:

$$\hat{mse}(\hat{R}) = \sum_{i=2}^{I}\hat{mse}\left(\hat{R}_i\right) + 2\sum\sum_{i<j}\rho_{i,j}\sqrt{\hat{mse}\left(\hat{R}_i\right)\hat{mse}\left(\hat{R}_j\right)}$$

And we can see that the correlations between the reserve estimates $\hat{R}_i$ are:

$$\rho_{i,j}^{Ultimate} = \frac{\hat{C}_{i,I}\hat{C}_{j,I}\sum_{k=I+1-i}^{I-1}\dfrac{\hat{\sigma}_k^2/\hat{f}_k^2}{\sum_{n=1}^{I-k}C_{n,k}}}{\sqrt{\hat{mse}(\hat{R}_i)\hat{mse}(\hat{R}_j)}}$$

These correlations are valid at the ultimate. In the case of our stepwise simulation, it is necessary to have a correction on these correlations so as to reflect the fact that, at step k, the view is not yet on the ultimate. As a matter of fact, if we denote $\rho_{i,j}^{(k)}$ the correlation at step k , it is equal to:

$$\rho_{i,j}^{(k)} = \frac{\hat{C}_{i,k}\hat{C}_{j,k}\sum_{m=I+1-i}^{I-1}\dfrac{\hat{\sigma}_m^2/\hat{f}_m^2}{\sum_{n=1}^{I-m}C_{n,m}}}{\sqrt{\hat{mse}(\hat{C}_{i,k})\hat{mse}(\hat{C}_{j,k})}}$$

where:

$$\hat{mse}(\hat{C}_{i,k}) = \hat{C}_{i,k}\sum_{m=I+1-i}^{k-1}\frac{\hat{\sigma}_m^2}{\hat{f}_m^2}\left(\frac{1}{\hat{C}_{i,m}} + \frac{1}{\sum_{j=1}^{I-m}C_{j,k}}\right)$$

In order to have a correct correlation on an ultimate view, we have the following:

$$Correction_{i,j}^{(k)} = \frac{\rho_{i,j}^{(k)}}{\rho_{i,j}^{Ultimate}} = \frac{1}{\prod_{m=k}^{I}\hat{f}_m^2}\sqrt{\frac{\hat{mse}(\hat{R}_i)\hat{mse}(\hat{R}_j)}{\hat{mse}(\hat{C}_{i,k})\hat{mse}(\hat{C}_{j,k})}}$$

The correlations $\rho_{i,j}^{Ultimate}\times Correction_{i,j}^{(k)}$ are the characteristics of the Gaussian copulas that are used in our stepwise simulation.

5. NUMERICAL EXAMPLES

In Appendix B, the triangles used for the application of the above methodology are shown. In this section, we will therefore focus on the results and their interpretations.

In the table below, for each triangle (see Appendix B) and each development year, the chain-ladder coefficients, the Mack variance, the skewness and the kurtosis are provided:

		k	1	2	3	4	5	6	7	8	9
Private Motor	Farmers Alliance	$\hat{f}_k$	1.076	0.992	0.994	0.996	0.996	1.001	1.001	0.995	1.000
		$\hat{\sigma}_k^2$	52.813	7.223	3.966	2.462	2.311	0.946	0.199	0.161	0.130
		$\hat{Sk}_k^3/(\hat{\sigma}_k^2)^{3/2}$	0.611	-0.256	-0.349	-0.090	1.049	0.477	0.273	NA	NA
		$\hat{Kt}_k^4/(\hat{\sigma}_k^2)^2$	317.02%	177.30%	166.23%	146.75%	340.69%	165.38%	NA	NA	NA
Private Motor	NC Farm Bureau	$\hat{f}_k$	1.290	1.038	0.997	0.988	0.991	0.998	0.998	0.998	0.996
		$\hat{\sigma}_k^2$	412.435	55.605	12.434	2.448	0.685	1.093	0.058	0.605	0.058
		$\hat{Sk}_k^3/(\hat{\sigma}_k^2)^{3/2}$	0.703	0.412	0.727	-0.047	-0.058	-0.769	0.500	NA	NA
		$\hat{Kt}_k^4/(\hat{\sigma}_k^2)^2$	223.54%	201.65%	182.26%	78.01%	144.84%	237.28%	NA	NA	NA
Private Motor	New Jersey Manufacturers	$\hat{f}_k$	1.331	1.158	1.070	0.990	0.980	0.990	0.994	1.001	1.000
		$\hat{\sigma}_k^2$	105.139	45.008	14.086	26.096	4.394	2.036	0.133	0.829	0.133
		$\hat{Sk}_k^3/(\hat{\sigma}_k^2)^{3/2}$	0.583	0.187	0.414	-0.565	-0.141	0.230	-0.008	NA	NA
		$\hat{Kt}_k^4/(\hat{\sigma}_k^2)^2$	204.32%	229.15%	207.06%	192.21%	103.56%	104.25%	NA	NA	NA
Product Liability	Pennsylvania	$\hat{f}_k$	1.451	1.407	1.091	0.963	1.046	1.051	0.950	0.992	1.030
		$\hat{\sigma}_k^2$	190.687	220.494	8.509	155.235	10.557	3.010	4.250	0.094	0.002
		$\hat{Sk}_k^3/(\hat{\sigma}_k^2)^{3/2}$	0.774	1.716	-0.540	-1.059	-0.620	0.164	0.111	NA	NA
		$\hat{Kt}_k^4/(\hat{\sigma}_k^2)^2$	324.61%	620.07%	298.72%	369.19%	164.25%	119.36%	NA	NA	NA
Product Liability	West Bend	$\hat{f}_k$	1.692	1.487	1.269	1.016	1.150	1.130	0.862	1.007	1.000
		$\hat{\sigma}_k^2$	31.078	66.326	70.197	33.319	23.011	3.421	14.919	0.015	0.000
		$\hat{Sk}_k^3/(\hat{\sigma}_k^2)^{3/2}$	-0.008	1.060	0.525	-0.507	-0.030	-0.484	-0.113	NA	NA
		$\hat{Kt}_k^4/(\hat{\sigma}_k^2)^2$	205.13%	411.28%	302.64%	200.90%	125.54%	121.50%	NA	NA	NA
Mack 1993		$\hat{f}_k$	3.491	1.747	1.457	1.174	1.104	1.086	1.054	1.077	1.018
		$\hat{\sigma}_k^2$	160280	37737	41965	15183	13731	8186	447	1147	447
		$\hat{Sk}_k^3/(\hat{\sigma}_k^2)^{3/2}$	0.137	0.215	0.638	-0.433	0.402	-0.026	-0.497	NA	NA
		$\hat{Kt}_k^4/(\hat{\sigma}_k^2)^2$	184.92%	170.29%	265.62%	162.92%	185.65%	123.00%	NA	NA	NA

Table 1: Chain-ladder coefficients, Mack variance, Skewness and Kurtosis estimators for 10 x 10 triangles

Kurtosis and skewness estimation for non-life reserve risk distribution

		k	1	2	3	4	5	6	7	8	9	10	11	12	13	14
SCOR	Casualty Prop	$\hat{f}_k$	6.82	1.22	1.04	1.01	0.99	1.01	1.02	1.00	1.00	1.02	1.00	1.01	0.99	1.00
		$\hat{\sigma}_k^2$	131'803'806	459'208	565'956	323'842	278'372	120'753	75'674	34'013	15'870	98'673	5'598	429	122	35
		$\hat{Sk}_k^3/(\hat{\sigma}_k^2)^{3/2}$	1.031	0.070	-0.372	-0.119	-0.390	0.537	-0.334	0.863	-0.842	0.868	-0.843	0.518	NA	NA
		$\hat{Kt}_k^4/(\hat{\sigma}_k^2)^{2}$	233.59%	291.73%	339.24%	193.39%	194.71%	260.08%	226.84%	314.55%	294.64%	266.31%	257.91%	NA	NA	NA
SCOR	Motor NonProp	$\hat{f}_k$	3.17	1.31	1.16	1.10	1.06	1.04	1.03	1.03	1.02	1.03	1.02	1.03	1.02	1.01
		$\hat{\sigma}_k^2$	8'756'735	1'115'133	126'117	349'261	207'405	79'654	70'839	33'789	37'388	24'469	2'983	78'442	3'144	126
		$\hat{Sk}_k^3/(\hat{\sigma}_k^2)^{3/2}$	1.491	0.286	0.348	-0.155	0.621	0.177	0.920	0.411	0.658	0.713	0.865	-0.212	NA	NA
		$\hat{Kt}_k^4/(\hat{\sigma}_k^2)^{2}$	631.06%	201.17%	403.49%	234.55%	294.37%	167.42%	292.94%	194.28%	291.68%	245.21%	265.76%	NA	NA	NA

Table 2: Chain-ladder coefficients, Mack variance, Skewness and Kurtosis estimators for 15 x 15 triangles

As expected, the skewness is positive in many instances. In addition, the kurtosis is often below 300%. This should reflect the platykurtic nature of the non-life reserves. As a matter of fact, the potential of having extreme values in an incurred triangle is mechanically remote: In general, the case reserves input by the claim adjusters will reflect the standard movements of all the claims. Only in rare cases could there be an extreme event in a triangle and these extreme events will be present on specific lines of business like Property NatCat. In addition, an extreme event may have a smaller impact in the triangle if the portfolio of risk is big enough to absorb the unusual movement. As a consequence, the distribution should be the result of frequent modestly sized deviations and have a platykurtic nature.

Following the analysis by development year, we now look at the results of the simulations below:

Triangle size	LoB	Company	Chain-ladder reserves	Chain-ladder stdev	CoV	Overall simulated skewness	Overall simulated kurtosis	LogN				Gamma			
								Mu	Sigma2	Resulting skewness	Resulting kurtosis	k	Theta	Resulting skewness	Resulting kurtosis
10 x 10	Private Motor	Farmers Alliance	-374	1493	-400%	-0.01	297%	NA	NA	NA	NA	0.06	-5'966	7.99	9882%
10 x 10	Private Motor	NC Farm	19'415	9'528	49%	0.32	298%	9.77	0.216	1.59	781%	4.15	4'676	0.98	445%
10 x 10	Private Motor	New Jersey Manuf.	109'719	11'961	11%	0.07	295%	11.60	0.012	0.33	319%	84.15	1'304	0.22	307%
10 x 10	Product Liab.	Pennsylvania	1'474	1'784	121%	0.06	350%	6.84	0.903	5.41	8222%	0.68	2'161	2.42	1180%
10 x 10	Product Liab.	West Bend	2150	1899	88%	0.35	384%	7.38	0.577	3.34	2784%	1.28	1'678	1.77	768%
10 x 10	Mack 1993 triangle		18'680'856	2'447'095	13%	0.13	292%	16.73	0.017	0.40	328%	58.28	320'557	0.26	310%
15 x 15	WW Casualty Prop	SCOR	219'461'925	79'722'452	36%	0.14	300%	19.14	0.124	1.14	539%	7.58	28960237	0.73	379%
15 x 15	WW Motor NP	SCOR	402'645'321	53'078'447	13%	0.17	289%	19.80	0.017	0.40	328%	57.55	6997030	0.26	310%

Table 3: Simulation (10 000 scenarios) of skewness and kurtosis of the overall reserves for the triangles shown in Appendix B

Note on Table 3: CoV stands for Coefficient of Variation and equals the standard deviation divided by the reserves.

Looking at the above table, we can see that the higher the CoV, the more the Lognormal and the Gamma distributions have their skewness and kurtosis far away from the simulated skewness and

kurtosis. Such deviations between the simulated distribution and Lognormal or Gamma distribution starts to be seen for CoV as low as 36% which are not uncommon CoVs in practice. For CoVs higher than 36%, the deviations get bigger and bigger. In such cases, the Lognormal and Gamma distributions do not fit anymore with the third and fourth simulated moments.

In order to estimate the impact of the deviations between the simulated distribution and the Lognormal distribution, we followed the steps below:

1. Fit the simulated distribution to a Johnson distribution (see Johnson 1949) on the first 4 moments using the software R (function JohnsonFit).

2. Estimate the Value At Risk (VaR) at 99% resulting from the application of the Johnson distribution.

3. Compare the above VaR with the VaR resulting from the Lognormal distribution fitted on the basis of the Best Estimate and CoV.

We recall that the family of Johnson distributions has the following properties (see also Johnson 1949):

$z = \gamma + \delta \, f\left(\dfrac{x - \xi}{\lambda}\right)$ where f is a function of simple form and z is a unit normal variable.

Depending on f, the Johnson distribution is noted as follows:

$f = \log$: Distribution SL

$f = \sinh^{-1}$: Distribution SU

$z = \gamma + \delta \log\left(\dfrac{x - \xi}{\xi + \lambda - x}\right)$: Distribution SB

$z = \gamma + \delta\left(\dfrac{x - \xi}{\lambda}\right)$: Distribution SN

As described above, the results of the fitting of the simulated distributions with the Johnson family on the basis of the first 4 moments are shown in the table below:

LoB	Company	Chain-ladder reserves	Chain-ladder stdev	CoV	Overall simulated skewness	Overall simulated kurtosis	Johnson fitting									
							Type	Gamma	Delta	Xi	Lambda	VaR 99%	Fitted Mean	Fitted Stdev	Fitted Skewness	Fitted Kurtosis
Private Motor	Farmers Alliance	-374	1493	-400%	-0.01	297%	SN	-	1.000	-374	1'493	3'099	-374	1'493	-	300%
Private Motor	NC Farm	19'415	9'528	49%	0.32	298%	SB	2.253	3.040	-24'390	134'033	43'432	19'355	9'406	0.19	287%
Private Motor	New Jersey Manuf.	109'719	11'961	11%	0.07	295%	SN	-	1.000	109'719	11'961	137'544	109'719	11'961	-	300%
Product Liab.	Pennsylvania	1'474	1'784	121%	0.06	350%	SU	-0.174	3.085	1'165	5'208	5'864	1'474	1'784	0.06	350%
Product Liab.	West Bend	2150	1899	88%	0.35	384%	SU	-0.785	2.715	707	4'603	7'214	2'150	1'899	0.35	348%
Mack 1993 triangle		18'680'856	2'447'095	13%	0.13	292%	SB	1.161	3.370	4'224'926	34'719'087	24'555'541	18'645'236	2'428'748	0.18	278%
WW Casualty Prop	SCOR	219'461'925	79'722'452	36%	0.14	300%	SL	-456.116	21.456	-1'490'114'663	1.000	411'994'159	219'461'925	79'722'452	0.14	303%
WW Motor NP	SCOR	402'645'321	53'078'447	13%	0.17	289%	SB	1.241	3.043	125'267'964	690'347'082	531'340'556	401'885'733	52'707'797	0.21	278%

Table 4: Results of the fitting of simulated distributions to Johnson distributions

Generally, the fitted distribution shows moments which reconcile pretty well with the simulated distributions. As a next step, the table below provides the comparison of the VaR 99% coming from the Lognormal and the Johnson distributions:

LoB	Company	VaR 99%		Difference LogN Johnson VaR 99%
		Johnson	Lognormal	
Private Motor	Farmers Alliance	3'099	NA	NA
Private Motor	NC Farm	43'432	51'358	18%
Private Motor	New Jersey Manuf.	137'544	140'453	2%
Product Liab.	Pennsylvania	5'864	8'556	46%
Product Liab.	West Bend	7'214	9'430	31%
Mack 1993 triangle		24'555'541	25'089'172	2%
WW Casualty Prop	SCOR	411'994'159	467'889'645	14%
WW Motor NP	SCOR	531'340'556	541'742'729	2%

Table 5: Comparison of VaR 99% coming from Johnson and Lognormal distributions

As expected, when the CoVs are relatively high, the capital requirements under the risk measure Value at Risk 99% using the lognormal assumption are much higher than the capital requirements using the Johnson distribution.

Finally, the graph below shows how far the lognormal distribution can be from the simulated distribution (shown as an histogram) and from the Johnson distribution in the case of the Product Liability triangle of West Bend:
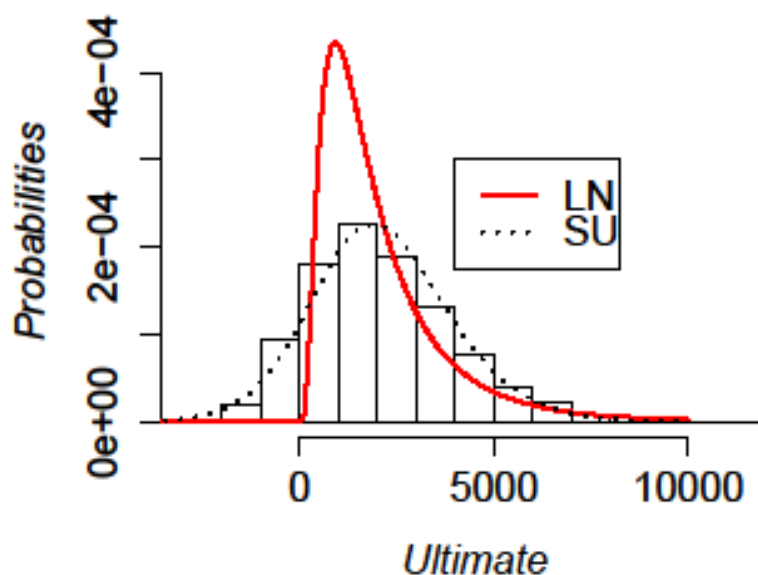
It is interesting to note that the simulations as well as the Johnson fitted distribution anticipate some possibilities for negative IBNRs. Looking into the specificities of the West Bend Product Liability triangle, it can be seen that the chain-ladder coefficient between development year N+6 and N+7 is equal to 0.86. Therefore, in certain cases, the company West Bend seems to have positive reserve development on this line of business. As mentioned, this is reflected in both the simulations and the Johnson distribution. In the case of the simulations, the maximum level of negative IBNRs is 2322 while the case reserves are 2683. Even though it is unlikely that there would be 2322 negative IBNRs on the Product Liability line of West Bend, this maximum level of negative IBNRs is consistent with the amount of case reserve in a "best case" scenario. As a conclusion on this case, we can see that the use of the Lognormal distribution does not anticipate the possibilities of negative IBNRs which is not consistent with the past results of West Bend. Therefore, the use of the Lognormal distribution should be avoided on this case.

6. CONCLUSIONS

Based on the results of simulations on eight triangles, it appears that the representation of the reserve risk distribution by a Lognormal or a Gamma distribution can have major drawbacks in the case of CoV of 36% and above. In particular, the use of such distribution is likely to bring higher and unnecessary capital requirements under VaR 99%. In order to avoid such unnecessary capital requirement, it is advisable to turn to other distributions fitting the skewness and kurtosis estimated through the simulations described in this paper. One possible choice of such distributions fitting with the simulated skewness and kurtosis could be the Johnson distribution.

7. REFERENCES

[1]   ARBENZ P., SALZMANN R., 2010: "A robust distribution-free loss reserving method with weighted data- and expert-reliance", http://www.risklab.ch/hclmethod

[2]   CRAMER H., 1946: "Mathematical methods of statistics", Princeton: Princeton University Press

[3]   DAL MORO ERIC, 2012 "Application of skewness to non-life reserving", Paper presented to the ASTIN Colloquium of Mexico-City, 2 October 2012

[4]   ENGLAND, P.D., VERRALL, R.J., 2002: "Stochastic claims reserving in general insurance", Paper presented to the Institute of Actuaries, 28 January 2002

[5]   ENGLAND, P.D., VERRALL, R.J., 2006: "Predictive distributions of outstanding liabilities in General Insurance", Annals of Actuarial Science, Volume 1, Issue 02, September 2006, pp 221-270

[6]   JOHNSON N.L., 1949: "Systems of Frequency Curves Generated by Methods of Translation", Biometrika, Vol 36 No 1/2

[7]   MACK Thomas, 1993a : "Distribution-free calculation of the standard error of chain-ladder reserve estimates", ASTIN Bulletin Vol 23, No2, 1993

[8]   MACK Thomas, 1993b : "Measuring the variability of Chain-Ladder Reserve Estimates", Casualty Actuarial Society Prize Paper Competition

[9]   MACK Thomas, 1999: "The standard error of chain-ladder reserve estimates, Recursive calculation and inclusion of a tail factor", ASTIN Bulletin Vol 29, No2, 1999

[10] MACK Thomas, 2008 : "The Prediction Error of Bornhuetter/Ferguson", ASTIN Bulletin Vol 38, No1, 2008

[11] WÜTHRICH, M. V., MERZ, M., 2008: "Stochastic Claims Reserving Methods in Insurance", Wiley, Chichester

Biography of the Author

Eric Dal Moro is chief reserving actuary at SCOR Switzerland. He is a Fellow of the Swiss Actuarial Association and of the French Institute of Actuaries. He is also Vice-Chaiman of the ASTIN Committee.

Contact: eric_dal_moro@yahoo.com

Appendix A

We want to show that $E(\hat{K}t_k^4) = Kt_k^4$. Therefore, we will, in a first step, concentrate on $E(\hat{K}t_k^4 \mid B_k)$. We have:

$$(Eq) = E\left[\left(\left(\sum_{i=1}^{I-k}\left(1 - \frac{C_{i,k}}{\sum_{i=1}^{I-k}C_{i,k}}\right)\right)^4 + \frac{\left(\sum_{i=1}^{I-k}C_{i,k}^2\right)^2 - \sum_{i=1}^{I-k}C_{i,k}^4}{\left(\sum_{i=1}^{I-k}C_{i,k}\right)^4}\right)\hat{K}t_k^4 \mid B_k\right]$$

$$= E\left[\sum_{i=1}^{I-k}C_{i,k}^2\left(\frac{C_{i,k+1}}{C_{i,k}} - \hat{f}_k\right)^4 - 3(\sigma_k^2)^2\left(2 - 6\frac{\sum_{i=1}^{I-k}C_{i,k}^2}{\left(\sum_{i=1}^{I-k}C_{i,k}\right)^2} + 4\frac{\sum_{i=1}^{I-k}C_{i,k}^3}{\left(\sum_{i=1}^{I-k}C_{i,k}\right)^3}\right) \mid B_k\right] \text{ for } 1 \le k \le I\text{-}4$$

where $B_k = \{C_{i,j} \mid i + j \le I + 1, j \le k\}, 1 \le k \le I\text{-}1$

In the following, we will concentrate on the term:

$$(Eq1) = E\left[\sum_{i=1}^{I-k}C_{i,k}^2\left(\frac{C_{i,k+1}}{C_{i,k}} - \hat{f}_k\right)^4 \mid B_k\right] \text{ for } 1 \le k \le I\text{-}4$$

$$= \sum_{i=1}^{I-k}\left[\frac{1}{C_{i,k}^2}E(C_{i,k+1}^4 \mid B_k)\right] - 4\sum_{i=1}^{I-k}\left[\frac{1}{C_{i,k}}E(\hat{f}_kC_{i,k+1}^3 \mid B_k)\right] + 6\sum_{i=1}^{I-k}\left[E(\hat{f}_k^2C_{i,k+1}^2 \mid B_k)\right] \qquad (9)$$

$$- 4\sum_{i=1}^{I-k}C_{i,k}E(\hat{f}_k^3C_{i,k+1} \mid B_k) + \sum_{i=1}^{I-k}C_{i,k}^2E(\hat{f}_k^4 \mid B_k)$$

For each of the 5 elements of the above equation, a formula is going to be established. For the first element of equation (9), we have:

$$KT(C_{i,k+1} \mid B_k) = Kt_k^4C_{i,k}^2 = E(C_{i,k+1}^4 \mid B_k) - 4E(C_{i,k+1} \mid B_k)E(C_{i,k+1}^3 \mid B_k) + 6E(C_{i,k+1}^2 \mid B_k)E(C_{i,k+1} \mid B_k)^2$$

$$- 3E(C_{i,k+1} \mid B_k)^4$$

From Mack (1993b), we have:

$$E(C_{i,k+1}^2 \mid B_k) = C_{i,k}\sigma_k^2 + C_{i,k}^2 f_k^2 \text{ and}$$

$$E(C_{i,k+1} \mid B_k) = C_{i,k}f_k$$

We also established that:

$$E\left(C_{i,k+1}^3 \mid B_k\right) = Sk_k^3 C_{i,k}^{3/2} + 3f_k C_{i,k}^2 \sigma_k^2 + C_{i,k}^3 f_k^3 \tag{10}$$

Therefore, we have:

$$E\left(C_{i,k+1}^4 \mid B_k\right) = Kt_k^4 C_{i,k}^2 + 4f_k Sk_k^3 C_{i,k}^{5/2} + 6f_k^2 C_{i,k}^3 \sigma_k^2 + C_{i,k}^4 f_k^4 \tag{11}$$

For the fifth element of equation (9), we have:

$$KT\left(\hat{f}_k \mid B_k\right) = \frac{1}{\left(\sum_{i=1}^{I-k} C_{i,k}\right)^4} \sum_{i=1}^{I-k} KT\left(C_{i,k+1} \mid B_k\right) = \frac{Kt_k^4}{\left(\sum_{i=1}^{I-k} C_{i,k}\right)^4} \sum_{i=1}^{I-k} C_{i,k}^2$$

$$= E\left(\left[\hat{f}_k - E\left(\hat{f}_k \mid B_k\right)\right]^4 \mid B_k\right) = E\left(\hat{f}_k^4 \mid B_k\right) - 4E\left(\hat{f}_k \mid B_k\right)E\left(\hat{f}_k^3 \mid B_k\right) + 6E\left(\hat{f}_k \mid B_k\right)^2 E\left(\hat{f}_k^2 \mid B_k\right) - 3E\left(\hat{f}_k \mid B_k\right)^4$$

From Mack (1993b), we have:

$$E\left(\hat{f}_k^2 \mid B_k\right) = \frac{\sigma_k^2}{\sum_{i=1}^{I-k} C_{i,k}} + f_k^2$$

We also know that:

$$E\left(\hat{f}_k^3 \mid B_k\right) = \frac{Sk_k^3}{\left(\sum_{i=1}^{I-k} C_{i,k}\right)^3} \sum_{i=1}^{I-k} C_{i,k}^{3/2} + 3\frac{f_k \sigma_k^2}{\sum_{i=1}^{I-k} C_{i,k}} + f_k^3$$

Therefore, we have:

$$E\left(\hat{f}_k^4 \mid B_k\right) = Kt_k^4 \frac{\sum_{i=1}^{I-k} C_{i,k}^2}{\left(\sum_{i=1}^{I-k} C_{i,k}\right)^4} + 4f_k Sk_k^3 \frac{\sum_{i=1}^{I-k} C_{i,k}^{3/2}}{\left(\sum_{i=1}^{I-k} C_{i,k}\right)^3} + 6\frac{f_k^2 \sigma_k^2}{\sum_{i=1}^{I-k} C_{i,k}} + f_k^4 \tag{12}$$

For the second element of equation (9), we have:

$$E\left(\hat{f}_k C_{i,k+1}^3 \mid B_k\right) = \frac{1}{\sum_{j=1}^{I-k} C_{j,k}} E\left[\sum_{\substack{j=1 \\ j \neq i}}^{I-k} \left(C_{j,k+1} C_{i,k+1}^3\right) + C_{i,k+1}^4 \mid B_k\right]$$

Due to the independence of $C_{i,k+1}$ and $C_{j,k+1}$, we have:

$$E\left(\hat{f}_k C_{i,k+1}^3 \mid B_k\right) = \frac{1}{\sum\limits_{j=1}^{I-k} C_{j,k}}\left[\sum\limits_{\substack{j=1 \\ j\neq i}}^{I-k} E\left(C_{j,k+1}\right)E\left(C_{i,k+1}^3 \mid B_k\right) + E\left(C_{i,k+1}^4 \mid B_k\right)\right]$$

$$= \frac{1}{\sum\limits_{j=1}^{I-k} C_{j,k}}\left[\sum\limits_{\substack{j=1 \\ j\neq i}}^{I-k} f_k C_{j,k}\left(Sk_k^3 C_{i,k}^{3/2} + 3f_k C_{i,k}^2 \sigma_k^2 + C_{i,k}^3 f_k^3\right) + Kt_k^4 C_{i,k}^2 + 4f_k Sk_k^3 C_{i,k}^{5/2} + 6f_k^2 C_{i,k}^3 \sigma_k^2 + C_{i,k}^4 f_k^4\right]$$

$$\Rightarrow E\left(\hat{f}_k C_{i,k+1}^3 \mid B_k\right) = \frac{Kt_k^4 C_{i,k}^2}{\sum\limits_{j=1}^{I-k} C_{j,k}} + \frac{3f_k^2 C_{i,k}^3 \sigma_k^2}{\sum\limits_{j=1}^{I-k} C_{j,k}} + \frac{3f_k Sk_k^3 C_{i,k}^{5/2}}{\sum\limits_{j=1}^{I-k} C_{j,k}} + f_k^4 C_{i,k}^3 + f_k Sk_k^3 C_{i,k}^{3/2} + 3f_k^2 \sigma_k^2 C_{i,k}^2 \quad (13)$$

For the third element of equation (9), we have:

$$E\left(\hat{f}_k^2 C_{i,k+1}^2 \mid B_k\right) = \frac{1}{\left(\sum\limits_{j=1}^{I-k} C_{j,k}\right)^2} E\left[C_{i,k+1}^2 \left(\sum\limits_{j=1}^{I-k} C_{j,k+1}\right)^2 \mid B_k\right]$$

$$= \frac{1}{\left(\sum\limits_{j=1}^{I-k} C_{j,k}\right)^2} E\left[C_{i,k+1}^2 \left(\sum\limits_{j=1}^{I-k} C_{j,k+1}^2 + 2\sum\limits_{n=1}^{I-k} \sum\limits_{\substack{m=1 \\ m\neq n}}^{n} C_{n,k+1} C_{m,k+1}\right) \mid B_k\right]$$

$$= \frac{1}{\left(\sum\limits_{j=1}^{I-k} C_{j,k}\right)^2} E\left[C_{i,k+1}^4 + \sum\limits_{\substack{j=1 \\ j\neq i}}^{I-k} C_{j,k+1}^2 C_{i,k+1}^2 + 2\sum\limits_{\substack{n=1 \\ n\neq i}}^{I-k} \sum\limits_{\substack{m=1 \\ m\neq n \\ m\neq i}}^{n} C_{n,k+1} C_{m,k+1} C_{i,k+1}^2 + 2\sum\limits_{n=1}^{I-k} C_{n,k+1} C_{i,k+1}^3 \mid B_k\right]$$

Due to the independence of $C_{i,k+1}$, $C_{j,k+1}$, $C_{n,k+1}$ and $C_{m,k+1}$ and on using the formulas for $E\left(C_{i,k+1}^4 \mid B_k\right)$ (equation (11)), $E\left(C_{i,k+1}^3 \mid B_k\right)$ (equation (10)) and $E\left(C_{i,k+1}^2 \mid B_k\right)$, we have:

$$E\left(\hat{f}_k^2 C_{i,k+1}^2 \mid B_k\right) = \frac{Kt_k^4 C_{i,k}^2}{\left(\sum_{j=1}^{I-k} C_{j,k}\right)^2} + \frac{5 f_k^2 C_{i,k}^2 \sigma_k^2}{\sum_{j=1}^{I-k} C_{j,k}} + \frac{2 f_k Sk_k^3 C_{i,k}^{5/2}}{\left(\sum_{j=1}^{I-k} C_{j,k}\right)^2} + \frac{2 f_k Sk_k^3 C_{i,k}^{3/2}}{\sum_{j=1}^{I-k} C_{j,k}} + f_k^4 C_{i,k}^2 + f_k^2 \sigma_k^2 C_{i,k}$$

$$+ \frac{\left(\sigma_k^2\right)^2 C_{i,k}\left(\sum_{j=1}^{I-k} C_{j,k} - C_{i,k}\right)}{\left(\sum_{j=1}^{I-k} C_{j,k}\right)^2} \tag{14}$$

Finally, for the fourth element of equation (9), we have:

$$E\left(\hat{f}_k^3 C_{i,k+1} \mid B_k\right) = \frac{1}{\left(\sum_{j=1}^{I-k} C_{j,k}\right)^3} E\left[C_{i,k+1}\left(\sum_{j=1}^{I-k} C_{j,k+1}\right)^3 \mid B_k \right]$$

$$= \frac{1}{\left(\sum_{j=1}^{I-k} C_{j,k}\right)^3} E\left[C_{i,k+1}\left(\sum_{j=1}^{I-k} C_{j,k+1}^3 + 3\sum_{n=1}^{I-k}\sum_{\substack{m=1 \\ m\neq n}}^{I-k} C_{n,k+1}^2 C_{m,k+1} + 6\sum_{n=1}^{I-k-2}\sum_{\substack{m=n+1 \\ m\neq n}}^{I-k-1}\sum_{\substack{o=m+1 \\ m\neq o \\ n\neq o}}^{I-k} C_{n,k+1} C_{m,k+1} C_{o,k+1} \right) \mid B_k \right]$$

$$= \frac{1}{\left(\sum_{j=1}^{I-k} C_{j,k}\right)^3} E\left[\begin{array}{l} C_{i,k+1}^4 + \sum_{\substack{j=1 \\ j\neq i}}^{I-k} C_{j,k+1}^3 C_{i,k+1} + 3\sum_{\substack{n=1 \\ n\neq i}}^{I-k} C_{n,k+1} C_{i,k+1}^3 + 3\sum_{\substack{j=1 \\ j\neq i}}^{I-k} C_{j,k+1}^2 C_{i,k+1}^2 + 3\sum_{\substack{n=1 \\ n\neq i}}^{I-k}\sum_{\substack{m=1 \\ m\neq n \\ m\neq i}}^{n} C_{n,k+1}^3 C_{m,k+1} C_{i,k+1} \\[2em] 6\sum_{\substack{n=1 \\ n\neq i}}^{I-k-2}\sum_{\substack{m=n+1 \\ m\neq n \\ m\neq i}}^{I-k-1}\sum_{\substack{o=m+1 \\ m\neq o \\ n\neq o \\ o\neq n}}^{I-k} C_{n,k+1} C_{m,k+1} C_{o,k+1} C_{i,k+1} + 6\sum_{\substack{n=1 \\ n\neq i}}^{I-k-1}\sum_{\substack{m=1 \\ m\neq n \\ m\neq i}}^{I-k} C_{i,k+1}^2 C_{n,k+1} C_{m,k+1} \mid B_k \end{array} \right]$$

Due to the independence of $C_{i,k+1}$, $C_{j,k+1}, C_{n,k+1}, C_{m,k+1}$ and $C_{o,k+1}$ and on using the formulas for $E\left(C_{i,k+1}^4 \mid B_k\right)$ (equation (11)), $E\left(C_{i,k+1}^3 \mid B_k\right)$ (equation (10)) and $E\left(C_{i,k+1}^2 \mid B_k\right)$, we have:

$$E\left(\hat{f}_k^3 C_{i,k+1} \mid B_k\right) = \frac{1}{\left(\sum_{j=1}^{I-k} C_{j,k}\right)^3} \left[\begin{array}{l} Kt_k^4 C_{i,k}^2 + f_k^4 C_{i,k}\left(\sum_{j=1}^{I-k} C_{j,k}\right)^3 + f_k Sk_k^3\left(C_{i,k}\sum_{j=1}^{I-k} C_{j,k}^{3/2} + 3C_{i,k}^{3/2}\sum_{j=1}^{I-k} C_{j,k}\right) \\ + 6 f_k^2 C_{i,k}\sigma_k^2\left(\sum_{j=1}^{I-k} C_{j,k}\right)^2 + 3\left(\sigma_k^2\right)^2 C_{i,k}\left(\sum_{j=1}^{I-k} C_{j,k} - C_{i,k}\right) \end{array}\right] (15)$$

On inserting equations (11), (12), (13), (14) and (15) into equation (9), we have:

$$(Eq1) = \sum_{i=1}^{I-k}\left[\frac{1}{C_{i,k}^2}\left(Kt_k^4 C_{i,k}^2 + 4 f_k Sk_k^3 C_{i,k}^{5/2} + 6 f_k^2 C_{i,k}^3\sigma_k^2 + C_{i,k}^4 f_k^4\right)\right]$$

$$-4\sum_{i=1}^{I-k}\left[\frac{1}{C_{i,k}}\left(\frac{Kt_k^4 C_{i,k}^2}{\sum_{j=1}^{I-k} C_{j,k}} + \frac{3 f_k^2 C_{i,k}^3\sigma_k^2}{\sum_{j=1}^{I-k} C_{j,k}} + \frac{3 f_k Sk_k^3 C_{i,k}^{5/2}}{\sum_{j=1}^{I-k} C_{j,k}} + f_k^4 C_{i,k}^3 + f_k Sk_k^3 C_{i,k}^{3/2} + 3 f_k^2\sigma_k^2 C_{i,k}^2\right)\right]$$

$$+6\sum_{i=1}^{I-k}\left[\begin{array}{l} \frac{Kt_k^4 C_{i,k}^2}{\left(\sum_{j=1}^{I-k} C_{j,k}\right)^2} + \frac{5 f_k^2 C_{i,k}^2\sigma_k^2}{\sum_{j=1}^{I-k} C_{j,k}} + \frac{2 f_k Sk_k^3 C_{i,k}^{5/2}}{\left(\sum_{j=1}^{I-k} C_{j,k}\right)^2} + \frac{2 f_k Sk_k^3 C_{i,k}^{3/2}}{\sum_{j=1}^{I-k} C_{j,k}} + f_k^4 C_{i,k}^2 + f_k^2\sigma_k^2 C_{i,k} \\ \\ + \frac{\left(\sigma_k^2\right)^2 C_{i,k}\left(\sum_{j=1}^{I-k} C_{j,k} - C_{i,k}\right)}{\left(\sum_{j=1}^{I-k} C_{j,k}\right)^2} \end{array}\right]$$

$$-\frac{4}{\left(\sum_{j=1}^{I-k} C_{j,k}\right)^3}\sum_{i=1}^{I-k} C_{i,k}\left[\begin{array}{l} Kt_k^4 C_{i,k}^2 + f_k^4 C_{i,k}\left(\sum_{j=1}^{I-k} C_{j,k}\right)^3 + f_k Sk_k^3\left(C_{i,k}\sum_{j=1}^{I-k} C_{j,k}^{3/2} + 3C_{i,k}^{3/2}\sum_{j=1}^{I-k} C_{j,k}\right) \\ + 6 f_k^2 C_{i,k}\sigma_k^2\left(\sum_{j=1}^{I-k} C_{j,k}\right)^2 + 3\left(\sigma_k^2\right)^2 C_{i,k}\left(\sum_{j=1}^{I-k} C_{j,k} - C_{i,k}\right) \end{array}\right]$$

$$+\sum_{i=1}^{I-k} C_{i,k}^2\left(Kt_k^4\frac{\sum_{i=1}^{I-k} C_{i,k}^2}{\left(\sum_{i=1}^{I-k} C_{i,k}\right)^4} + 4 f_k Sk_k^3\frac{\sum_{i=1}^{I-k} C_{i,k}^{3/2}}{\left(\sum_{i=1}^{I-k} C_{i,k}\right)^3} + 6\frac{f_k^2\sigma_k^2}{\sum_{i=1}^{I-k} C_{i,k}} + f_k^4\right)$$

After simplifications, the above equation is given as:

$$(Eq1) = \left(I - k - 4 + 6 \frac{\sum_{i=1}^{I-k} C_{i,k}^2}{\left(\sum_{i=1}^{I-k} C_{i,k}\right)^2} - 4 \frac{\sum_{i=1}^{I-k} C_{i,k}^3}{\left(\sum_{i=1}^{I-k} C_{i,k}\right)^3} + \frac{\left(\sum_{i=1}^{I-k} C_{i,k}^2\right)^2}{\left(\sum_{i=1}^{I-k} C_{i,k}\right)^4} \right) Kt_k^4 + 3(\sigma_k^2)^2 \left(2 - 6 \frac{\sum_{i=1}^{I-k} C_{i,k}^2}{\left(\sum_{i=1}^{I-k} C_{i,k}\right)^2} + 4 \frac{\sum_{i=1}^{I-k} C_{i,k}^3}{\left(\sum_{i=1}^{I-k} C_{i,k}\right)^3} \right)$$

As we have the equality below:

$$I - k - 4 + 6 \frac{\sum_{i=1}^{I-k} C_{i,k}^2}{\left(\sum_{i=1}^{I-k} C_{i,k}\right)^2} - 4 \frac{\sum_{i=1}^{I-k} C_{i,k}^3}{\left(\sum_{i=1}^{I-k} C_{i,k}\right)^3} + \frac{\left(\sum_{i=1}^{I-k} C_{i,k}^2\right)^2}{\left(\sum_{i=1}^{I-k} C_{i,k}\right)^4} = \sum_{i=1}^{I-k} \left(1 - \frac{C_{i,k}}{\sum_{i=1}^{I-k} C_{i,k}} \right)^4 + \frac{\left(\sum_{i=1}^{I-k} C_{i,k}^2\right)^2 - \sum_{i=1}^{I-k} C_{i,k}^4}{\left(\sum_{i=1}^{I-k} C_{i,k}\right)^4}$$

we finally get the result:

$$\hat{K}t_k^4 = \frac{\left[\sum_{i=1}^{I-k} C_{i,k}^2 \left(\frac{C_{i,k+1}}{C_{i,k}} - \hat{f}_k \right)^4 - 3(\hat{\sigma}_k^2)^2 \left(2 - 6 \frac{\sum_{i=1}^{I-k} C_{i,k}^2}{\left(\sum_{i=1}^{I-k} C_{i,k}\right)^2} + 4 \frac{\sum_{i=1}^{I-k} C_{i,k}^3}{\left(\sum_{i=1}^{I-k} C_{i,k}\right)^3} \right) \right]}{\left(\sum_{i=1}^{I-k} \left(1 - \frac{C_{i,k}}{\sum_{i=1}^{I-k} C_{i,k}} \right)^4 + \frac{\left(\sum_{i=1}^{I-k} C_{i,k}^2\right)^2 - \sum_{i=1}^{I-k} C_{i,k}^4}{\left(\sum_{i=1}^{I-k} C_{i,k}\right)^4} \right)} \quad \text{for } 1 \leq k \leq I\text{-}4$$

Appendix B

Schedule P – Farmers Alliance – Incurred claim – Private motor

	N	N+1	N+2	N+3	N+4	N+5	N+6	N+7	N+8	N+9
1988	10'798	11'595	11'724	11'820	11'746	11'641	11'557	11'552	11'525	11'522
1989	11'313	13'743	13'621	13'666	13'352	13'182	13'186	13'159	13'061	
1990	15'110	15'143	15'401	14'915	14'998	14'858	14'811	14'887		
1991	15'163	15'253	14'577	14'269	14'456	14'721	14'898			
1992	14'232	14'999	14'932	14'933	14'915	14'788				
1993	14'063	15'468	15'052	15'263	15'042					
1994	12'050	12'907	13'156	13'016						
1995	12'163	13'086	12'847							
1996	11'624	13'122								
1997	12'942									

Schedule P – NC Farm Bureau – Incurred claim – Private Motor

	N	N+1	N+2	N+3	N+4	N+5	N+6	N+7	N+8	N+9
1988	48900	65079	70815	71880	71384	71034	70442	70383	70379	70127
1989	55918	80979	85951	87272	86546	85784	85620	85427	85084	
1990	63493	88613	93872	93363	91731	90592	90599	90378		
1991	69112	91167	94452	93792	92186	91270	91187			
1992	76922	96890	100163	98974	98405	97721				
1993	86811	112247	114502	113842	112357					
1994	101257	123451	127532	125923						
1995	107582	134485	135931							
1996	113997	141203								
1997	108190									

Schedule P – New Jersey Manufacturers – Incurred claim – Private Motor

	N	N+1	N+2	N+3	N+4	N+5	N+6	N+7	N+8	N+9
1988	50395	66425	77609	82841	83400	82368	81691	81296	81592	81590
1989	56022	75703	90838	96847	96690	94344	93918	93233	93165	
1990	61079	85761	100131	105776	106002	104401	103106	102485		
1991	70857	97925	113696	123809	121641	119372	117638			
1992	84104	109443	126585	136514	136238	132453				
1993	95669	128682	150353	161064	156112					
1994	104360	134860	152751	161981						
1995	116839	153337	174393							
1996	136837	181052								
1997	152180									

Schedule P – Farmers Alliance – Incurred claim – Product Liability

	N	N+1	N+2	N+3	N+4	N+5	N+6	N+7	N+8	N+9
1988	199	246	360	404	401	321	328	340	332	342
1989	1312	1181	1347	1390	1511	1686	1860	1689	1680	
1990	493	700	817	740	912	903	897	902		
1991	391	487	1472	1640	611	593	597			
1992	586	741	1251	1509	1864	2040				
1993	892	1285	1556	1720	1828					
1994	654	1644	2060	2270						
1995	379	906	1255							
1996	705	950								
1997	384									

Schedule P – West Bend – Incurred claim – Product Liability

	N	N+1	N+2	N+3	N+4	N+5	N+6	N+7	N+8	N+9
1988	71	127	220	305	221	185	186	186	186	186
1989	185	318	360	377	425	382	389	258	261	
1990	124	225	348	394	218	216	243	261		
1991	299	344	531	1028	1266	1675	1959			
1992	204	399	431	647	625	709				
1993	208	459	566	714	766					
1994	452	665	910	805						
1995	243	532	1197							
1996	319	493								
1997	412									

Triangle from Mack 1993a

i	C_{i1}	C_{i2}	C_{i3}	C_{i4}	C_{i5}	C_{i6}	C_{i7}	C_{i8}	C_{i9}	C_{i10}
1	357'848	1'124'788	1'735'330	2'218'270	2'745'596	3'319'994	3'466'336	3'606'286	3'833'515	3'901'463
2	352'118	1'236'139	2'170'033	3'353'322	3'799'067	4'120'063	4'647'867	4'914'039	5'339'085	
3	290'507	1'292'306	2'218'525	3'235'179	3'985'995	4'132'918	4'628'910	4'909'315		
4	310'608	1'418'858	2'195'047	3'757'447	4'029'929	4'381'982	4'588'268			
5	443'160	1'136'350	2'128'333	2'897'821	3'402'672	3'873'311				
6	396'132	1'333'217	2'180'715	2'985'752	3'691'712					
7	440'832	1'288'463	2'419'861	3'483'130						
8	359'480	1'421'128	2'864'498							
9	376'686	1'363'294								
10	344'014									

Kurtosis and skewness estimation for non-life reserve risk distribution

SCOR Triangle – Investors' day 2011 – Casualty Proportional Worldwide

i	C_{i1}	C_{i2}	C_{i3}	C_{i4}	C_{i5}	C_{i6}	C_{i7}	C_{i8}	C_{i9}	C_{i10}	C_{i11}	C_{i12}	C_{i13}	C_{i14}	C_{i15}
1	1'235'909	28'084'788	36'277'376	40'475'255	43'219'640	44'663'902	45'326'054	48'497'970	49'233'469	49'550'622	49'690'990	49'831'357	50'360'419	49'973'713	49'862'952
2	1'377'332	35'709'204	43'213'075	51'753'287	56'669'470	57'444'929	59'800'911	61'566'321	62'121'477	62'876'588	62'731'073	63'128'635	63'478'522	63'123'300	-
3	2'227'264	38'477'741	50'151'725	53'966'008	54'671'130	57'692'644	60'487'201	61'814'232	61'750'678	62'642'901	62'622'432	61'843'626	62'208'577	-	-
4	6'063'523	77'188'565	106'217'291	113'964'643	120'412'934	126'472'874	124'422'859	128'806'243	127'982'137	125'512'850	134'339'552	135'190'968	-	-	-
5	17'046'209	118'335'598	149'317'461	158'231'476	169'819'782	171'982'468	171'861'629	180'041'898	178'223'074	178'218'074	180'474'109	-	-	-	-
6	19'133'946	129'368'180	158'428'296	161'509'739	164'007'156	161'788'099	163'095'804	160'465'113	165'298'176	166'467'315	-	-	-	-	-
7	15'294'531	146'782'533	181'032'613	182'220'943	171'358'201	171'997'288	170'664'146	171'561'713	169'645'162	-	-	-	-	-	-
8	18'460'498	117'030'644	146'526'379	132'681'719	130'134'713	124'721'542	133'009'259	136'533'267	-	-	-	-	-	-	-
9	23'542'477	113'067'832	125'103'424	143'385'679	143'172'588	130'016'839	125'328'953	-	-	-	-	-	-	-	-
10	17'341'979	122'713'595	142'017'246	154'409'430	144'452'872	134'267'844	-	-	-	-	-	-	-	-	-
11	19'850'416	134'682'628	163'649'787	168'464'316	170'525'178	-	-	-	-	-	-	-	-	-	-
12	16'449'001	125'616'958	155'143'936	158'477'086	-	-	-	-	-	-	-	-	-	-	-
13	27'153'648	144'171'990	164'553'451	-	-	-	-	-	-	-	-	-	-	-	-
14	30'745'977	141'700'503	-	-	-	-	-	-	-	-	-	-	-	-	-
15	14'911'754	-	-	-	-	-	-	-	-	-	-	-	-	-	-

SCOR Triangle – Investors' day 2011 – Motor Non Proportional Worldwide

i	C_{i1}	C_{i2}	C_{i3}	C_{i4}	C_{i5}	C_{i6}	C_{i7}	C_{i8}	C_{i9}	C_{i10}	C_{i11}	C_{i12}	C_{i13}	C_{i14}	C_{i15}
1	10'423'937	38'199'157	56'447'001	65'220'946	69'971'047	75'438'044	82'032'154	88'917'308	90'783'545	95'891'585	96'943'325	99'646'098	104'075'520	105'688'533	107'081'042
2	19'822'329	62'295'888	67'897'838	78'668'550	87'437'467	101'244'678	105'098'260	106'590'001	113'349'293	115'150'660	117'761'459	119'647'411	125'137'281	128'007'498	-
3	16'289'990	85'826'755	104'560'866	121'691'315	142'404'493	148'725'556	153'628'342	160'474'637	165'538'675	169'936'112	173'201'144	176'369'649	177'284'240	-	-
4	17'124'281	57'203'907	88'346'499	109'965'887	118'628'540	126'194'567	135'670'962	137'597'240	139'736'825	139'880'019	142'610'492	145'133'035	-	-	-
5	20'515'740	74'875'340	112'483'709	132'861'474	155'637'649	169'622'852	178'165'835	184'091'134	191'246'803	195'323'123	203'866'747	-	-	-	-
6	33'482'519	92'014'269	129'305'456	150'572'754	163'499'855	165'457'892	172'624'920	174'772'280	182'143'897	184'582'942	-	-	-	-	-
7	27'392'045	96'643'323	123'004'812	144'947'690	145'787'570	148'529'812	151'376'613	153'343'145	156'330'256	-	-	-	-	-	-
8	19'607'049	61'499'050	81'282'483	88'308'523	91'759'024	92'962'986	92'866'953	92'770'919	-	-	-	-	-	-	-
9	14'194'503	36'060'482	42'749'759	48'177'878	56'986'475	61'193'539	61'669'910	-	-	-	-	-	-	-	-
10	19'089'137	53'803'738	66'496'545	76'630'418	87'210'128	93'216'641	-	-	-	-	-	-	-	-	-
11	22'663'556	54'418'824	73'046'329	82'194'820	89'086'080	-	-	-	-	-	-	-	-	-	-
12	24'538'834	71'839'634	88'048'826	99'525'306	-	-	-	-	-	-	-	-	-	-	-
13	19'272'742	54'901'732	68'136'199	-	-	-	-	-	-	-	-	-	-	-	-
14	14'389'656	45'362'116	-	-	-	-	-	-	-	-	-	-	-	-	-
15	17'306'007	-	-	-	-	-	-	-	-	-	-	-	-	-	-

The Impact of Different Forms of Decision-Aids on User Best Assessments

Marc-André Desrosiers, FCAS, MBA

Abstract: In a world where information can be gathered, analyzed, interpreted and diffused much faster than for prior generations, we inquire about optimal schemes for the presentation of predictive models. We asked subjects to make predictions of quarterback ratings by presenting them with different information sets, some information sets including a predictive model prediction (with some different ways, some direct, some indirect, of presenting this information). We investigate (1) actual and optimal model usage by model users, (2) preferences of model users over ways to present information to them, and (3) inter-personal consistency of predictions. We find that (i) subjects are over-confident in their predictions, (ii) that this over-confidence is reduced when the subjects are presented with the predictive model, (iii) that subjects prefer an indirect presentation of the predictive model, where the model is presented as a deviation from a base statistic that is perceived to be relevant and credible, (iv) that their preferences are aligned with their own informal predictive model, and (v) that inter-personal consistency of predictions is fostered by the indirect presentation of the model (as a proposed deviation from a base statistic that is perceived to be credible and relevant).

Keywords: Clinical *versus* actuarial controversy, clinical synthesis, behavioral economics, information processing, business engineering, over-confidence, anchoring effects.

1. INTRODUCTION

When peering through a book like *Macrowikinomics* (Tapscott and Williams 2012), it becomes apparent that the ease with which information can be accumulated, processed, interpreted and used has greatly increased with the advent of new information technologies. While the authors may not focus on this, we believe that the exact nature of information processing by individuals, acting on their own or as part of a group, needs to be understood to be able to guide or better use the new possibilities arising out of increased access to (processed) information.

From a business perspective, in many fields, predictive models have long been distributed to users. For example, for insurance pricing purposes, rating manuals have been extent at least since the beginning of the 20th century. With the advent of the powerful central computers and databases, businesses could begin to gather massive amounts of information from their activities, process this information with statistical technology that was becoming more and more powerful, and harness the results in tools that could be deployed to field operatives. That trend has continued to the point that, with today's technology, it is relatively easy to present field operatives with "informational dashboards" to assist them in their decision-making.

However, more information is not always better information and, even if all the pieces of available information were always relevant and pertinent, in the end, human beings still need to comprehend it, process it, interpret it and, ultimately, take decisions in reaction to it. Therefore, it is

worthwhile to understand the influence of the presentation of information has on human decision-making.

To investigate this issue, we decided to ask subjects to make predictions with access to different information sets: we have asked American undergraduate business students to make predictions of quarterback ratings. As such, we have decided to focus on the cognitive aspect of the influence of information sets; more specifically, the cognitive influence (in human decision-making) of the different ways a predictive model can be communicated.

Because of known behavioral effects such as over-confidence and anchoring, issues of actual and optimal model usage are especially interesting. How much do the subjects use the model when it is presented to them? What variables influence that usage? Can the subjects bring information that can help them beat the model? How much should the subjects be using the model? What variables influence how much the subjects should be using the model? Are the subjects bringing in this information in an optimal way? Moreover, is the across subjects consistency of predictions affected by the way the information is presented, keeping in mind that consistency places an upper bound on reliability of predictions?

Also, because information technology is not evolving in a vacuum, it is also worthwhile to ask about the preferences of the subjects over different ways information is presented to them. Do the subjects perceive differently information sets, in terms of the credibility and the relevance of the information they are provided with? How does this perception match up with the informal models subjects can use to make predictions? Does this preference line up with their own confidence in their predictions?

We find the following.

1. The subjects display massive over-confidence in their own predictive abilities, as optimal model usage should be about 70% (for the population of subjects as a whole) while actual model usage is bounded by 30% (again, for the population of subjects as a whole) in our experimental design, where the subjects that were attracted to the experiment tended to generally perceive themselves to be knowledgeable about the subject-matter.

2. The subjects can bring valuable information to the predictive models, as predictive accuracy could be increased by adding human inputs. However, the subjects do not bring in this information optimally as their actual predictive accuracy when they are presented with the model is less than optimal.

3. Presenting the subjects with the predictive model as the only piece of extra information most favors model usage: inducing about 30% model usage. Presenting the model indirectly, as a proposed deviation from a selected contextually relevant historical mean reduces model usage to about 15%. We interpret this extra use of the model when it is presented alone to be the result of an anchoring effect.

4. The only variable that significantly affects actual model usage is the Judging-Perceiving MBTI personality dimension where Judging individuals actually use the model more. The only variable that affects optimal model usage is the self-reported familiarity with football (that is, expertise in the field of interest): subjects more knowledgeable about the subject-matter need the model less.

5. Across subjects prediction consistency is fostered by presenting the predictive model indirectly as a proposed deviation from a base statistic that is perceived to be relevant and credible.

6. The subjects perceive more favorably base statistics that match better with their own informal predictive models.

7. When they are presented with the predictive model, the subjects prefer it to be presented indirectly, as a proposed deviation from a favorably perceived base statistic.

8. The perception of credibility and relevance of the provided information lines up well with the self-reported confidence the subjects have in their own predictions.

1.1. Research Context

The present research can be conceived to exist at the confluence of three major research programs: (1) the actuarial *versus* clinical controversy, (2) the research on information processing arising out of behavioral economics, and (3) the business engineering applied research stream. We will discuss each in turn.

1.1.1. Clinical *versus* actuarial controversy

In psychology and medicine, the clinical *versus* actuarial controversy relates to the extent to which a clinician can do and actually does better (or worse) than an available model based on predictive modeling. In "Man versus models of man: A rationale, plus some evidence, for a method of improving on clinical inferences" (Goldberg 1970), when examining how clinical psychologists, physicians, and other professionals that are typically called on to combine cues to arrive at some diagnostic or prognostic decision, it was found that, for the diagnostic task, models of the human decision-making were generally more valid than the humans themselves, even when models were developed on a relatively small set of cases and then humans and model competed on a completely new set. Such was the case because mathematical representations of such clinical judges can often be constructed to capture critical elements of their judgment strategies. Keeping in mind that judgment consistency sets an upper bound on judgment reliability, models of human decision-making, that are free from inconsistencies, outperforming actual human decision-making is indicative that judgment consistency is empirically more important than the ability of humans to incorporate information in more complex ways or using qualitative information that cannot be incorporated in the model of human decision-making. In "Effect of input from a mechanical model on clinical judgment" (Peterson and Pitz 1986), when exploring clinical synthesis[1], it was found that performance of subjects improved when the model was provided, but subjects still did less well than the model.

Peterson and Pitz further researched the topic. They found that it is worthwhile to make a distinction between the belief that a prediction is correct (*i.e.* confidence) and the ability of a subject making a prediction to imagine scenarios in which a prediction is not realized (*i.e.* uncertainty). They find that confidence increases as more information is provided to the subjects but that it was decreased when the difficulty of the task was increased. On the other hand, uncertainty increased as subjects were provided with more information. (Peterson and Pitz, Confidence, Uncertainty, and the Use of Information 1988, 85) They also found tendencies of over-confidence, with over-confidence a decreasing function of quantity of information provided. (Peterson and Pitz, Effects of Amount of Information on Predictions of Uncertain Quantities 1986, 229)

---

[1] Clinical synthesis is giving the decision makers predictions from a model as input but let him or her make the final judgment.

Finally, in "A comparison of the predictive accuracy of loan officers and their linear-additive models" (Zimmer 1981), a clinical *versus* actuarial-type study was conducted on loan officers and found materially the same effects as when an health care practitioner population is examined.

An important insight that emerges from the actuarial *versus* clinical literature is that, because of the "broken leg" problem, where a user of the model may have a material insight into the problem at hand that cannot be easily or at all integrated into the model prediction, it is often sensible, for business purposes, to allow users of decision-aids to make the final selection.

1.1.2. Behavioral economics: over-confidence and anchoring

In economics, following the work of Simon (A Behavioral Model of Rational Choice 1955), significant efforts have been dedicated to refining theoretical models of preferences and human information processing. A seminal work in this area was "Judgment under Uncertainty: Heuristics and Biases" (Tversky and Kahnemen 1974) where the authors explore three families of heuristics that tend to induce individuals to misestimate probabilities:

(1) the representativeness heuristic, exemplified by the following biases:
    (a) insensitivity to prior probability of outcome,
    (b) insensitivity to sample size,
    (c) misconception of chance, *e.g.*, gambler's fallacy,
    (d) insensitivity to predictability,
    (e) illusion of validity, and
    (f) misconception of regression [towards the mean];
(2) the availability heuristic, exemplified by the following biases:
    (a) biases due to the retrievability of instances,
    (b) biases of imaginability, and
    (c) illusory correlation; and
(3) the adjustment and anchoring heuristic, exemplified by the following biases:
    (a) insufficient adjustment,
    (b) biases in the evaluation of conjunctive and disjunctive events, and
    (c) anchoring in the assessment of subjective probability distribution.

Thus, two factors that can influence actual and optimal model usage are (1) subject over-confidence due to subjects using a heuristic scheme to evaluate the relevant probabilities where the heuristic scheme leads to estimated probabilities that can be quite different from those that would be obtained under an optimal scheme (like Bayes' theorem) and (2) anchoring which is particularly relevant for our purposes since it predicts that we can influence the selections made by the mere presentation of a numerical piece of information to the subjects. In "Framing, Probability Distortions, and Insurance Decisions" (Johnson, et al. 1993), potential business consequences of these information processing effects were explored. Among the findings of the study, it was found that sophisticated subjects can be subject to the same imperfect information processing bias as subjects in the general population are subject to.

1.1.3. Business engineering

Finally, there is an array of applied business research that examines how (imperfect) information processing effects influence business outcomes and whether or not business practices can be adapted to take into account these behavioral effects to help businesses attain their objectives. An example of applied research that examines reliance on statistical models is "Decision Aid Reliance: A Longitudinal Field Study Involving Professional Buy-Side Financial Analysts" (Hunton, Arnold and Reck 2010): this study examines the discretionary decision-aid reliance behavior of professional buy-side financial analysts. The researchers find that "the most interesting finding from theoretical and practice perspectives is that increased task ability, as determined by objective historical task performance, was associated with increased reliance on the DA." (Hunton, Arnold and Reck 2010, 1019) Another example of applied business research is "Decision making in an organizational setting: Cognitive and organizational influences on risk assessment in commercial lending" (McNamara and Bromiley 1997), a field study of decision-making, where it was found that "organizational effects appear to dominate cognitive ones" (McNamara and Bromiley 1997, 1083). Finally, in "Bridging the marketing theory-practice gap with marketing engineering" (Lilien, et al. 2002), the authors "provide several illustrations of the successful application of the marketing engineering concept" (Lilien, et al. 2002, 111), that are based on marketing management support systems that enrich decision-making.

This literature focuses on the application of pure knowledge about human behavior to solve business problems/issues. This involves applied research, but note also that other pure research avenues are often opened. As noted above, common applications relate to banking, insurance, finance, marketing, accounting, *etc.* What is generally found is that there is a gap between pure knowledge and the necessary optimal design that requires field testing. As such, this often brings up the issue of external validity that allows one to bridge from results obtained in one context to another, presumably similar, context.

1.2. Outline

The remaining will go as follows. Section 2 will be dedicated to the background and methods. Section 2.1 will contain a description of the experiment. Section 2.2 will relate specifically to the construction of the decision-aid used in the experiment. Section 3 will present and discuss the results. In section 3.1, we will present a model-of-man that takes into account the treatment the subjects received. In section 3.2, actual model compliance will be examined; in section 3.2.1, the effect of presenting a proposed deviation will be presented; in section 3.2.2, the actual (implicit) deviation will be explored as a function of the proposed (implicit) deviation; in section 3.2.3, drivers of actual model compliance will be sought; in section 3.3, optimal model compliance will be explored; in section 3.3.1, drivers of optimal model compliance will be sought; in section 3.4, the

reported perceptions of credibility and relevance of the presented information will be explored; in section 3.5, the self-reported confidence of subjects regarding their prediction will be explored; in section 3.6, the interpersonal agreement of subject predictions will be presented; in section 3.7, the net compensation outcomes will be presented and discussed; and, in section 3.8, external validity considerations will be discussed. Section 4 will conclude.

2. BACKGROUND AND METHODS

In this section, we will describe the exact nature of the experiment and the associated predictive model-building activities that needed to take place before the experiment could go live.

2.1. Experiment Description

Targeting undergraduate business students from University of Wisconsin-Madison, an internet survey asking subjects to <u>make predictions about the quarterback rating results for the coming week of National Football League activity</u>[2] was distributed for three weeks in a row (NFL weeks 14, 15 and 16 of the 2012-2013 season). The target subjects were reached through in-class presentations of the experiment and targeted e-mail distribution lists. In particular, for week 16, subjects that had participated in weeks 14 and 15 were solicited to participate again.

Count	Repeats		
Week	No	Yes	Total
14	82		82
15	60	12	72
16	44	51	95
Total	186	63	249

Table 1: Number of subjects by week

2.1.1. Task description

The subjects were asked to make predictions about the quarterback ratings for the expected starters for the Sunday and Monday night games of the upcoming week of NFL games. The subject were then presented with some high level and general information concerning quarterback ratings: they were referred to the appropriate Wikipedia entry describing quarterback/passer rating and they were told about the possible values for the statistic, what a quarterback needed to accomplish to obtain a perfect score, the approximate league-wise average for the score, about an intra-week measure of league-wise quarterback rating dispersion, and about an across-weeks measure of intra-individual measure of quarterback rating dispersion.

---

[2] Note that no predictions were asked for the Thursday night game: only predictions relating to the Sunday and Monday night games were asked.

2.1.2. Compensation description

They were then told what the compensation scheme would be. Subjects could gain from 0 to 15 USD, or up to 50¢ per prediction. They were told that their total compensation would be a sum of their by quarterback prediction compensations. Their by quarterback prediction compensations were computed using the following function.
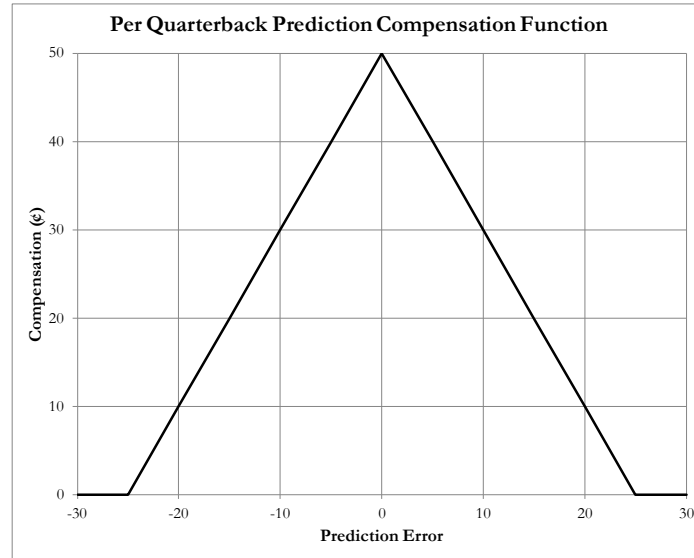
Graph 1: Per Quarterback Prediction Compensation Function

The compensation function was selected to incentivize subjects to get every prediction as right as possible. We believe that subject risk aversion, cautiousness, financial cautiousness, *etc.* would not play a material role in affecting subject predictions: we will come back to this when examining drivers of model compliance.

2.1.3. Treatments descriptions

Table 2 provides a high level description of the treatments. Table 3 presents for which weeks each treatment was ran. Table 4 presents the verbose that was presented to the subjects under each treatment. Note that for each week, the 30 predictions that needed to be made were separated in two batches of 15 predictions, which were the same 15 predictions under all the treatments of the week. The subjects randomly received a treatment for each batch, independently of the other batch of prediction as well as independently of other subjects.

After each prediction batch, the subjects were also asked to reflect back on the predictions they just made. They were asked, in order, (1) how confident they felt about their predictions, (2) how confident they felt about the credibility and relevance of the information they were provided with for their predictions, and (3) to voluntarily discuss, in a free-form field, the strategies that supported their predictions.

Provided Infomation	A·	B·	Notes
·1	Overall Average	Individual Average	(only differ in nature of the average provided)
·2	Predictive Model Only	Predictive Model Only	(same)
·3	Overall Average plus Predictive Model Presented as a Proposed Deviation	Individual Average plus Predictive Model Presented as a Proposed Deviation	(see line ·1)

Table 2: Descriptions of treatments
The possible treatments are A1, A2/B2, A3, B1, and B3.

Possible Treatments	Week 14	Week 15	Week 16
A1	YES	YES	
A2/B2	YES	YES	YES
A3	YES	YES	
B1			YES
B3	YES	YES	YES

Table 3: Time of the treatments

Note that the subjects were not described the predictive model beyond the details provided in Table 4. This was voluntary on our part as we wanted to make sure to reproduce a setting that is common in the deployment of predictive models: the predictive model is constructed using potentially complex statistical methodologies that are generally not described in details to the users of the predictive models. It is not uncommon that the users are only (1) told that the predictive model is the 'best' available model and (2) shown what variables enter the predictive model.

2.1.4. Subject population description

Demographic information was asked to the subjects in two batches: a batch of questions was asked to the subjects prior to their making their predictions and a second batch of questions was asked after the subjects completed their predictions.

The first batch of questions can be divided in three parts: (1) general demographic questions, (2) football trivia, and (3) self-assessed familiarity with football and fantasy sports. Under the heading of general demographics, questions were asked about: gender, age, level of study, study area, Grade Point Average, and mathematical abilities.

Common	For each of the listed quarterback below, please enter your forecast for their quarterback rating in the coming game. You will find in brackets, first, their own team, second, the team they are playing against.
A1	**You will also find the overall average of the quarterback rating for the starting quarterbacks (OAAvg).** Player (ForTeam, @AgainstTeam) **OAAvg**: aa.a
A2/B2	**You will also find the prediction from a statistical model (Stat. Model). The statistical model takes into account the following factors:**

the overall quarterback rating average for the starters	the individual's prior week quarterback rating result
the individual's season-to-date quarterback rating	the individual's two week prior quarterback rating result
the average season-to-date allowed quarterback rating for the opposing team	the own team win-loss record
the consistency of the quarterback rating for the individual	the opposing team win-loss record

Player (ForTeam, @AgainstTeam) **Stat. Model**: bb.b

A3	You will also find: 1. **the overall average of the quarterback rating for the starting quarterbacks (OAAvg), and** 2. **the prediction from a statistical model (Stat. Model Dev.) expressed as a deviation from the overall average. That is, Stat. Model = OAAvg + Stat. The statistical model takes into account the following factors:**

the overall quarterback rating average for the starters	the individual's prior week quarterback rating result
the individual's season-to-date quarterback rating	the individual's two week prior quarterback rating result
the average season-to-date allowed quarterback rating for the opposing team	the own team win-loss record
the consistency of the quarterback rating for the individual	the opposing team win-loss record

Player (ForTeam, @AgainstTeam) **OAAvg, Stat. Model Dev.**: aa.a, +/-cc.c

B1	**You will also find the individual average of the quarterback rating for the starting quarterbacks (IndAvg).** Player (ForTeam, @AgainstTeam) **IndAvg**: dd.d
B3	**You will also find:** 1. **the season-to-date quarterback rating for the individual (IndAvg), and** 2. **the prediction from a statistical model (Stat. Model Dev.) expressed as a deviation from the season-to-date quarterback rating for the individual The statistical model takes into account the following factors:**

the overall quarterback rating average for the starters	the individual's prior week quarterback rating result
the individual's season-to-date quarterback rating	the individual's two week prior quarterback rating result
the average season-to-date allowed quarterback rating for the opposing team	the own team win-loss record
the consistency of the quarterback rating for the individual	the opposing team win-loss record

Player (ForTeam, @AgainstTeam) **IndAvg, Stat. Model Dev.**: dd.d, +/-ee.e

Table 4: Treatment description
In the table above, one can find the way the information was presented to subjects for each of the treatments described in Table 2.

The football trivia section was originally designed as an entertainment section: that is, it was intended to be fun for the subjects. The questions asked were:

1. "Which team won the last Super Bowl (played in February 2012)?",
2. "Who is a quarterback for the Green Bay Packers?",
3. "Which of these players is a defensive end who, in the 2011-2012 season, was a member of the Super Bowl winning team, went to the Pro Bowl, and lead his team for the number of sacks in the season?", and,
4. "Which of these players has posted the most games with a perfect passer rating?".

Given that UW-Madison is in Wisconsin, it was expected that most subjects would accurately identify Aaron Rodgers as the quarterback for the Green Bay Packers and this expectation was met. Given the publicity and level of public attention surrounding the Super Bowl, it was expected that most subjects would have known that the New York Giants had won the last played Supe Bowl at the time of the survey. More subjects picked the wrong answer, but no subject picked a team that had not played in the Super Bowl game. Questions 3 and 4 were of a higher level of difficulty for the subjects and the accuracy of their answers followed accordingly.

While the intent behind asking questions about football was to make the survey more fun for the subjects, it may have reduced the participation of subjects whose level of familiarity with football may have been average or less than average. This provided us with a subject population biased towards 'football experts' (as contrasted with the general American population), but it also potentially had the side-effect of making the subject population more homogeneous.

Under the heading of self-assessed familiarity with football and fantasy sports, questions were asked about:

- the self-reported familiarity of the subjects with fantasy sports,
- the self-reported familiarity of the subjects with football,
- the average number of days per week that the subjects watch or read sports news, and,
- whether the subjects has a fantasy football team and, if so, whether the team was doing well or not.

The second batch of questions can be divided in two parts: (1) a simplified Myers-Briggs personality assessment, and (2) other demographic questions. The Myers-Briggs questions covered the following four personality dimensions (in parenthesis, the approximate proportion of the population that reported the dimension):

- the Judging (65-70%) *versus* Perceiving dimension,
- the Thinking (65-70%) *versus* Feeling dimension,
- the Sensing (55-60%) *versus* Intuiting dimension, and,
- the Introversion (60-65%) *versus* Extroversion dimension.

The other questions covered cautiousness, financial cautiousness, locus of control, and other questions about the statistics and sports statistics.

With all this demographic information available, it was easy for us to verify that the randomizing scheme integrated with our web-survey was adequate, as the generated sample appeared (materially and statistically) well-balanced on all dimensions.

2.2. Preparation of the Predictive Model of Quarterback Rating

The process to prepare predictive model quarterback rating predictions for the upcoming week of NFL activity went as follows. First, using a fantasy football website[3], the expected starting quarterbacks were retrieved. Because of ease of access of information and to ensure consistency of information processing with sources of information that the subjects could access on their own, some pieces of information were also retrieved from the said website: the individual season-to-date quarterback rating and detailed fantasy football experts' predictions. The expert predictions were retrieved but ultimately not used in predictive modeling, because we wanted to ensure that the subjects could bring forward valuable qualitative information not reflected in the model, and not just information that would have arisen between the time the experts entered their predictions and the time when the subjects made their predictions. Still, the detailed expert predictions were translated into a predicted quarterback rating by first averaging the details of the predictions across the (three) experts and then converting to a predicted quarterback rating using the appropriate formula[4]. Second, using football statistics websites[5], the remaining necessary information was retrieved:

- the overall quarterback rating average for the starters,
- the individual's season-to-date quarterback rating,
- the average season-to-date allowed quarterback rating for the opposing team,
- the individual's prior week quarterback rating result,
- the individual's two week prior quarterback rating result,
- the consistency of the quarterback rating for the individual,

---

[3] http://fantasynews.cbssports.com/fantasyfootball/playerindex/POS_QB.

[4] Note that the order of averaging and applying the quarterback rating formula does not affect materially the predicted quarterback rating.

[5] http://sports.yahoo.com/nfl/stats/ and www.nfl.com/.

- the own team win-loss record, and,
- the opposing team win-loss record.

Each week, using the latest available information, the predictive model was calibrated by maximizing the compensation function across all (known) weeks (played at that time), using a weighting scheme that put more weight on the more recent weeks.[6]

It is worthwhile to note that, from a predictive modeling point of view, both the overall quarterback rating average for the starters and the individual's season-to-date quarterback rating are basically equally powerful one-variable predictors: they both would have yielded an average compensation of 16.1¢ per prediction had they been used for all predictions from weeks 6 to 17 (358 predictions). This is greatly due to the significant regression-to-the-mean that happens both at the league-wise level and at the intra-individual level. Note also that the predictive model would have generated a higher average compensation (on retrodicted quarterback ratings) than either (1) using only the overall mean, (2) using only the individual mean, (3) using only the predictions from the panel of experts on football, or (4) using a weighted average of the overall mean and the predictions of the panel of experts. The following table details these elements.

---

[6] Please note that there was a significant revision to the predictive model between week 15 and 16: thus, the coefficients of the predictive model changed materially between the two weeks.

Week	14				15				16			
Including Panel of Experts Predictions?	No		Yes		No		Yes		No		Yes	
	Estimate	St. Error	Estimate	St. Error	Estimate	St. Error	Estimate	St. Error	Estimate	St. Error	Estimate	St. Error
Starters Season-to-date	-0.42	0.43	0.19	0.47	-0.17	0.43	0.28	0.42	0.91	0.43	0.98	0.31
Individual Season-to-date	-0.19	0.30	-0.63	0.27	-0.21	0.26	-0.62	0.34	-0.75	0.34	0.46	0.39
Allowed QB Rating Opposing Team	1.11	0.26	0.16	0.35	0.85	0.38	0.01	0.44	0.58	0.45	-0.97	0.39
Prior Week Result	0.07	0.12	-0.07	0.13	0.09	0.10	-0.05	0.10	-0.10	0.14	-0.03	0.12
Prior Week Missing Indicator	6.97	5.72	-0.20	2.55	7.61	3.79	0.99	4.17	-1.26	6.80	0.26	4.79
Two Prior Week Result	0.13	0.07	0.09	0.07	0.17	0.11	0.11	0.10	0.22	0.11	-0.08	0.12
Two Prior Week Missing Indicator	12.01	6.41	1.02	3.54	14.80	6.73	4.04	3.35	8.52	5.28	1.26	4.38
Own Team Record	4.18	4.61	1.87	2.59	3.91	6.18	0.80	5.14	-13.18	3.49	0.95	3.33
Opposing Team Record	29.10	5.93	6.86	5.53	28.46	5.64	8.32	5.48	3.25	5.07	0.33	4.41
Panel of Experts Predictions			1.11	0.24			1.10	0.27			0.57	0.28
Random-Effects-Like Parameter	8.17	5.63	0.00	0.06	7.06	6.11	0.00	3.08	11.43	5.89	0.00	0.14
Average Compensation on Retrodictions												
Across	8	Weeks	8	Weeks	9	Weeks	9	Weeks	10	Weeks	10	Weeks
Average Prediction of Predictive Model	0.176	0.014	0.183	0.020	0.181	0.015	0.188	0.015	0.177	0.014	0.185	0.013
Weighted Average of Panel Predictions and Overall Average	0.167	0.016			0.170	0.012			0.171	0.011		
Starters Season-to-date	0.162	0.018			0.166	0.014			0.163	0.014		
Panel of Experts Predictions	0.156	0.015			0.159	0.010			0.162	0.010		

Table 5: Coefficients of predictive models for weeks 14, 15 and 16, with and without predictions of panel of experts as an input.

The above table contains the coefficients used in the predictive models for weeks 14, 15 and 16 with associated standard errors estimated using block bootstrap (that is, a bootstrap procedure where complete weeks are re-sampled). Note that the coefficients of the predictive models are generally not precisely estimated. Note that there was a significant model change between weeks 15 and 16. Nonetheless, the predictive model fares better on the average compensation (over many weeks) for retrodicted quarterback ratings than just using the predictions of the panel of experts, just using the average season-to-date for all starters, or using a weighted average of the predictions of the panel of experts and the overall average of the starting quarterbacks. Note also that, when the predictions of the panel of experts are included in the model, the average compensation of optimized retrodictions is higher than when that information is not included in the predictive models. We interpret this as evidence that human input can be significant in improving predictive performance.

In Table 5 above, note that when the predictions of the panel of experts is added to the predictive model, the average compensation (optimized) for the retrodicted quarterback ratings can be improved from what was attainable without the human input. We interpret this as evidence that human inputs can materially improve predictive performance. Furthermore, this improvement in

predictive performance seems to be substantially related to the random-effects-like[7] parameter becoming unnecessary when the prediction of the panel of experts is added as a predictive variable.

3. RESULTS AND DISCUSSION

We can now turn our attention to the analysis of the results. First, we will quickly describe minor data manipulations that were done in order to ensure the interpretability of the data. Records where the quarterbacks the panel of experts predicted would play (that is, the quarterbacks for whom model predictions were built) but did not actually play were removed from the analysis and treated as 'not available'. Cases were the subjects were effectively predicting that the quarterbacks for whom a prediction was sought would not actually play were also removed and treated as 'not available': a prediction threshold of 12.5 was used to accomplish this, as this constituted a natural cutoff point in the data.

---

[7] Random-effects are meant to reflect fundamentally the same effects as those reflected in greatest accuracy credibility theory. Examples of predictive applications of random-effects-like predictive modeling applications can be found in (Fundamentals of Individual Risk Rating, Part I 1992). Random-effects-like formulas are meant to allow the predictive model to reflect individual differences, but only to the extent they are credible (or predictive).

3.1 Model-of-Man

Our first step in the analysis of the data was the construction of predictive models of subject predictions (that is, models-of-man).

	Predictive Model Coefficients			Model-of-Man (without Model)			Model-of-Man (with Model)			Model-of-Man (with Model and Cognitive Dissonance)		
	Week 14	Week 15	Week 16	Estimate	Std. Error		Estimate	Std. Error		Estimate	Std. Error	
(Intercept)	-36.56	-14.22	79.08	2.70	4.89		15.14	3.37	***	12.99	3.12	***
Individual Season-to-date	-0.19	-0.21	-0.75	0.53	0.04	***	0.52	0.06	***	0.42	0.06	***
Model							0.31	0.07	***	0.38	0.06	***
Own Team Record	4.18	3.91	-13.18	14.98	2.39	***	16.46	2.38	***	17.25	2.24	***
Opposing Team Record	29.10	28.46	3.25	-9.67	2.14	***	-13.42	2.05	***	-12.31	1.92	***
Individual Season-to-date:A2/B2							-0.25	0.07	***	-0.21	0.07	**
Individual Season-to-date:A3							-0.08	0.08		-0.07	0.08	
Individual Season-to-date:B1							-0.10	0.10		-0.02	0.10	
Individual Season-to-date:B3							-0.07	0.07		-0.03	0.07	
Model:A2/B2							0.24	0.08	**	0.22	0.07	**
Model:A3							0.08	0.09		0.07	0.08	
Model:B1							0.09	0.10		0.06	0.10	
Model:B3							0.06	0.08		0.05	0.07	
Allowed QB Rating Opposing Team	1.11	0.85	0.58	0.37	0.05	***						
Prior Week Result	0.07	0.09	-0.10	-0.01	0.02							
Prior Week Missing Indicator	6.97	7.61	-1.26	-20.83	4.85	***						
Two Prior Week Result	0.13	0.17	0.22	0.05	0.02	**						
Two Prior Week Missing Indicator	12.01	14.80	8.52	-11.87	2.57	***						
Individual Season-to-date:Large Cognitive Dissonance										0.09	0.02	***
Individual Season-to-date:A2/B2:Large Cognitive Dissonance										-0.05	0.03	.
Individual Season-to-date:A3:Large Cognitive Dissonance										-0.01	0.03	
Individual Season-to-date:B1:Large Cognitive Dissonance										-0.07	0.04	*
Individual Season-to-date:B3:Large Cognitive Dissonance										-0.04	0.03	.
Random-Effects-Like Parameter	8.17	7.06	11.43									
Average Weighted Compensation (Retrodiction)	0.176	0.181	0.177									

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 6: Model-of-man with the predictive models attached.

The table above contains both the predictive models that were used to provide the subjects with guidance in treatments A2/B2, A3 and B3 and models-of-man, both including and excluding the model as a factor. Note that, while the predictive models vary from week to week, the weighted average compensation they would have generated on past predictions remain approximately constant. The factor "Random-Effects-Like Parameter" is effectively a ballast term to allow the model to appropriately recognize the average residuals (of a model that does not take into account performance consistency) for a given individual. When the predictive model is incorporated in the model-of-man, only the individual mean and the model prediction seem to interact with treatment A2/B2. The coefficients suggest that, for treatment A2/B2, the subjects were using less of the individual mean and substituting it for the model. This is consistent with an anchoring effect. The model-of-man that does not incorporate the model is suggestive of why the subjects would substitute from the individual mean under A2/B2: the individual mean is the most important component of their internal model. This is consistent with an analysis of median deviation from the presented relevant mean for A1 and B1: the median deviation is about 10 for A1 (which presents only the overall mean) and only about 5 for B1 (which presents only the individual mean).

Also, note that when the subjects do not agree with the model (this is indicated by the 'Large Cognitive Dissonance' variable which is an indicator that the unaided prediction and the model prediction are more than their median distance apart from each other), they tend to rely more on the individual average, but the effect is attenuated when the subjects are explicitly presented with the model.

At a first glance, it is interesting to note that the subjects were effectively using a model that was dissimilar to the optimized predictive model. For example, the mental model of the subjects did not appear to incorporate well regression-to-the-mean effects, as evidenced by the negligible coefficient

attached to the intercept for the model-of-man that does not include the model as a predictive variable. This was the case even though the subjects were explicitly told that regression-to-the-mean was a significant feature of the data. Also, the subjects seemed to believe that the individual mean was a powerful predictor of future performance, as evidenced by both the 0.53 highly significant coefficient attached to the 'Individual Season-to-date' variable in the model-of-man without the model as a variable as well as the fact that the median absolute deviation (from the selected mean) from the treatment were the subjects only received the individual mean was about half of the median absolute deviation under the treatment received only the overall mean.

When the 'model' variable is added to the model-of-man, a clear effect of information substitution appears: when presented only with the model, the subjects materially and significantly substitute the model prediction for the individual mean, as evidenced by +0.24 highly significant coefficient and -0.25 highly significant coefficient attached to the 'Model:A2/B2 ' and 'Individual Season-to-date:A2/B2', respectively. Thus, we can speculate the following mental process could be occurring in the subjects. When they are presented with the model as a proposed deviation (that is, in treatments either A3 or B3), the subjects examine the information that the predictive model is providing them with and incorporate it in the best assessment predictions. The coefficient of 0.31 attached to the 'Model' variable in the model-of-man suggests that the subjects have a somewhat low valuation of the predictive model predictions. But, when they are presented only with the model as an extra piece of information (that is, in treatments A2/B2), then the subjects get anchored to the predictive model prediction and use the model more; that is to say, the data seems to suggest that the extra use of the predictive model by the subjects could well come from an anchoring effect where the mental model of the subjects gives extra weight to the model prediction because it is a (unique) number that has been floated to them just before they have to make their predictions. That the subjects substitute from the 'Individual Season-to-date' variable seems natural given its importance in their uninfluenced mental model.

Furthermore, we are led to wonder about the following. What happens to subject predictions when their own uninfluenced predictions would be 'far away' from the predictive model predictions; that is, what do subjects do when the predictive model appears to them as particularly less relevant? In the 'Model-of-Man (with Model and Cognitive Dissonance)', we added a 'Large Cognitive Dissonance' indicator variable: the indicator was set to 1 when the distance between the uninfluenced subject (average) prediction and the predictive model prediction was more than the median distance between the two predictions. From the coefficients of the 'Individual Season-to-date' crossed with the 'Large Cognitive Dissonance' crossed with the treatments variables, we see that the 'Individual Season-to-date' variable received a little bit more weight when the subjects agreed less with the predictive model. This leads us to suppose that the subjects gave increased weight to their own internal model when they agreed less with the predictive model prediction. This can be interpreted as a cognitive dissonance effect where the subjects ignore the model when they (particularly) do not agree with it.

3.2. Actual Model Compliance

We can now turn our attention towards quantifying model compliance under different treatments.

3.2.1. Effect of presenting a proposed deviation

From Graphs 2 and 3, we see that model compliance is greatest under treatments A2/B2 (at about 30% model compliance). The model compliance is reduced by about half when the subjects are presented with the model as a proposed deviation: there is about 15% model compliance under treatments A3 and B3. This is consistent with the effects found in the model-of-man analysis where (1) the subjects were found not to be utilizing the predictive model very much and (2) the subjects gave extra weight to the predictive model when that was all they were presented with.
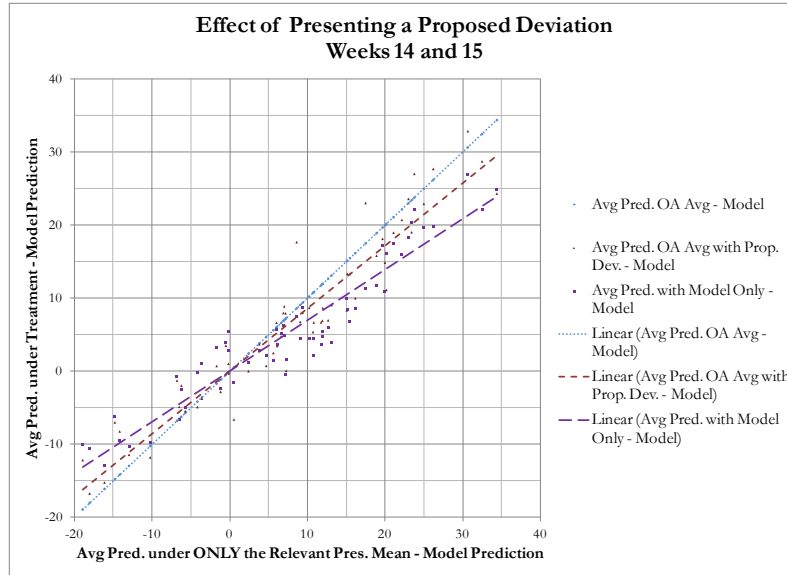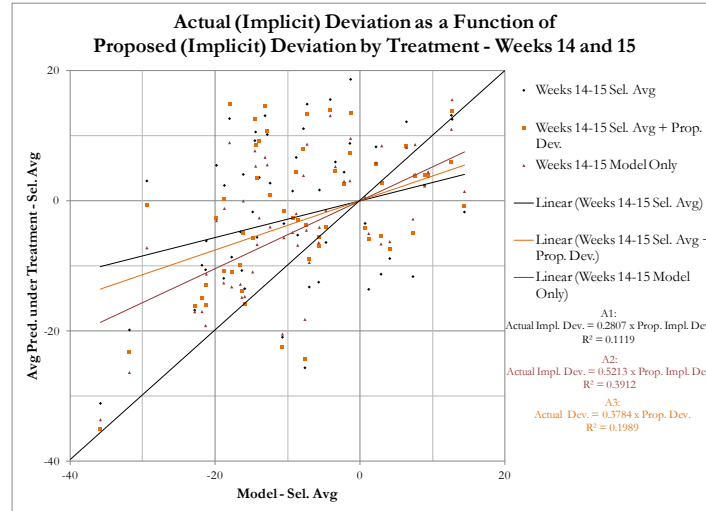
Graph 2: Model compliance, expressed as a relationship between the actual average prediction for a given treatment against the average prediction when only the overall mean is presented, both axes normalized by the model prediction, under treatments A1, A2 and A3 for weeks 14 and 15.

A slope of 1 indicates no model usage: which necessarily happens for A1. A slope of 0 would indicate complete model compliance. The line corresponding to treatment A2 has a slope of approximately 0.7: thus, implying a model usage of about 30%. The line corresponding to treatment A3 has a slope of about 0.85: thus, implying a model compliance of about 15%. If the fitted line for treatment B3 had been added to the graph above, it would materially and statistically overlap the fitted line for treatment A3.
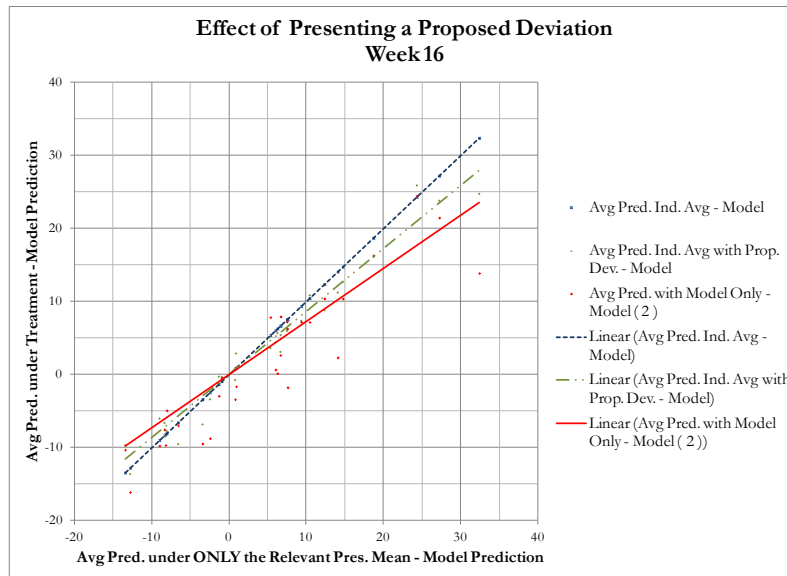
Graph 3: Model compliance, expressed as a relationship between the actual average prediction for a given treatment against the average prediction when only the overall mean is presented, both axes normalized by the model prediction, under treatments B1, B2 and B3 for week 16.

Similarly to Graph 2, the line corresponding to treatment B2 has a slope of approximately 0.7: thus, implying a model usage of about 30%; and, the line corresponding to treatment B3 has a slope of about 0.85: thus, implying a model compliance of about 15%.

3.2.2. Actual (implicit) deviation as a function of the proposed (implicit) deviation

Graphs 4 and 5 present the actual deviation as a function of the proposed deviation. Note that only under treatments A3 and B3 is the model presented explicitly as a proposed deviation. For models A1, B1 and A2/B2, this proposed deviation is implicit as the subjects either were not informed of the predictive model prediction (A1//B1) or were not presented the predictive model predictions in this way (A2/B2). A slope of 1 with a R^2 of 1 in Graphs 4 and 5 would indicate complete model compliance. As was found in the previous section, the subjects tended to use more of the proposed (implicit) deviation under treatments A2/B2, less so under models A3 and B3, and even less so under treatments A1 and B1. What is also apparent in Graphs 4 and 5 is that the proportion of the subjects actual (implicit) deviation that can be explained by the proposed (implicit) deviation increases from treatments A1//B1 to A3//B3 to A2/B2. This is consistent with our working hypothesis that the model 'takes up more mental place' when the subjects are only presented the predictive model as supplementary information.

It is interesting to note that one may measure the causal impact on subject predictions of changing the provided information from that of an initial information set to another information set by using the experimental methodology laid out above. Of particular interest to the experimenter was the causal impact of changing the base from which the model is presented as a proposed deviation; that is, the causal impact of changing the information set from that of A3 (overall average with model as proposed deviation) to B3 (individual average with model as proposed deviation). With the data obtained under the experiment, it is difficult to conclude the following definitely (because of the high leverage points in Graphs 4 and 5), but, visually, it appears that the actual deviation as a function of the proposed deviation is not overly affected by the change of basis from which the model is presented as a deviation. This may suggest that the subjects could be selecting an average level of the proposed deviation that they will use and that this level may be a function of the model and information set only and not of the basis from which the model is presented as a deviation. For business purposes, identifying if that effect was actually present would be relevant if one was attempting to forecast the impact of a change in information set on the ultimate usage of the model by users. In a marketing or sales setting, this could arise if users were presented with an array of prices (*e.g.* the cost at target profit, with the walk-away price, the suggested price of a first offer, the predicted valuation of the best alternative offer available to the customer, *etc.* presented as deviations).

Actual (Implicit) Deviation as a Function of Proposed (Implicit) Deviation by Treatment - Weeks 14 and 15

Graph 4: The actual (implicit) deviation from the model expressed as a function of the proposed (implicit) deviation, for treatments A1, A2, and A3, for weeks 14 and 15.

Note that the deviations from the model are implicit for treatments A1 and A2, as the subjects were not explicitly presented a proposed deviation. A slope of 0 means no agreement with the model, while a slope of 1 indicates complete agreement with the model. Similarly to what was found in Graph 2, the fitted line for treatment A3 approximately bisects the angle formed by the lines for A1 and A2. Note also that the proportion of the subjects' predictions that can be explained by the (implicit) proposed deviation going from treatment A1 to A3 to A2 increases.

Actual (Implicit) Deviation as a Function of Proposed (Implicit) Deviation by Treatment - Week 16
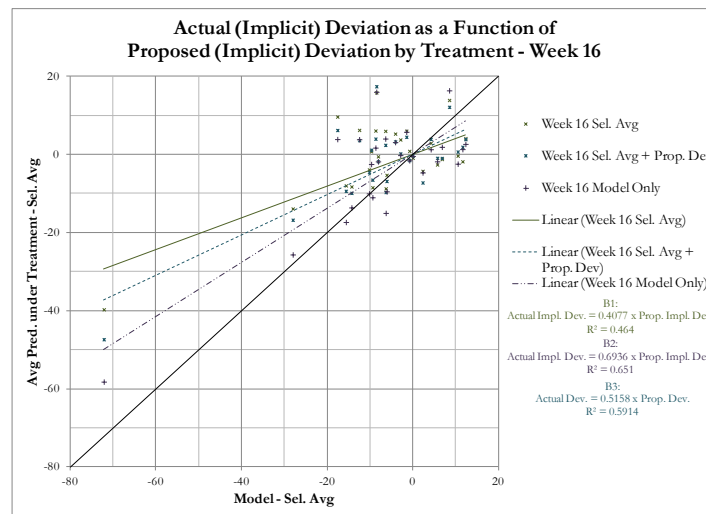
Graph 5: The actual (implicit) deviation from the model expressed as a function of the proposed (implicit) deviation, for treatments B1, B2, and B3, for week 16.

Note that the deviations from the model are implicit for treatments B1 and B2, as the subjects were not explicitly presented a proposed deviation. Similarly to what was found in Graph 3, the fitted line for treatment B3 approximately bisects the angle formed by the lines for B1 and B2. Note that the slope for treatment B1 is closer to 1 than the slope for treatment A1. The change in slope obtained by going from B1 to B3 is numerically similar to that obtained from going from A1 to A3; similarly, the change in slope obtained by going from B3 to B2 is numerically similar from that obtained by going from A3 to A2. As in Graph 4, note also that the proportion of the subjects' predictions that can be explained by the (implicit) proposed deviation going from treatment B1 to B3 to B2 increases.

3.2.3. Driver of actual model compliance

We examined if we could identify demographic characteristics of the subjects that could help us predict which subjects would follow the model more or less. As mentioned earlier, the prudence or

risk aversion of the subjects could have potentially affected the response of the subjects such that the exact nature of the compensation function (as opposed to its essence which was to encourage subjects to provide us with their best assessment) could have influenced the outcome of the experiment.

The only demographic dimension that seemed to influence the behavior of the subjects was the indicator for the Judging inclination of our (simplified) Myers-Briggs test. For that dimension, subjects that self-reported the Judging inclination tended to make greater use the model: note that the effect was not consistently observed as can be seen in Table 7.

Actual Deviation as a function of Proposed Deviation	Weeks 14 & 15				Weeks 16	
	A3		B3		B3	
	Coefficient	St. Err.	Coefficient	St. Err.	Coefficient	St. Err.
Intercept	0.25	0.06	0.33	0.04	0.35	0.06
Indicator for Judging Inclination	**0.19**	**0.07**	0.01	0.04	**0.18**	**0.06**

Table 7: Actual deviation expressed a function of the proposed deviation for those treatments with explicitly provided proposed deviations: that is, weeks 14 and 15 A3 and B3 as well as B3 for week 16.

For A3 in weeks 14 and 15 and B3 for week 16, the proportion of the proposed deviation actually used by the subject is materially and statistically higher. This could be explained by individuals with a judging inclination seeing "the need for most rules" and liking "to make & stick with plans" (PersonalityType.com/LLC n.d.).

A working hypothesis to explain that effect is that subjects self-reporting the Judging inclination reported, among other things, to see the need for most rules and prefer to stick with plans. It is then no great stretch to see model compliance as either thoughtful rule abiding (presuming a neutral or positive perception of the predictive model) or as a form of disciplining of the subject's own predictions.

3.3. Optimal Model Compliance

Given known over-confidence behavioral effects, we examined whether the subjects, that were somewhat using the model, were doing so optimally or whether the subjects were ignoring valuable insights from the model and over-weighting their own mental model.

For subjects that were not explicitly presented with the predictive model (that is, subjects that experienced the A1 and B1 treatments), we formed the convex combination of the subjects' predictions and the predictive model predictions that generated the most favorable compensation for the subjects. To obtain an estimated distribution for the level of optimal model compliance, we then repeated the exercise with bootstrapped samples. This distribution is presented below.

The identified optimal model compliance is about 70%. This suggests massive over-confidence of the subjects in their own differential predictive abilities as actual model usage is much lower than 70% when the subjects are presented with the model (30% for A2/B2 and 15% for A3//B3).

**Estimated Density Function of
Compensation at Optimal Model
for Subjects Given A Relevant Mc**
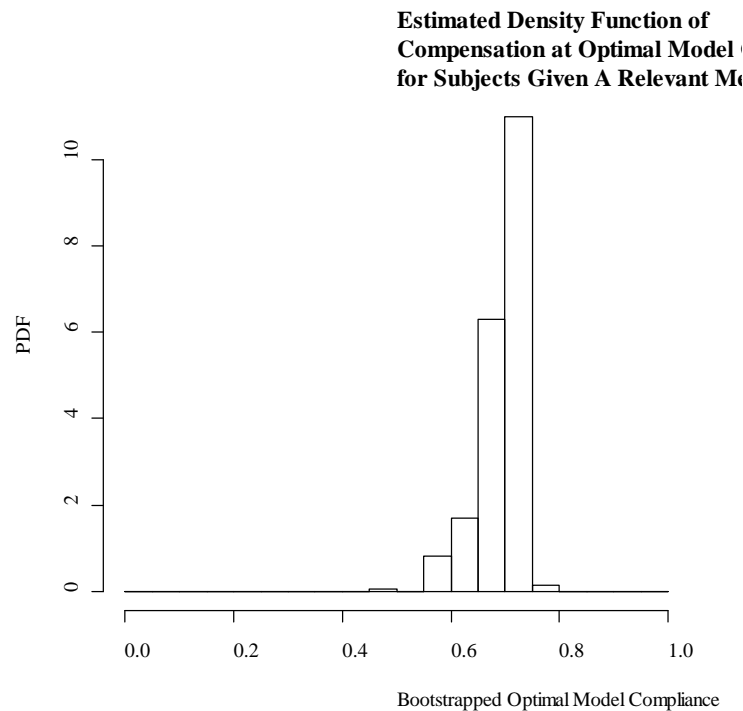
Graph 6: Estimated density function of the optimal model compliance.
The distribution of the estimated optimal compliance levels were obtained using (1) bootstrapped samples of predictions of subject receiving the A1 and B1 treatments, and (2) optimizing the average prediction compensation by taking a weighted average of the subject prediction and the model prediction. The full sample fitted value is about 70%.

Note that our estimation procedure presumed that subjects not explicitly presented with the model were effectively not influenced by the predictive model. While this effective assumption cannot be practically improved upon, it is less than perfect as the overall mean and the individual mean were variables that were part of the predictive model and thus correlated with the model.

3.3.1. Drivers of optimal model compliance

Contrasted with the drivers of actual model compliance are the drivers of optimal model compliance. Here, we are attempting to identify which subjects <u>should</u> have used the model more or less.

Here again, only one demographic dimension appears important: the subject self-reported familiarity with football. Subjects that self-reported they knew football the most were also the ones that should be using the model the least. However, even subjects that needed the model the least were still over-confident in their abilities: their model usage is bounded above by 30% and their optimal model usage was about 50%.

Bootstrapped Optimal Model Compliance	Self-Reported Familiarity with Football				
Familiarity Level:	1	2	3	4	5
Min.	0.00	0.00	0.00	0.00	0.00
1st Qu.	0.28	0.69	0.50	0.59	0.37
Median	0.69	0.71	0.61	0.69	0.49
Mean	**0.53**	**0.70**	**0.63**	**0.68**	**0.48**
3rd Qu.	0.79	0.72	0.78	0.78	0.65
Max.	1.00	1.00	1.00	1.00	1.00
Number of Subjects	94	32	7	8	7
Prop. of Subjects	64%	22%	5%	5%	5%

Welch Two Sample t-test of whether the mean of the optimal model usage is the same for 1 (very familiar) as for the other self-reported levels of familiarity with football.

Test statistic: -8.1172

Degrees of Freedom: 1179.003

p-value: 1.191×10^{-15}

Table 8: Optimal model compliance as a function of the self-reported level of familiarity with football. The distribution of the estimated optimal compliance levels were obtained using (1) bootstrapped samples of predictions of subject receiving the A1 and B1 treatments, and (2) optimizing the average prediction compensation by taking a weighted average of the subject prediction and the model prediction. Visually, level 1 (very familiar), levels 2 to 4, and level 5 should be binned together. Because the number of subjects reporting level 5 (very unfamiliar) is small, levels 2 to 5 are binned together. Note that the subject count only covers treatments A1 and B1, but counts them once per time within a week where they get treatment: keep in mind that, for each week, each subject received two treatments. Using the proposed binning, the mean optimal model compliance for level 1 is materially and statistically different from that of the other levels.

Note that subjects that repeated the experiment also needed the model less than other subjects, but it is also the case that subjects that repeated the experiment also had greater self-reported familiarity with football. This is a possibly natural association as subjects familiar with football should also be subjects that found the experiment more fun and thus worthy of repetition.

3.4. Perceptions of Credibility and Relevance

As mentioned in the 'Treatments descriptions' section, the subjects were asked about their perceptions of the information sets. Subjects' reported perceptions of the provided tools was thought to be relevant because, in a business setting, if employees feel that they do not have access to the necessary tools to do their work well, the irritation that the employees feel towards the employer may become so dramatic as to cause significantly lower levels of employee engagement and thus lead to decreased productivity.

PMF	Weeks 14 and 15 - C1	Weeks 14 and 15 - C2	Weeks 14 and 15 - C3	Weeks 14 and 15 - C4	Weeks 14 and 15 - C5	Week 16 - C1	Week 16 - C2	Week 16 - C3	Week 16 - C4	Week 16 - C5
A1//B1	11%	24%	25%	23%	16%	17%	41%	28%	12%	3%
A2/B2	5%	38%	37%	17%	3%	4%	35%	42%	18%	2%
A3	9%	25%	35%	25%	7%					
B3	9%	51%	27%	11%	2%	17%	38%	36%	9%	0%

CDF (from C5 to C1)	Weeks 14 and 15 - C1	Weeks 14 and 15 - C2	Weeks 14 and 15 - C3	Weeks 14 and 15 - C4	Weeks 14 and 15 - C5	Week 16 - C1	Week 16 - C2	Week 16 - C3	Week 16 - C4	Week 16 - C5
A1//B1	100%	89%	65%	39%	16%	100%	83%	42%	14%	3%
A2/B2	100%	95%	57%	20%	3%	100%	96%	62%	20%	2%
A3	100%	91%	67%	32%	7%					
B3	100%	91%	40%	13%	2%	100%	83%	45%	9%	0%

Table 9: Subject reported perception of credibility and relevance of provided information for A1, A2, A3 and B3 for weeks 14 and 15 and for B1, B2 and B3 for week 16. The table at the top is an empirical probability mass function over all possible responses from C5 (not at all confident in credibility and relevance) to C1 (highly confident in credibility and relevance). The bottom table is an empirical cumulative distribution function starting at C5.

Note that, if the provided information was unequivocally perceived as being more credible and relevant by the subjects than another alternative, then the empirical cumulative distribution function of credibility and relevance would be first-order stochastically dominated by that of the alternative. We see that providing only the model (A2) generates less negative perceptions of credibility and relevance than either of only providing the overall mean (A1) or providing the overall mean and the model as a proposed deviation (A3). Note, however, how providing the individual mean (as in B1 and B3) decreases the perceptions of negative credibility and relevance of information. This suggests the following simplified rule for predicting the perception of credibility and relevance: receiving the overall mean is less preferred than receiving the model only which is less preferred than receiving the individual mean. When a selected mean is provided, receiving the model as supplementary information does not appear to affect materially the perception of credibility and relevance.

The results from Table 9 suggest that subjects preferred least the information set under A1; that is, being only provided with the overall mean. Note that the subjects had a much less negative perception of the information set under B1 (only the individual mean). This alignment of subject preferences and internal subject model suggests that subjects prefer being provided with a model with which they agree even if, as is the case here, the model is actually not more predictive than other models. Recall that the overall mean and the individual mean are actually equally valuable individual pieces of information when presented alone and that this is largely due to the large regression to the mean at the inter-individual and intra-individual levels.[8]

---

[8] It is worthwhile to note that both the overall average and the individual average represent valid perspectives on the predictive problem at hand, as is evidenced by the fact that they are both as powerful one-way predictors of quarterback

When the model is presented, the subjects appear to prefer it when it is presented as a proposed deviation from the individual mean but the subjects appear to prefer it to be presented directly if the alternative is to receive the model as a proposed deviation from the overall mean. This, then, suggests that the base (from which the model is presented as a deviation of) is an important driver of subject preferences (over information sets).

Recalling that subjects appeared to revert back to their internal model when they experienced cognitive dissonance with the predictive model, this preference for a base in which they believe may be rationalized by the ease with which subjects can revert back to their own internal model when they do experience cognitive dissonance.

Considering only week 16, it does appear that the subjects do (marginally) prefer also having the model (as opposed to not having it) when they are presented with the individual mean.

3.5. Self-Reported Confidence

The other side of subject preference is the induced self-reported confidence of the subjects after they make their predictions.

The findings relating to self-reported confidence (in predictions) line up pretty well with findings concerning the perceptions of credibility and relevance of the provided information sets. As a general rule, when the information set was thought to be more relevant and credible, the self-reported confidence in predictions also improved.

As can be seen in Table 10, the slight exception to the general rule can be found for week 16 in the comparison of treatments B1 and B3 were the B1 subjects reported less negative self-confidence than the B3 subjects but where the B3 information set was perceived less negatively than the B1 information set.

PMF	Weeks 14 and 15 - C1	Weeks 14 and 15 - C2	Weeks 14 and 15 - C3	Weeks 14 and 15 - C4	Weeks 14 and 15 - C5	Week 16 - C1	Week 16 - C2	Week 16 - C3	Week 16 - C4	Week 16 - C5
A1//B1	3%	33%	46%	11%	8%	3%	51%	35%	12%	0%
A2/B2	5%	39%	39%	15%	3%	5%	40%	36%	16%	2%
A3	0%	40%	43%	17%	0%					
B3	1%	44%	41%	13%	0%	6%	39%	39%	15%	0%

CDF (from C5 to C1)	Weeks 14 and 15 - C1	Weeks 14 and 15 - C2	Weeks 14 and 15 - C3	Weeks 14 and 15 - C4	Weeks 14 and 15 - C5	Week 16 - C1	Week 16 - C2	Week 16 - C3	Week 16 - C4	Week 16 - C5
A1//B1	100%	98%	65%	19%	8%	100%	97%	46%	12%	0%
A2/B2	100%	95%	56%	17%	3%	100%	95%	55%	18%	2%
A3	100%	100%	60%	17%	0%					
B3	100%	99%	55%	13%	0%	100%	94%	55%	15%	0%

Table 10: Subject self-reported perception of confidence in their predictions for A1, A2, A3 and B3 for weeks 14 and 15 and for B1, B2 and B3 for week 16. The table at the top is an empirical probability mass function over all possible

ratings. As such, this experiment does not explore the effect of presenting a biased or voluntarily distorted statistic as a guide for prediction or as a base from which a model is presented as a proposed deviation.

responses from C5 (not confident at all) to C1 (very confident). The bottom table is an empirical cumulative distribution function starting at C5.

Similarly to Table 9, a treatment that induces greater self-reported confidence than another treatment would be stochastically dominated by the other treatment. The data above can be approximately summarized as follows: receiving the overall mean induces more negative self-reported confidence than receiving the overall mean and the model as a proposed deviation or just the model, which induces more negative self-reported confidence than receiving the individual mean and the model as a proposed deviation, with only receiving the individual mean inducing least negative self-reported confidence in the predictions.

3.6. Interpersonal Agreement

Before moving on to the net effect of treatment on the average compensation to the subjects, we wanted to examine the effect of the treatments on inter-personal agreement. This question is interesting because it has been found in the actuarial *versus* clinical debate that increased prediction consistency is an important factor in increased prediction accuracy, as consistency sets an upper bound on reliability.

As can be seen in Table 11, the inter-personal prediction consistency is roughly similar across treatments. However, subjects given the B3 treatments seem to have made more consistent predictions. The results of Table 11A confirm this diagnostic as the linear model of across subject standard deviation and variance of predictions for a given quarterback/week has a highly significantly negative average treatment effect for the B3 treatment.

Having seen in the "Actual Model Compliance" section that subjects given the B3 treatment use the predictive model about half as much as subjects given the A2/B2 treatments, we can now wonder if the increased consistency of predictions in the B3 treatment can empirically counter-balance the decreased model usage and allow the average compensation of the B3 treatment to be similar to that of the A2/B2 treatment.

Across Subjects Standard Deviation of Predictions	Weeks 14 and 15		Week 16	
	Average	St. Dev.	Average	St. Dev.
A1//B1	11.9	2.7	9.7	1.6
A2/B2	10.5	3.0	10.4	2.6
A3	11.3	3.1		
B3	9.6	2.4	10.4	2.5

Table 11: The average and standard deviation of the standard deviation of across subjects standard deviation of prediction for A1, A2, A3 and B3 for weeks 14 and 15 and for B1, B2 and B3 for week 16.

The inter-personal agreement of predictions is roughly similar across weeks and treatments. However, a potentially interesting difference is between B3 and A1 for weeks 14 and 15: B3 subjects seem to have been more consistent than A1 subjects.

	Standard Deviation			Variance		
Treatment #	Estimate	St. Error		Estimate	St. Error	
Base Level	11.15	0.29	***	130.86	6.67	***
A2/B2	-0.68	0.40	.	-12.98	9.44	
A3	0.12	0.45		5.53	10.55	
B3	-1.27	0.40	**	-27.53	9.44	**

Table 11A: A linear model of across subject standard deviation and variance across quarterback/weeks prediction sets, by treatment.

The linear model of across subject standard deviation (and variance) for quarterback/week predictions is consistent with the results found in Table 11: treatment B3 induces increased consistency of predictions across subjects. Increased consistency of predictions can be a driving force of prediction accuracy, as was empirically found in the actuarial *versus* clinical debate.

3.7. Net Compensation Outcomes

We are now ready to analyze the net compensation outcomes by treatment. Because there are only three weeks of data, some caution is necessary in the interpretation of the results. In particular, there were only 90 quarterback/weeks in the sample; therefore, the performance comparison between the subjects and the predictive model must be examined with prudence. Moreover, even across treatments comparisons must be examined with care as the particular observed outcome differential was obtained under non-experimentally designed variations in information sets: that is, the observed difference in outcomes may reflect a particular mix of information sets induced by chance. In particular, the information sets were not designed to induce the optimal variation in variables that could have been of interest such as in generating a wide and balanced variety of (1) proposed deviation, (2) difference between the uninfluenced (average) subject prediction and predictive model predictions, (3) overall mean, (4) individual mean, *etc*.

Compensation (per prediction) ($)	Weeks 14 and 15			Week 16			All Weeks		
	Average	Nb of Pred.	St. Dev.	Average	Nb of Pred.	St. Dev.	Average	Nb of Pred.	St. Dev.
A1//B1	0.1770	1 137	0.0051	0.1734	1 004	0.0052	0.1753	2 141	0.0036
A2/B2	0.1826	1 080	0.0052	0.1816	795	0.0060	0.1822	1 875	0.0039
A3	0.1803	992	0.0053				0.1803	992	0.0053
B3	0.1921	1 160	0.0050	0.1701	951	0.0054	0.1822	2 111	0.0037
Total	0.1832	4 369	0.0026	0.1746	2 750	0.0032	0.1799	7 119	0.0020
Total excl. A1/B1	0.1853	3 232	0.0030	0.1754	1 746	0.0040	0.1815	4 978	0.0024
Model	0.1700			0.1820			0.1746		

Table 12: The average compensation per prediction (with the associated number of predictions and standard deviation of compensation) for A1, A2, A3 and B3 for weeks 14 and 15 and for B1, B2 and B3 for week 16; the table includes sub-totals for those treatments where the subjects saw the model and the compensation that would have been obtained under the model.

There is a significant difference in compensation between A1 and B3 for weeks 14 and 15. The sources for the difference could be (1) increased prediction consistency, (2) increased perception of credibility and relevance of information, (3) increased quality of information provided. Note, however, that other differences do not appear as significant. This suggests that subjects needed to be provided with a reference that they believed in more and a model prediction to significantly increase their compensation performance. Keep in mind that the compensation scheme was constructed to incentivize subjects to attempt to be as correct as possible with regards their predictions.

When the subjects had access to the model, they performed better than without access to the model. In weeks 14 and 16, they did not (on average) beat the model; but they did (on average) beat the model in week 15.

From Table 12, we can see that treatments A2/B2 and B3 generated the same average compensation across the three weeks. Treatment A3 comes next. The clear loser in terms in generating favorable compensation for the subjects was treatment A1 with average compensation significantly and materially lower than the average compensation for treatment B3 in weeks 14 and 15. Compared to the compensation that would have been generated if the subjects had perfect model compliance, the subjects fared better than the predictive model over the course of the three weeks: even those subjects that were not presented with the model. Again, some caution is necessary here since the predictive model would have beat the subjects (on average) under all treatments for weeks 14 and 16, but the predictive model performed poorly in week 15. The evidence does suggest that many of the model predictions were useful to the subjects (even for week 15) as subjects that had access to the model fared better than those that did not have access to it.

It is also worthwhile to re-examine the results obtained in the 'Optimal Model Compliance' section: what would have been the average compensation of the subjects had they been using the model optimally? Table 13 answers this question. We can see that, under optimal model usage, the subjects would have generated an average compensation (0.1858) 2% higher than what they generated on average under treatments A2/B2 and B3 (0.1822). While a 2% improvement may not appear large at a first glance, if one thinks about the effect on Return on Equity of a 2% increase in the net income ratio, then this improvement appears much more substantial.

	Average Compensation per Prediction at Best Model Usage ($)
Mean	0.1858
St. Dev.	0.0036
1st Q	0.1834
Median	0.1857
3rd Q	0.1880

Table 13: Estimated features of the distribution of compensation of subjects if they used the model optimally.

This table represents the compensation the subjects would have attained by following the optimal weighted average of their unaided assessment and the model. Notice that this is statistically and materially (being a 2% performance improvement over the average compensation obtained under A2/B2 and B3) significant.

The results from Table 13 beg the question of what would have been the optimal (average) compensation for subjects that were already influenced by the predictive model.

(Further) Model Use	Relative Gains			
	A1//B1	A2/B2	A3	B3
0%	0.0%	3.9%	2.9%	3.9%
25%	2.9%	4.3%	4.3%	4.6%
50%	5.2%	4.7%	5.3%	5.0%
75%	5.8%	4.7%	5.7%	4.9%
100%	0.1%	0.1%	-3.0%	0.1%

Table 14: Estimated gains from further model use by treatment.

The results shown in the above table can be interpreted in the following way. At no supplementary model use, the best outcomes are achieved under A2/B2 and B3, consistent with table with Table 12. At full model use, we obtain that there would be a gain compared to the unaided prediction under every treatment but A3, because of the poor performance of the model in week 15, as noted in the comments of Table 12. Consistent with the result of Table 13, in the table above, the best performance for treatments A1//B1 is achieved at 75% model for a 5.8% performance improvement. At best model usage, however, the optimal attainable performance is highest under the A1//B1 treatments. This may occur because the subject predictions correlate more with the model predictions when the subjects are presented with the model. Compare this with the engineering problem of identifying the best mix of instruments and weights assigned to instruments to generate the 'best' measurement: one would pick instruments whose measurement errors would be as little positively correlated as possible. In this case, it appears that the subjects are 'destroying' some of the statistical signal that they would be picking if they did not see the model when they are explicitly presented with the model.

Table 14 suggests that subjects that were already influenced by the predictive model could not possibly achieve a better (average) compensation by using a weighted average of the predictions under the treatment and the predictive model predictions. One working hypothesis one might have had is that the optimal attainable (average) compensation should not be a function of the treatment. However, think of the following analog problem. Suppose we are attempting to take a measure of an empirical reality. First, suppose that we have only two (unbiased) instruments, each with their own level of precision. If measurement errors of the two instruments were independent, then, to obtain the least variable and unbiased measurement, we should weight together the measurements of the two instruments such that more weight is assigned to the more precise instrument. Holding constant the precision of the two instruments, one would prefer the measurement errors of the two instruments to be as negatively correlated as possible, such that the error of one instrument should be naturally corrected by the error of the other instrument. However, in our case, we get a case where the measurements taken by the subjects under A2/B2, B3 and A3 are positively correlated with the predictive model predictions. Thus, because of this, at the optimal compensation, the optimal compensation is higher under the A1//B1 treatments. Nonetheless, the net model usage at optimum (average) compensation [not shown] are quite similar across treatments.

3.8. External Validity Considerations

Just as is the case for any social science experiment, one needs to examine the potential transferability of the results of the experiment to other (often, non-experimental) circumstances.

3.8.1. Non-neutrality of the compensation scheme and environment

In "An Effort Based Analysis of the Paradoxical Effects of Incentives on Decision-Aided Performance" (Samuels and Whitecotton 2011), the researchers found that

> [i]n contrast to the findings of prior research, our study shows that incentives do not necessarily decrease performance in the presence of decision aids. Rather, we demonstrate that the effect of incentives on decision-aided performance depends on other contextual factors such as the absence or presence of additional contextual information." (345)

Thus, it is quite possible that changing the compensation scheme or the context of the experiment may affect the findings and, therefore, may make the findings non-transferable. Here is a potential example of such non-transferability. Imagine we go back to our example of the use of predictive models in a marketing or sales context. Now, imagine that the objective of the users of the predictive model is not to provide their best assessment of a future statistic but, instead, to try to optimize sales (subject to some profitability constraints). Even further, imagine that there are negotiations with a third party involved. In that case, while the final selection of the user is expected to be influenced by the model output (when provided), there is no representation that the user selection is a best assessment and may instead represent the impact of external constraints imposed by the third party. There is *a priori* no reason to think that these two contexts would generate similar results. However, in a context where the final price of a transaction is rationally related to the expectation of a future profitability statistic and where compensation to the users of the decision-aid is related to their (individual) accuracy in predicting that future profitability statistic, then the context of the experiment and the pricing situation become alike enough to expect some level of transferability. For example, this would naturally occur in an insurance pricing context, but also in the pricing of many financial products.

4. CONCLUSION

At this point, we wish to interpret the results of the experiment from a business perspective (in particular for insurance or retail financial products, like mortgages). Assume that the interest of the business is to have the users of the decision-aid be as accurate as possible in predicting the future performance of a profitability statistic attached to a (sold) contract. In our closely related experimental setting, which decision-aid would the business prefer to provide to the users? Assuming that the difference in costs in providing the different decision-aids were not material, then the business would prefer to provide a decision-aid that presents either only the predictive model or the predictive model presented as a proposed deviation from a statistic that the users of the

decision-aid find relevant and credible (in the case of the experiment, that was the individual mean) as they generate the same accuracy in user predictions. Which decision-aid would the users prefer to receive? Presumably, the users would prefer to receive higher compensation, be more confident in their predictions and find the decision-aid they receive relevant, and it is unclear *a priori* in what order. For our purposes, the choice of the subject should be insensible to the exact nature of the preference because the same basic ranking comes up under any weighting of the preferences: the users would prefer to receive either only a statistic they perceive as relevant and credible or the model presented as a proposed deviation from the same statistic, as these two decision-aids generate similar compensation, similar self-confidence and similar perceptions of relevance of information. So, altogether, this implies that the decision-aid that should be deployed is the decision-aid with the predictive model presented as a proposed deviation from a statistic that is perceived by the subjects to be relevant and credible.

Further, applied research needs to be done at the business level: our research has not identified what features of the individual mean statistic made it so attractive to subjects and, even if we had done that, it is unclear that this piece of the research would be transferable. Note, however, that the experimental framework that we used should be implementable in an applied (business) research framework at relatively little costs. This means that the applied researchers can conduct meaningful applied (business) research, with an experimental inclination, without needing to develop tools to the point where they are ready to be deployed in a production environment: it should be apparent from our research that significant insights can be gathered in a simplified, even to the point of being skeletal, framework.

The current research does leave open many empirically interesting questions. For example, what would be the best way to seek out the subjective opinion of the subjects to arrive at the most predictive best assessment? Should the opinion of the user be sought and then the final prediction be generated mechanically from the recorded prediction and the output of the predictive model? If the user always needs to have a final say in the recorded prediction, should the users be limited in their ability to deviate from the predictive model? For example, should the users be limited in their freedom to deviate from the model only globally or should there be a limit on the ability of users to deviate from the model that applies prediction by prediction (or both)? Is there a way to ensure that the deviations from the model reproduce some form of distribution known to hold for the underlying population (that the predictions relate to) as a whole?

Acknowledgments

Appendix A. On-line Survey Tool

Consent Block

UNIVERSITY OF WISCONSIN-MADISON

Subject CONSENT to Participate in Research Study "Quarterback Rating Experiment"

Title of the Study: Quarterback Rating Forecasts
Principal Investigator (*PI*): Justin Sydnor (phone: 608-263-2138, email: jsydnor@bus.wisc.edu)
Mailing Address: 5287 Grainger Hall, Wisconsin School of Business, 975 University Ave., Madison, WI, 53706

Introduction

You are invited to participate in this research study about forecasting.
We are studying how people make predictions about quarterback ratings in upcoming NFL games. You are invited to take part because you are a student at UW-Madison. Note that you must be a citizen of the United States of America to participate in the study: this is because we can only provide compensation to American citizens. Your participation is voluntary.

Procedures

If you decide to participate in this research, you will be asked to forecast the quarterback rating of quarterbacks expected to start in the coming weekend of NFL activity. We will also collect information about you for this research study. This information includes gender, year of birth, citizenship, attained education level, major, GPA. We will also ask you about your familiarity with, interest in and understanding of football, sports statistics, and fantasy sports. We will also ask questions to assess some of your personality traits. This questionnaire will be conducted with an online Qualtrics-created survey.

Risks/Discomforts

The only risk of taking part in this study is that your study information could become known to someone who is not involved in performing or monitoring this study.

Benefits

You are not expected to benefit directly from participating in this study. Your participation in this research study may benefit other people by helping us learn more about how individuals make decisions. There are no direct benefits to you from participating in this research.

Compensation

You will receive a compensation that will be determined as a function of your forecasts and the actual outcomes in the coming weekend of NFL football. Your compensation will range from zero (0) to fifteen (15) dollars. If you choose to participate, the exact way to compute your compensation will be described within the survey. Once the actual values for the forecasts you made are known, we will tally up your compensation and send you an e-mail to let you know where you can pick them up on campus. Once you have received your compensation, your name and e-mail address will be removed from the databases related to the experiment.

Confidentiality

For compensation purposes, we will ask you to provide us with your name and your preferred e-mail address: we need this information because your compensation is determined based on your answers to the survey. We may provide your name and preferred e-mail address to support staff so that these persons can give you your compensation. You will also need to complete a Participant Payment Disclosure Form (i.e., Subject Log) in order to be paid. Once you receive your compensation, we will delete your name and e-mail address from our records. After that, all other data will be stored indefinitely on a secure location on campus in a faculty member or graduate student computer.

Participation

Your participation is voluntary. You do not have to continue with this on-line survey and you may refuse to do so. If you refuse to continue, however, you cannot take part in this research study. You may completely withdraw from the study at any time without penalty. You also may choose to cease participation or skip any questions that you do not feel comfortable answering.

Questions about the Research

Please take as much time as you need to think over whether or not you wish to participate. If you have any questions about this study at any time, contact the Principal Investigator Justin Sydnor at 608-263-2138. If you are not satisfied with response of research team, have more questions, or want to talk with someone about your rights as a research participant, contact the Social and Behavioral Science Institutional Review Board at the University of Wisconsin-Madison: 310 Lathrop Hall, 1050 University Avenue, Madison, WI 53706, phone: 608-263-2320.

1. STATEMENT OF CONSENT

By entering the information below, you acknowledge that:
_*You have read the above information.
You have received answers to the questions you have asked.
You consent to participate in this research.
You are an American citizen.
You are at least 18 years of age.*_

Type your first and last names

Type your preferred email address (so that we can communicate to you the exact details of where and when you can pick up your compensation).

2. Do you agree to participate in this study?

* Yes
* No

Demographics Block

The following questions are about demographics.

What is your citizenship?

* United States of America
* Other

What is your gender?

* Male
* Female

What year were you born?

Are you a ___?

* Freshman
* Sophomore
* Junior
* Senior
* Graduate
* Other

What is your major? (If more than one choice may apply, pick the one
you most enjoy.)

What is your current Grade Point Average?

On a scale of 1 (very comfortable) to 5 (having difficulty), how would you rate your mathematical abilities?

Football Trivia

The following questions are football trivia.

Which team won the last Supe Bowl (played in February 2012)?

* Cleveland Browns
* Green Bay Packers
* New York Giants
* University of Wisconsin-Madison Badgers
* New England Patriots

Who is a quarterback for the Green Bay Packers?

* Phil Esposito
* Aaron Rodgers
* Tom Brady
* Peyton Manning
* Tim Tebow

Which of these players is a defensive end who, in the 2011-2012 season, was a member of the Super Bowl winning team, went to the Pro Bowl, and lead his team for the number of sacks in the season?

* Jason Pierre-Paul

* Osi Umenyiora
* Russell Martin
* J.J. Watt
* Clay Matthews

Which of these players has posted the most games with a perfect passer rating?

* Dan Marino
* Joe Montana
* Peyton Manning
* Ben Roethlisberger
* Steve Young

Familiarity with Football and Fantasy Sports

In the following questions, you will be asked about your familiarity with football and fantasy sports.

On a scale of 1 to 5 (where 1 is very familiar and 5 is very unfamiliar), how would rate your own familiarity with fantasy sports?

1 (very familiar)   2   3   4   5 (very unfamiliar)

On a scale of 1 to 5, how would rate your own familiarity with football?

1 (very familiar)   2   3   4   5 (very unfamiliar)

On average, how many days per week do you watch or read sports news?

1   2   3   4   5   6   7

Do you currently have a fantasy football team?

Yes   No

If you do have a fantasy football team, please rate your ability at fantasy football from 1 (very good) to 5 (very poor).

1 (very good)   2   3   4   5 (very poor)

Task Description and Compensation Scheme

This page describes the task we ask you to complete in this survey. We will also describe the exact formula that will be used to compute your compensation.

*Task Description*:

For the 30 or so quarterbacks that are expected to start the game in the upcoming weekend of NFL football, you will be asked to provide your forecast of the quarterback rating for the week for each of these quarterbacks.

Let us share with you some information about the statistic that we are asking you to forecast.

According to Wikipedia, passer rating
<http://en.wikipedia.org/wiki/Passer_rating> is a measure of the performance of quarterbacks. Passer rating is calculated using each quarterback's completion percentage, passing yardage, touchdowns and interceptions. A perfect passer rating in the NFL is 158.3. A perfect rating requires at least a 77.5% completion rate, at least 12.5 yards per attempt, a touchdown on at least 11.875% of attempts, and no interceptions.

Here are some _*facts about quarterback ratings*_ for this season:
- the _*average*_ for starting quarterbacks is about *_90_*,
- on any given week, _*about half of the quarterback rating will be between 75 and 125*_, and
- about _*75% of the time ratings fall within about 30 of the quarterback's own season average*_.

*Compensation Scheme:*

You will be compensated for your participation based on how accurately you are able to forecast quarterback ratings this week. Here we describe the exact formula we will use to determine your compensation.

Your earnings in this experiment will be based on how accurate your predictions are. You can earn up to $0.50 for each prediction, for a total possible earnings of $15 if you manage to perfectly predict the rating of each quarterback this week. Any prediction that is off by more than 25 points will earn you no money for that prediction. So if you miss each prediction by 25 points or more, you will earn nothing in the experiment. Your goal here should simply be to try to make each prediction as accurately as you think possible.

The following table describes the compensation scheme. For the cases in between, the compensation will be obtained by interpolating between the values in the following table.

Error = Your Forecast - Actual Value	Your Compensation ($)
+/- 30 (or more)	0.00
+/- 25	0.00
+/- 20	0.10
+/- 15	0.20
+/- 10	0.30
+/-5	0.40
0	0.50

Finally, your total compensation will be the sum of your compensation for each of your individual forecasts.

   Suppose that you have made the following forecasts with the attached actual values, what would be your compensation?

Player	*Your Forecast*	*Actual QB Rating*
Quarterback #1	100	133
QB #2	60	50
QB #3	90	95

 * 0.1
 * 0.3
 * 0.5
 * 0.7
 * 0.9

The correct answer was 0.7.

Sensitize Info Set

In the coming pages, we'll ask to make two sets of predictions. The survey will randomly select what extra information, if any, you'll be provided with to make your forecasts. The extra information will be described at the top of the page.

Forecasts - Treatment 1

You will now be asked to make your forecast for half of the quarterbacks.

For each of the listed quarterback below, please enter your forecast for their quarterback rating in the coming game.

You will find in brackets, first, their own team, second, the team they are playing against.

_*You will also find the overall average of the quarterback rating for the starting quarterbacks (OAAvg).

*_ Tom Brady (Patriots, @49ers) _*OAAvg*_: 85.8
(…)

Forecasts - Treatment 1 - Retrospective Confidence

How confident do you feel about your first series of predictions?

1 (very confident) 2   3   4   5 (not confident at all)

How confident do you feel about the credibility and relevance of the information you were provided with for the first series of predictions?

1 (highly confident in credibility and relevance)   2   3   4   5 (not at all confident in credibility and relevance)

We would be interested in learning about the strategies you used in making your predictions. Feel free to tell us about the strategies that supported your prediction choices.

Sensitize 2

In the next page, we'll ask to make the second set of predictions. The survey will randomly select what extra information, if any, you'll be provided with to make your forecasts. The extra information will be described at the top of the page.

_*Note that this extra information may *__*(or may not)*__*differ from the extra information you were provided with for the first selection*_

Forecasts - Treatment 2

You will now be asked to make your forecast for the other half of the quarterbacks.

For each of the listed quarterback below, please enter your forecast
for their quarterback rating in the coming game.

You will find in brackets, first, their own team, second, the team they are playing against.

_*You will also find the overall average of the quarterback rating for the starting quarterbacks (OAAvg).

*_ Jay Cutler (Bears, @Packers) _*OAAvg*_: 85.8
(…)

Forecasts - Treatment 2 - Retrospective Confidence

How confident do you feel about your second series of predictions?

1 (very confident) 2   3   4   5 (not confident at all)

How confident do you feel about the credibility and relevance of the information you were provided with for the second series of predictions?

1 (highly confident in credibility and relevance)   2   3   4   5 (not at all confident in credibility and relevance)

We would be interested in learning about the strategies you used in making your predictions. Feel free to tell us about the strategies that supported your prediction choices.

Ex Post Demographics - MBTI

The following questions are aimed at facilitating the understanding of your personality type.

Below, you will find two descriptions. Select the set that best corresponds to you.

A
    * I Make decisions objectively
    * I appear cool and reserved
    * I am most convinced by rational arguments
    * I am honest and direct
    * I value honesty and fairness
    * I take few things personally
    * I am good at seeing flaws
    * I am motivated by achievement
    * I argue or debate issues for fun

B
    * I decide based on my values & feelings
    * I appear warm and friendly
    * I am most convinced by how I feel
    * I am diplomatic and tactful
    * I value harmony and compassion
    * I take many things personally
    * I am quick to compliment others
    * I am motivated by appreciation
    * I avoid arguments and conflicts

A   B

Below, you will find two descriptions. Select the set that best corresponds to you.

A

* I focus on details & specifics
* I admire practical solutions
* I notice details & remember facts
* I am pragmatic
* I live in the here-and-now
* I trust actual experience
* I like to use established skills
* I like step-by-step instructions
* I work at a steady pace

B
* I focus on the big picture & possibilities
* I admire creative ideas
* I notice anything new or different
* I am inventive
* I think about future implications
* I trust my gut instincts
* I prefer to learn new skills
* I like to figure things out for myself
* I work in bursts of energy

A   B

Below, you will find two descriptions. Select the set that best corresponds to you.

A
* I have high energy
* I talk more than listen
* I think out loud
* I act, then think
* I like to be around people a lot
* I prefer a public role
* I can sometimes be easily distracted
* I prefer to do lots of things at once
* I am outgoing & enthusiastic

B
* I have quiet energy
* I listen more than talk
* I think quietly inside their heads
* I think, then act
* I feel comfortable being alone
* I prefer to work "behind-the-scenes"
* I have good powers of concentration
* I prefer to focus on one thing at a time
* I am self-contained and reserved

A   B

Below, you will find two descriptions. Select the set that best corresponds to you.

A
* I like to have things settled

  * I take responsibilities seriously
  * I pay attention to time & am usually prompt
  * I prefer to finish projects
  * I work first, play later
  * I seek closure
  * I see the need for most rules
  * I like to make & stick with plans
  * I find comfort in schedules

B
  * I like to keep my options open
  * I am playful and casual
  * I am less aware of time and may run late
  * I prefer to start projects
  * I play first, work later
  * I may have difficulty making some decisions
  * I question the need for many rules
  * I like to keep plans flexible
  * I want the freedom to be spontaneous

A   B

Ex Post Demographics - Other

For the following questions, you are asked to find the applicability of the statement to you on a scale of 1 ("I'm very much like that") to 5 ("I am not at all like that").

  My friends would say I am cautious.

1 ("I'm very much like that")   2 3 4 5 ("I am not at all like that")

  Being financially cautious is important to me.

1 2 3 4 5

  I like statistics.

1 2 3 4 5

  I like SPORT's statistics.

1 2 3 4 5

  I enjoy reading the sports news.

1 2 3 4 5

I enjoy reading about sport statistics.

1   2   3   4   5

I quite often feel that things are set in my life and I can't change them.

1   2   3   4   5

I'm aware that, while I can't always control what happens around me, I do control my own reaction to said events.

1   2   3   4   5

I believe that when people find themselves in bad situations, it's usually due more to unlucky circumstances.

1   2   3   4   5

5. REFERENCES

[1]   Gillam, W. R., and R. H. Snader. "Fundamentals of Individual Risk Rating, Part I." *National Council on Compensation Insurance (Study Note)*. 1992.

[2]   Goldberg, L. R. "Man versus models of man: A rationale, plus some evidence, for a method of improving on clinical inferences." *Psychological Bulletin* 73 (1970): 422-432.

[3]   Hunton, James, Vicky Arnold, and Jacqueline L. Reck. "Decision Aid Reliance: A Longitudinal Field Study Involving Professional Buy-Side Financial Analysts." *Contemporary Accounting Research* 27, no. 4 (2010): 997-+.

[4]   Johnson, Hershey, Meszaros, and Kunreuther. "Framing, Probability Distortions, and Insurance Decisions." *Journal of Risk and Uncertainty* 7 (1993): 35-51.

[5]   Lilien, GL, G. L., A. Rangaswamy, G. H. van Bruggen, and et al. "Bridging the marketing theory-practice gap with marketing engineering." *Journal of Business Research* 55, no. 2 (02 2002): 111-121.

[6]   McNamara, G., and P. Bromiley. "Decision making in an organizational setting: Cognitive and organizational influences on risk assessment in commercial lending." *Academy of Management Journal* 40, no. 5 (10 1997): 1063-1088.

[7]   PersonalityType.com/LLC. *Take Our Free PersonalityType.com Assessment*. http://www.personalitytype.com/career_quiz (accessed 10 01, 2012).

[8]   Peterson, D. K., and G. F. Pitz. "Confidence, Uncertainty, and the Use of Information." *Journal of Experimental Psychology - Learning, Memory, and Cognition* 14, no. 1 (January 1988): 85-92.

[9]   Peterson, D. K., and G. F. Pitz. "Effect of input from a mechanical model on clinical judgment." *Journal of Applied Psychology* 71 (1986): 163-167.

[10] Peterson, D. K., and G. F. Pitz. "Effects of Amount of Information on Predictions of Uncertain Quantities." *Acta Psychologica* 61, no. 3 (April 1986): 229-241.

[11] Samuels, Janet A., and Stacy M. Whitecotton. "An Effort Based Analysis of the Paradoxical Effects of Incentives on Decision-Aided Performance." *Journal of Behavioral Decision Making* 24, no. 4 (10 2011): 345-360.

[12] Simon, Herbert. "A Behavioral Model of Rational Choice." *Quarterly Journal of Economics* 69, no. 1 (February 1955): 99-118.

[13] Tapscott, Don, and Anthony D. Williams. *Macrowikinomics: New solutions for a Connected Planet*. Toronto, ON: Portfolio Penguin, 2012.

[14] Tversky, and Kahnemen. "Judgment under Uncertainty: Heuristics and Biases." *Science* 185 (1974): 251-284.

[15] Zimmer, I. "A comparison of the predictive accuracy of loan officers and their linear-additive models." *Organizational Behavior and Human Performance* 27 (1981): 69-74.

Biographies of the Authors

Marc-André Desrosiers is a Property/Casualty actuary working for a large insurer in Canada as a Research and Development project manager. He is interested in all aspects of ratemaking.

Weaving Actuarial Stories

Marc-André Desrosiers, FCAS, MBA

Abstract: Given that actuaries are using (informal) economic theories in their work (to build useful databases, to support the implementation process of proposed policies, etc.), it is worthwhile to understand the way these (informal) theories function and to evaluate their quality. We will construct a view of economic thinking that shows that economic theories fundamentally function like stories, like narratives. With that view in hand, we will then be able to highlight common mistakes in actuarial work as well as propose alternate views for the future work of actuaries.

Keywords: Credible Worlds, Imitation of an Action, Philosophy of Science, Epistemology of Economics, Narrative Theory.

1. INTRODUCTION

During the course of my career as a ratemaking actuary for Property/Casualty insurance products, I have come to notice that the practice of actuarial pricing may rely heavily on statistical methodology, but it also depends heavily on the capacity of the actuary to make sense of the underlying economics of the purchase of insurance. For instance, when it comes time for the actuary to select which variables to examine whether in a costing model, a buying ratio predictive model, a predictive model of deviations from system rates or a model of price elasticity, the modeling actuary needs to form, at least, an informal theory of the mechanisms in play to be able to dig through immense databases to isolate and compose variables that have the potential to be informative. Moreover, since ratemaking is an activity that does not end until the prices make their way to the markets, and maybe not even then, the pricing actuary needs to worry about whether or not intermediate users of the models, such as underwriters and brokers, will accept the proposed pricing models. To pass that hurdle, the pricing actuary often will present the rationale for broad and specific elements of the model by appealing to, sometimes, informal theories that the stakeholders of the insurer have about what drives the buying and claiming behaviors of the insureds. In effect, what I have come to observe is that significant efforts of theorizing of choice behavior under risk are taken by practicing field actuaries.

While the theorizing activities mentioned above often take place at an informal level, their cumulative resulting effects can be material to the success of the activities of the insurer: largely because they are so deeply related to policy-making with regards to pricing, the key driver of

profitability (and thus solvability and existence) of the insurer.

But, for some practicing and many academic actuaries, it is not even obvious that economic modeling does or should enter their practice. To them, the actuary should be focusing on predictive modeling with an absolutely agnostic attitude regarding the nature of the processes driving the data[1]. This attitude can be contrasted with two other 'schools' of statistical/probabilistic work oriented towards economic applications. One is the causal modeling orientation to estimation[2]. Two is the structural approach to estimation[3]. In both of these approaches, the modeler gets heavily involved with theorizing. Under the causal approach, the modeler gets involved because some form of *a priori* reasoning needs to take place in the selection of statistical instruments (given the humanly or economically relevant selected question of interest). Under structural modeling, the modeler needs to become quite familiar with the underlying (economic) theoretical model to be able to write down conditions the models predict would be observed in empirical reality. Either way, agnosticism about the underlying reality that generated the data is a no-go.

While we are not proposing that actuaries stop using the (predictive modeling) tools that have made them successful in their environment, we are proposing that something important is lost with such pure agnosticism. That loss even affects the statistical aspect of the work. Take the example of building and calibrating a rating algorithm. With a completely agnostic attitude, all the variables that are statistically significant would be kept. Reflecting on the underlying (economic) reality that generated the data would help the actuary (1) make reasonableness checks about the nature and strength of the relationships found, (2) avoid over-fitting of the data because of the work done in (1), and (3) go through a healthy dose of story-telling about the selected models that can help the stakeholders of the actuarial work wrap their head about the significance of the model.

We will thus explore here what makes a theoretical economics model 'good'. We do so in the hope that actuaries will be better equipped with answering the question of whether or not the (economic) theories they are using are well constructed and relevant to the problems at hand. With that exploration in hand, we will then explore implications relating to: (1) the asymmetry between

---

[1] As examples of work that are much in line with the said attitude, see (Regression Modeling with Actuarial and Financial Applications 2010) and (Loss Models: From Data to Decisions 1998).

[2] A good example of that type of work can be found in (Mostly Harmless Econometrics: An Empiricist's Companion 2009). One unfortunate aspect of causal-style estimation is that it is never obvious to what extent the identified causal factor will reproduce in situations that are not exactly like the situations in which the causal factor was identified.

[3] Good examples of that type of approach can be found in the examples of (Analysis of Panel Data 2003). An unfortunate aspect of structural estimation is that, at times, only factors of profoundly unrealistic models can be estimated due to mathematical tractability issues.

past and future, (2) the difference between regularity and causality, (3) why actuarial science needs to move beyond statistics and into econometrics, (4) how to identify model blind spots and comment on their importance, (5) what are alternative considerations into building a good rating system, and (6) how actuaries can get involved into the human side of claims.

1.1 RESEARCH METHODOLOGY

To explore how (many) good economic theories are constructed, we will start by relating some commonly known economic theories that actuaries sometimes share with each other. These stories relate to adverse selection, moral hazard and the pricing cycle: all subjects deep at the heart of actuarial practice.

We will then introduce the concepts of narrative causality and of the imitation of an action. On that subject, we will hear thoughts coming from different traditions: from economists reflecting on the trade of economic theorizing and its relationship to economics at large, from Paul Ricoeur (through the words of one of his commentators[4]) reflecting on narrative theory and on the practice of the historical science, through Ricoeur's work, from a range of thinkers that reflected on narrative theory, the historical science, the functioning of language and models, etc., and from applied mathematicians building concepts that allow them to build models relating to human activity.

With these tools in hand, we will be able to turn to the applications that we mentioned above.

1.2 Outline

The remainder of the text will go as follows. Section 2 will be dedicated to recalling commonly shared actuarial economic models: in section 2.1, we will discuss the economist's view of adverse selection; in section 2.2, we will discuss the actuary's view of adverse selection; in section 2.3, we will discuss moral hazard; and, in section 2.4, we will discuss the rationality of the pricing (and reserving) cycles. Section 3 will be dedicated to understanding how (many) good economic narratives work: in section 3.1, we will discuss the importance for story telling of being able to interpret human action; in section 3.2, we will discuss the inside view of the characters; as contrasted as the discussion of the outside view of the (implied) narrator we will have in section 3.3; in section 3.4, we will discuss features of fictional time; in section 3.5, we will discussed how narrative causality differs from both

---

[4] We will mostly use the commentary by (Dowling 2011). Another relevant commentary is (Discussion: Ricoeur on Narrative 1991).

sufficient reason and efficient cause; at that point, the reader may choose to skip to section 4 relating to a first set of applications of the work done so far and come back to section 3.6 after that; in section 3.6, we will explore how narratives serve as laboratories to identify unintended consequences; in section 3.7, we will discuss how good a story induces a (mini-)paradigm shift and how that may differ from the story being true; in section 3.8, we will explore how collective entities may properly enter economic stories; and, in section 3.9, we will mention how the mathematical language enters economic story telling. Section 4 will discuss the basic applications: in section 4.1, we will discuss how the past is not necessarily representative of the future; and, in section 4.2, we will discuss how regular succession can be quite different from causality. Section 5 will be dedicated to more advanced applications: in section 5.1, we will discuss how the statistical work of the actuary needs to be supplemented with story-telling; in section 5.2, we will discuss how a narrative approach to economic theorizing can assist the actuary in identifying model blind spots and assessing their importance; in section 5.3, we will use the framework built above to construct an alternative view of what a good rating system could look like; and, finally, in section 5.4, we will discuss why and how the actuary can get involved in the claims process.

2. ACTUARIAL STORIES

With a view of understanding how good economic theorizing functions, we propose to start by relating a couple of examples of economic models that commonly enter actuarial education and are commonly in the mind of many practitioners when going through their ratemaking exercises. The hope is that the features of good narratives that will be identified in section 3 will be readily apparent to the reader when considering how those commonly related stories were built.

2.1 The Economist's View of Adverse Selection

This particular story motivates why, even if the insurer was a monopolist, if the purchasing of coverage was not mandatory, it could happen that large sub-populations could be left out of the insurance system.

> Let's provide a typical example of adverse selection as the economic literature has considered it. The setting can be a life insurance pool in the 18th century (Bühlmann 1997). Suppose that there is initially only one insurer: basically it's the insurer for the county, and there is no competition yet. The insurer is not sophisticated yet, so the 'local actuary' sets the rate such that premium cover losses: effectively, the group starts with a pooled rate. Now, for random reasons (maybe the insureds are temporarily acting less than fully rationally, maybe because of the temporary influence of preferences other than that of terminal wealth, ...), a small subset of insureds decides to leave the pool. Since rating is basically projecting past trends into the future, the rate for the next year will decrease if the proportion of high-risk individuals that left the pool is greater than that of low-risk individuals. If that's the case, most of the people that left have no incentive to not come back in as the price for the same coverage has decreased. There's also no reason to believe that people attracted to buy in the pool will have a significantly different proportion of high-risk/low-risk than the proportions already in the pool. But,

what if, at the more likely extreme, only low-risk people are tempted to leave? Then, the rate for the next year will increase, and most low-risk people not already in the pool won't be tempted to get (back) in, but the high-risk people won't feel the disincentive and the pool will actually attract more high-risk people than the rate anticipated. The insurer will have to increase rates even further the period after. The same phenomenon can apply in many successive periods. In the limit, if the insurer hasn't filed for bankruptcy already, only high-risk individuals will be left and they will be purchasing the full amount of insurance at their own adequate rate. [Freely inspired by (Akerlof 1970)]

2.2 The Actuary's View of Adverse Selection

This particular story motivates why rate segmentation (based on cost differentials) is a sensible action to take in a competitive insurance market.

> Consider the situation in which a company (e.g., Simple Company) charges an average rate for all risks when other competing companies have implemented a rating variable that varies rates to recognize the differences in expected costs. In this case, Simple Company will attract and retain the higher-risk insureds and lose the lower-risk insureds to other competing companies where lower rates are available. This results in a distributional shift toward higher-risk insureds that makes Simple Company's previously "average" rate inadequate and causes the company to be unprofitable. Consequently, Simple Company must raise the average rate. The increase in the average rate will encourage more lower-risk insureds to switch to a competing company, which causes the revised average rate to be unprofitable. This downward spiral will continue until Simple Company improves their rate segmentation, becomes insolvent, or decides to narrow their focus solely to higher-risk insureds and raises rates accordingly. This process is referred to as adverse selection. (Werner and Modlin 2010, 151)

2.3 Moral Hazard

This particular story motivates why, as long as the actions that the insureds take to avoid or contain losses are not observable by the insurer, including risk sharing features (like deductibles, coinsurance or limits) in insurance contracts is a sensible action to take by the insurer, even if the insured is loss averse and the insurer is basically and relatively risk neutral.

> Assume that there is an insurer that approaches the market with insurance contracts. The insured, then, can accept one of the proposed contracts. Once coverage starts, the insured has a choice to follow two generic paths: (1) the insured can put in all the necessary efforts to avoid losses, or (2) the insured can stop putting in the necessary efforts to prevent losses. The insurer cannot become aware of whether or not the insured is taking the necessary steps to avoid losses. Clearly, if the insurer was able to become aware of the level of effort put in by the insured to avoid losses, the risk neutral insurer could rate accordingly and provide risk-averse insureds with contracts that they prefer and that protect them fully against insured losses. When the insured has more information than the insurer about the level of effort displayed to avoid losses, incentive compatibility forces the insurer to offer contracts where the insured faces some of the risk; for, imagine the insurer did not force the insured to face some of the risk, then the insured would receive the same coverage no matter what level of effort was put in to avoid losses, and the insured would rationally put in as low a level of costly effort as possible to avoid the loss and, thus, claim, on average, more. [Freely inspired by (Chavas 2004, 192)]

2.4 The Rationality of the Pricing (and Reserving) Cycle(s)

This story motivates why it could be the case that pricing cycles arise even in the absence of exogenous shocks to insurer capital, like large catastrophes.

Underwriting cycles, like profit fluctuations in other industries, reflect the interdependence of rival firms. Strong policyholder loyalty and demand inelasticity hold the allure of large returns for incumbent firms, but the apparent ease of entry into insurance, the lack of market concentration, and the difficulty of monitoring competitors' prices preclude excessive profits. The interaction of these forces keeps the market in disequilibrium, with continuing price oscillations. (Feldblum 1990, 175)

The rationality of the reserving cycle could follow in an environment where rates are regulated so that charged rates are based on the insurers' expected losses and expenses plus a set profit margin. In that case, insurers' reserves (which are presumably set within a margin of error that is often economically material), that do enter the expected losses, are one of the only ways left for the insurer to endogenously affect insurance prices. That being said, this does not rationalize the reserving cycle in a non-rate regulated environment.

3. HOW DO (MANY) GOOD ECONOMIC NARRATIVES WORK

The question of the nature and evaluation of theories in economic theory is a question that received a fair bit of attention in the wake of some successful and some disastrous economic interventions in the recent past: whether it be the Eastern transition from a control to market economy, the massive deregulation of banking and finance markets that many suspect to be at least a contributing cause for the recent Great Recession, or the massive auctions of the cellular 3G capacity in the USA and UK. In particular, (The Puzzle of Modern Economics: Science or Ideology 2010) written by Roger E. Backhouse addresses these issues. The first part of his work deals with fact finding relating to recent involvement of economists in public policy (Backhouse 2010, 15-96), while the 6th chapter "Creating a 'Scientific' Economics" (Backhouse 2010, 99-116) and 7th chapter "The Quest for a Rigorous Microeconomics" (Backhouse 2010, 117-136) deal with the nature of modeling in economics.

As a matter of course, Backhouse will be mainly interested in macroeconomic policy; however, for our purpose which is to relate to actuarial practice, we are much more interested in microeconomic policy-making. Therefore, we will follow a different path. Our path will instead take us much closer to understanding how economic theorizing draws upon our ability to follow a story, if only indirectly at times.

It is our thesis that (many) sound economic arguments are fundamentally 'good' imitations of an action. We are thus very close to the view proposed by Robert Sugden in (Credible Worlds: The Status of Theoretical Model in Economics 2008) and (*Credible Worlds, Capacities and Mechanisms* 2009).

At the heart of an 'imitation of an action', we find a text (either literally written or rendered verbally) which unfolds an imaginary world for the audience: we find

> (...) the idea that literary works are self-contained worlds with their own laws and their own logic (...). (…) The best way to understand *mimesis preaxeos*[5] (...) is to begin by freeing the concept of 'imitation' from any narrowly conceived comparison of art work and object, as in the physical resemblance between a marble bust and its subject. (Dowling 2011, 2)

We will spend the rest of the section attempting to highlight the nature of a 'good' 'imitation of an action'.

3.1 Interpreting Human Action

> From the point of view of the economist, the condition of human existence exhibit four fundamental characteristics. The ends are various. The time and the means for achieving these ends are limited and capable of alternative application. At the same time the ends have different importance. Here we are, sentient creatures with bundles of desires and aspirations, with masses of instinctive tendencies all urging us in different ways to action. But the time in which these tendencies can be expressed is limited. The external world does not offer full opportunities for their complete achievement. Life is short. (…) Our fellows have other objectives. Yet we can use our lives for doing different things, our materials and the services of others for achieving different objectives. (Robbins 2008, 74)

The entry point of a 'good' 'imitation of an action' is the building of believable characters. As the passage above is made to illustrate, part of what makes characters believable is that they suffer from the human condition. For example, in the passage above, the presumption is that the large diversity of sought ends contrasted with the limited available means and time to achieve those ends is a human universal: so much so that persons from all cultures, backgrounds, ages and times are expected to recognize it to be true of themselves and of the persons that surround them.

Notice how the above characterization abstracts from a lot of the minutiae of our lives. In theorizing generally, it is commonly sound and necessary to abstract from some features of the problem at hand and instead focus (sometimes to the point of caricaturing) on other features of the problem. In fact, it was the opinion of Milton Friedman that, in the process of abstraction at the heart of modeling, some of the 'assumptions' may be entirely unrealistic: to the point that, very often, the most significant and useful theories are built on 'assumptions' that are wildly unrealistic.[6] In fact, it is one of our hopes here to be able to provide guidance about which assumptions must be

---

[5] *mimesis praxoes* roughly translates to the action of imitating.
[6] "In so far as a theory can be said to have 'assumptions' at all, and in so far as their 'realism' can be judged independently of the validity of the predictions, the relation between the significance of a theory and the 'realism' of its 'assumptions' is almost the opposite of that suggested by the view under criticism. Truly important and significant hypotheses will be found to have 'assumptions' that are wildly inaccurate descriptive representations of reality, and, in general, the more significant the theory, the more unrealistic the assumptions (in this sense)." (Friedman 2008, 153)

realistic to drive the point of the economic narrative and which ones can be simplified, even to the point of caricature.

Despite Friedman's insistence that assumptions about economic agents need not be realistic, significant efforts have been deployed to provide economics with a less idealized view of economic agents. Herbert Simon has been a great proponent of that view. In his essay (A Behavioral Model of Rational Choice 1955), Simon was proposing that we attempt to refine our assumptions about how human beings deal with situations of economic interest: Simon wanted us to replace the (rational) 'economic man' of traditional economics with a view of human beings that have preferences that may appear inconsistent from some point of view, that use inappropriate rule-of-thumbs for assessing relevant probabilities instead of Bayes' theorem and other probability and statistics theorem, that have imperfect access to relevant information, etc. (Simon 1955, 99)

Yet, some story telling requires the author and the audience to be able to interpret human actions within a specific cultural context. Take a simple example of understanding why it might be the case that the owners of automobiles with make "Lexus", "Mercedes", "BMW" may have higher than average price elasticity with regard to their automobile insurance coverage. In that case, the actuary needs to use cultural awareness to understand that automobiles with these makes are signs of wealth and luxury. The actuary should also understand that people do become rich by 'becoming good at negotiating prices'. Thus, the actuary could form a working hypothesis that the price elasticity for automobile insurance may be higher than average for owners of cars with those makes.

Bottom line, it is important to build believable characters that react (to hypothetical situations) in a way that we recognize as plausible "because those same reasons [to explain why we undertake an action] are necessarily the means we use to explain to ourselves the actions of other people." (Dowling 2011, 4)

3.2 The Inside View of the Characters

> What makes history radically different from the physical sciences, Collingwood argues, is that historical events have an 'inside' - how the historical actor understood themselves and their actions - as well as an 'outside', meaning a subjection to external forces such as climate, geography, social institutions, and the like. Collingwood's 'outside' corresponds (...) to any social and physical environment independent of consciousness, and his 'inside' to the thoughts and motives of human agents. (...) [T]he inside includes a great deal more than rational calculation, as when unconscious desires or undeclared animosities become, along with conscious motives, a mainspring of action. (Dowling 2011, 56)

To understand why the inside views of the characters is relevant to economic storytelling, it is relevant to contrast the intentions behind historic storytelling, as described by Collingwood, to (economic) policy-supporting story-telling. The (presumed) intention that supports discourse relating to the science of history is to say something true about the past. Contrasted with this, the (presumed) intention of (economic) policy-making understood in its positive, as contrasted with its normative, sense is to say something true about what could possibly, plausibly or probably happen if a given set of policies is implemented. In a basic way, history is aimed at the past, while policy-making is aimed at the future. Caution is required here, because a large motivation for historical science undertaking is the understanding of the past so as to avoid its pitfalls or foster its successes (in the future). But it is also the case that people need to be able to imagine how their actions will appear, in retrospect, either to their future selves or to future generations, to be able to best guide their policies about what to do.

The key to why the inside view of the characters is relevant to (economic) story-telling is that, while they are in the story, the characters are literally unable to grasp the meaning of their actions as set against the background of the whole story. To them, the situations of the story are experienced in a state of imperfect knowledge (with informational blind spots, with some elements of available information not understood at all or well, with poor forecasts about how our future selves or generations will value our current actions, etc.).[7]

Without the inside view of the characters, one could be tempted to ask of the characters to act in synch with the ultimate lesson that a modeler may be attempting to draw out of an economic theory. But, for us, meaning and lessons are not available at the beginning of a story, but at the end. Therefore, to connect with the characters, it is important for the audience of the (economic) story to be able to traverse the story as if they were the characters in the story, with their state of imperfect knowledge and foresight.

3.3 The Outside View of the (Implied) Narrator

We will refer to the outside view in a way different from that of Collingwood as reported above. For us, the outside view is the dual of the inside view of the characters: if the inside view is the view of the story as it is experienced by the characters as they suffer the story and react accordingly, the

---

[7] "(...) a chain of causal implication that must be traversed in time, and in a state of partial or imperfect knowledge (...)An important point for Ricoeur is that any audience outside the horizon of the events in the story (...) must make this traversal in just the same state of imperfect knowledge as those inside it." (Dowling 2011, 8-9)

outside view corresponds to that of a narrator that has already grasped the events of the story as a whole and can thus highlight the lessons implied by the story.

In essence, we are saying that many economic narratives employ implicit omniscient narrators. Omniscient because it is quite common for the implied narrator of economic stories to be able to "say" exactly what the state of knowledge of the different characters is, what they prefer, what they are trying to do, etc. It is implicit because, as contrasted with, say, many modern novels, the narrator does not necessarily have an explicit voice within the (economic) story. For example, in the way that we presented the four actuarial stories in section 2, the narrator can be identified explicitly because the audience is being asked to consider a scenario. However, as we will comment on more later on, those stories could easily have been written almost completely with a mathematical language: as a set of assumptions and theorems. Indeed, some of those same stories from section 2 were initially related in their mathematical formulation. When that happens, the effective assumption is that the audience of the theory is able to translate the mathematical formulation into a story about people, their goals, their values, their hidden agendas, their beliefs, etc., facing situations and taking actions in accordance. When they do so, the audience is invoking a narrator that reports to them the interior discourse of the characters, the events as they truly happen without the characters being necessarily aware of them, etc.

We call this an outside view because it is a view from outside the time of the story, outside of the time that the characters would be experiencing. It is a view from which the globality of the story can be grasped. From this point of view, it is possible to extract lessons concerning the policies under consideration, because, outside of the time of the story, one can grasp together the intentions of the characters with the ultimate consequences of their chosen actions.

3.4 The Fictional Time of the Story

> We live in a world in which, not only are the things that we want scarce, but their exact occurrence is a matter of doubt and conjecture. In planning for the future we have to choose, not between certainties, but rather between a range of estimated probabilities. It is clear that the nature of this range itself may vary, and accordingly there must arise not only relative valuation of the different kinds of uncertainties between themselves, but also of different ranges of uncertainty similarly compared. From such concepts may be deduced many of the most complicated propositions of the theory of economic dynamics. (Robbins 2008, 79)

The experience of time has to be one of the fundamental human universal. Across civilizations, cultures, ages, times, etc., human beings have formed wisdom about their experience of the passage of time and recorded that wisdom in idioms of their language. The quote above is meant to illustrate that, for human beings, the passage of time is not the neutral and regular swing of the pendulum.

Time takes meaning in the worries, concerns, projects, etc. that we experience. Now, how does that translate in the stories that interest us?

3.4.1 Calendar Time in the Story

It is worthwhile to note that most cultures have a calendar: a calendar that helps them keep track of the movement of celestial bodies, of seasons, etc.; a calendar that starts at a meaningful event; a calendar which helps mark important intimate, private and public moments. Now, calendars are imminently public: if calendars were not public, they could not help people synchronize their activities. In fact, the importance of calendar time has increased dramatically with the advent of the written language, of written contracts, and, even more so, of the digital age. For example, in my own life, I keep track of pay-days that come every two weeks, of rents due at every beginning of the months, of utility bills due every month, of classes I need to teach that take place twice a week, and so on.

So, in an economic narrative, even when it is written in a mathematical language, the reference to time is not a reference to the neutral physical time of Newtonian mechanics. It is not even a reference to the warped physical time of general relativity. It is a reference to the humanly meaningful time of calendars and clocks. Admittedly, calendars are built with a view on the movements of the celestial bodies; but, calendars are meaningful to humans and not to the Sun, the Moon, ... Calendars also help us explain to ourselves the actions of other people: whether it be understanding why farmers are sowing seeds, understanding why people rush to the nearest mall to buy a pine tree, why a man spends a full morning making reservations at a local restaurant, etc. Thus, in a story, we can invoke calendar time to justify, rationalize, the actions of the characters. In effect, this is saying that calendar time lives just as much in stories as it does in our lived lives. That being said, the calendar time of the story is, by construction, fictive, while the calendar time of our lives is not. What is important is that the characters react to it in a way that we can comprehend.

3.4.2 Available Information

In natural languages, it is quite common for some verb tenses to be commonly associated with the telling of stories. Take the example of the opening of Star Wars: "A long time ago in a galaxy far, far away...." Given that this is the opening of a science-fiction movie (which would generally indicate a futuristic inclination for the movie), the invocation of the past may seem odd. In fact, the invocation of the past is more meant to indicate that a fictional story will be related. The use of the imperfect verb tense also often serves the same aim.

In effect, within natural languages, many verb tenses are used in narratives and the reason why there are many verb tenses used in narratives is to allow the narrator to reflect the position of the characters in the time of the story. Just like in real life, the characters look backward and forward in time. So, natural languages have provided us with ways to say that somebody was reflecting about the meaning of some events or was considering attempting an action or considering the consequences of said attempted action.

In economic narratives, as they are often either based on or expressible in terms of game theory, the notion of information set and of allowable actions serve substantially the same purpose. For example, in a (repeated) game theory setting[8], at any given point in the time of the game, a given player has access to some information relating to past and current steps in the game, but not to others; the player has properly understood the meaning of some pieces of information, but not others: thus allowing the player to form a view about the future; the player is also attempting to achieve some ends, often expressed in terms of attempting to maximize some preference function, and is allowed to take on some actions, but not others.

---

[8] See (Repeated Games and Reputations 2006) for an advanced introduction to repeated game theory. Another variant of game theory where dynamics are crucial is evolutionary game theory. For an introduction, see (Evolutionary Games and Equilibrium Selection 1997).

3.4.3 Disproportions and Meaningfulness

> What necessarily emerges from the disproportion [between time taken to narrate and time as it passed in the story] is a structure of significance. (...) [I]ts primary importance is that it represents a break or rupture with linear time, a transformation of Aristotle's cosmic time into a time of human preoccupation or concern. (…) At the level of events (...) characters move forward in a world where choices must be made with only approximate guess about their consequences, where accidents might occur at any moment to alter fortunes of the individual or the community, and where people must be judged on the shifting and uncertain ground of social appearances. (Dowling 2011, 47-48)

Consider again the stories that were relayed in section 2. The stories narrate the events centering around insurer and insureds decisions. Yet, presumably, these decisions take a tiny fraction of calendar time. But, they take up almost the entirety of narrated time, the time taken to go through the story. Why is that so? Compare with plain-vanilla dramatic movies, for example. In those movies, again, much of the movie will not cover many events that occur daily to the characters: like time spent on personal hygiene, time spent eating, etc. The movie-maker is instead choosing to focus the attention of the audience on the events that relate to the story at hand: events that are significant and meaningful in the story.

So, many economic narratives will focus the attention of the audience on particular (often recurrent) decisions that need to be taken by individuals, firms, governments, etc. In effect, the entirety of the time of narration is taken to describe those moments of pondering, decision, and action.

3.5 Narrative Causality *vs*. Efficient Cause *vs*. Sufficient Reason

> Ordinary life, Aristotle said, is most often made up of actions and events that take place in meaningless succession: 'one thing after another'. But narrative always involves, due to the logic of emplotment, a strong implication of causality: 'one thing because of another'. (...) [E]mplotment permits an intuitive grasping together (...) of otherwise heterogeneous elements (...). (Dowling 2011, 5)

At this point, we are ready to introduce the concept of narrative causality. In effect, causality is at the heart of narration because the audience needs to grasp why events occur in the sequence that they do: a story is not a (mere) temporal sequence of events.

But, what makes narrative causality different from other forms of causality that we commonly encounter: how is narrative causality different efficient cause[9]? How is narrative causality different from sufficient reason[10]?

Let's first address how narrative causality incorporates argumentation by sufficient reason. Because, in a well constructed narrative, the actions of the characters make sense taking into account their motivations and their environment as they face it; necessarily, characters have to have sufficient reason to take the actions that they actually do take within the story. It is worthy to note that some economists have thought that some brands of economic arguments rest entirely on sufficient reason[11] and, when expressing that thought, they were referring to the theory of valuation.

Still, why is it that argumentation by sufficient reason comes so naturally in economics? It is simply because economics is concerned with human choices. So doing, economic argumentation must extend itself beyond the material world and "involve links in the chain of causal explanation which are psychical, not physical, and which are, for that reason, not necessarily observable by behaviourist methods." (Robbins 2008, 85)[12]

By way of contrast, as Ricoeur noted (Ricoeur 1983, 249-255), the explanation of the evolution of human societies (especially in the historical science) must also make room for efficient cause: it must make room for change occurring in the natural environment of human beings. This is one way that efficient cause must enter narrative causality: by providing the rules by which the natural environment of the characters is evolving. For actuarial purposes, this is especially relevant. Whether the actuary is considering the impact of epidemics, the impact of weather and climate, the impact of

---

[9] "The 'efficient cause' of an object is equivalent to that which causes change and motion to start or stop (such as a painter painting a house) (see Aristotle, Physics II 3, 194b29). In many cases, this is simply the thing that brings something about. For example, in the case of a statue, it is the person chiseling away which transforms a block of marble into a statue." (Wikipedia n.d.)

[10] "Fourth Form: The Principle of Sufficient Reason of Acting (*principium rationis sufficientis agendi*); briefly known as the law of motivation.[11] 'Any judgment that does not follow its previously existing ground or reason' or any state that cannot be explained away as falling under the three previous headings 'must be produced by an act of will which has a motive.' As his proposition in 43 states, 'Motivation is causality seen from within.'[12]" (Wikipedia n.d.)

[11] "The distinctive character given to this system of theory by these postulates and by the point of view resulting from their acceptance may be summed up broadly and concisely in saying that the theory is confined to the ground of sufficient reason instead of proceeding on the ground of efficient cause." (Veblen 2008, 133)

[12] Here's another take at the same basic points: "The first, inescapable in any thinking about human conduct, is fundamentally the problem of the reality of choice, or 'freedom of the will'. It involves the essence of the value problem in the sense of individual values, and is at bottom the problem of the relation between individual man and nature. The second basic problem has to do with the relation between the individual man and society. The crucial fact in connection with the first problem is that, if motive or end in any form is granted any role in conduct, it cannot be that of a cause in the sense of causality in natural science. (...) Motive cannot be treated as a natural event. A fundamental contrast between cause and effect in nature and end and means in human behaviour if of the essence of the facts which set the problem of interpreting behaviour. There seems to be no possibility of making human problems real, without seeing in human activity an element of effort, contingency, and, most crucially, of error, which must for the same reasons be assumed to be absent from natural processes." (Knight 2008, 101)

ground movements, etc., the actuary is commonly faced with taking into account how the natural world is affecting the human environment.

But efficient cause also enters narrative causality because the characters must take into account physical, chemical, biological, etc. laws in figuring out what they can do and how they will go about effecting their choices. Now, in most economic narratives, as opposed to, say, crime novels, there is generally little emphasis on the exact ways in which the characters take actions. Still, for an economic narrative to be instructive, it is often the case that the narrative must reflect, even on an abstract level, the actual potentialities that humans have to effect their intentions.

As mentioned in the outline, readers eager to get to applications of the work done so far should skip to section 4 "Basic Applications" and come back to complete this section before moving on to section 5 "Advanced Applications ".

3.6 A Laboratory to Explore Unintended Consequences

> As one looks back on a completed series of events in a plot, it does seem as though there is something like unity or simultaneity in the causal chain. (...) [N]o, it was entirely unforeseeable; yes, we now see that it was inevitable after all. (…) Every story (...) is in an important sense told forward and backward (...). The forward movement, which belongs to what Ricoeur calls the syntagmatic order of discourse, links a movement from event X to event Y in an irreducibly temporal way. (...) At the same time, any continuous implication that the story has already been grasped as a whole (...) mean that events must be moving toward a conclusion so far unforeseen by its characters and by us its audience. (Dowling 2011, 9-10)

We are here getting at one fundamental reason why human beings tell each other stories: so that not everybody has to <u>actually</u> go through an experience to be able to draw the lessons from that experience. In effect, stories allow us to <u>simulate</u> reality and draw conclusions about what would really happen if we made certain choices or decisions. This is particularly valuable when the events, were they actually experienced, would induce dramatic consequences, e.g., irreparable damages, points-of-no-return, etc.

Given that we have argued that economics (and actuarial science also) are heavily grounded in policy-making, which is intrinsically forward-looking, it is quite natural to wish to be able to be able to comment about the <u>actual</u>, as <u>contrasted with intended</u>, effects of a policy, if it were implemented. And, our daily lives inform us that what actually takes place may well be far away from what we intended: sometimes because of poor execution on our part, but sometimes because people that share the world with us form intentions to interfere with our projects. Presumably, they do so because they choose to, if only at a subliminal level. And, their choice is the result from their own

projects, their own possibilities, their own preferences, their own conjectures, etc. And, this is where stories become useful simulations: because all the characters have to act in accordance with their own beliefs and motivations, the story as a whole may unfold in a way that does not match any of the characters' stated or hidden intentions. Thus, the story serves as a (virtual) laboratory that allows people to explore the unintended consequences of their choices and actions.

What are the stories from section 2 allowing us to experience without actually having to experience it? The story on the economist's view of adverse selection is meant to help us understand that it may be that all the people from a population may either gain or be no worse off from having a central power mandate insurance, when the distortions from adverse selection are too severe. The story of the actuary's view of adverse selection is meant to help us understand that, in a competitive insurance market, an insurer that does not invest in (cost and) rate segmentation may have its solvability threatened. The story about moral hazard is meant to illustrate that risk sharing is an invaluable feature of insurance contracts because, without it, claim inflation may make insurance unaffordable. And, the story about the pricing cycle is meant to illustrate that pricing cycles may not be the result of insurer irrationality and, so, suggest that education may not be the way out of the cycles. In all of the cases, the audience of the story is able to (virtually) experience the 'bad consequences' (unavailable coverage, insurer insolvency, claim inflation, continuing pricing cycles) without having necessarily to go through the experience themselves.

3.7 (Mini-)Paradigms Shifts vs. Truth

Are the stories of section 2 true? While the question may seem harmless, in fact, it is a source of great embarrassment[13]. If one meant by true that there is a correspondence between the events of the stories and events in the real (empirical) world, then a sensible answer may well be that the question is undecidable as, in the real (empirical world), the setups of the stories have never actually been encountered. If one meant that, if, in the real (empirical) world, we set up a situation that is like the initial conditions of the stories, then events would unfold as in the stories, then the question may still be undecidable, not as a matter of principle, but because it may be actually impossible to effect such setups. Such is the case because to construct those stories many circumstances from life have been abstracted away (and replaced by implicit and, yet, potentially not neutral assumptions).

---

[13] As the stories are not intended as stories about what actually happens (or, even for that matter, what <u>literally</u> will or would actually happen), the question asked is not actually that far from the question of asking whether a fictional work, like *War and Peace*, is actually true.

A potentially more promising avenue is to think about stories as being metaphorically true. But, then one has to come up with a view about what makes a metaphor true. A metaphor is meant to highlight how a certain aspect of something is like a certain aspect of some other thing. Is the likeness then true?[14]

The question then becomes whether or not the above question needs to receive a definite answer to allow us to make progress. Supposing we remain agnostic about the truth of fictional narratives, which would include the stories from section 2 as they were not stories meant to relay how (real) events actually unfolded, then it may still be sensible to ask if those stories were powerful, useful, ...

Here, we are pointing towards the idea that good economic narratives induce change in perspectives in the audience that allows them to better deal with their own lives. So, we say that good economic narratives are useful and that the channel for their usefulness is a change in perspectives from which events, data, choices, policies, etc. are considered. In effect, we are saying that good economic narratives induce (mini-)paradigms shifts in their audience.

In fact, all of the stories above were presented with the shift in vision-of-reality in mind: the stories were meant to induce (mini-)paradigms shift. And, of all of the presented stories, maybe it is the story initially presented by Akerlof that had the more lasting effect. In fact, Akerlof won the Nobel prize for that work[15].

3.8 The Place of Collective Entities

Examining the stories of section 2, we find that some of the characters are not, not even in principle, individuals: case in point, when insurers are brought in as characters of the story. What we have here is a case where a collective entity is a character in the story. But, then, how are we to make sense of assessments, motivations, intentions, etc. of collective entities? As the summary of the assessments, motivations, intentions, etc. of the composing individuals? Take the case of insurers. What is the objective of an insurer? To exist? To remain solvent? To provide insurance coverage at a fair price? To make profit? In fact, there are some groups of stakeholders of the insurer that would

---

[14] It is tempting to say that the likeness cannot be actually be true because truth has no degree while likeness does. And, when we say that truth has no degree, we are aware that we say things like 'half-truths': but, we say such things when the things told were literally true (and recognized to be such) but they were presented to be _interpreted_ in a misleading way.

[15] "Akerlof is perhaps best known for his article, 'The Market for Lemons: Quality Uncertainty and the Market Mechanism', published in Quarterly Journal of Economics in 1970, in which he identified certain severe problems that afflict markets characterized by asymmetrical information, the paper for which he was awarded the Nobel Prize." (Wikipedia n.d.)

focus on subsets of these possible answers. So, while it is certainly true that the way the insurer acts as a character in a story is a function of the individuals that exist in the sphere of the insurer, the insurer does not appear reducible to that of those individuals. Yet, collective entities commonly enter narratives: not just economic narratives, but fictional narratives, historic narratives, etc.

Let us examine a little bit more closely how collective entities enter historic narratives.

> What gives such terms [singular collectives, like 'Germany', 'Americans', etc.] meaning is that they refer to concepts - family, tribe, clan, nation - that are constitutive of individual consciousness, and which therefore have a real existence in what Husserl called the *Lebenswelt* or 'life world' of men and women. (…) Taken together, Ricoeur argues, these factors [civilizational forces] generate 'an ethics of participatory belonging,' meaning that individuals themselves as belonging to the group (...) but as sharing a common destiny. (...) This is the category of what Ricoeur will call first-order entities, meaning that they are directly rooted in the real life of men and women. (…) Any first-order entity, Ricoeur thinks, may legitimately be treated as a quasi-character (...) in historical narrative. (...) The point is that such entities have a semi-autonomous status in historical reality. (Dowling 2011, 66-67)

The key point for our purposes is that it is sensible to take collective entities as characters in a story because, as people, we are able to understand how these entities enter our lives. They enter our lives by our (potentially subliminal) choices. While it is true that actuaries, many of whom are employees or contractors of insurers, have their own individual agenda, they also understand themselves to act, in their capacities as employees or contractors, <u>for</u> the insurer. The same can be true of many stakeholder groups of the insurer, in many circumstances. In effect, the insurer can be properly treated as character in a story because, as a collective entity, it has much of the same characteristics of individuals. It has tendencies, it processes and assesses information, it has a diversity of aim, it has limited access to resources to achieve these aims, and it suffers from internal conflict. Thus, we think of collective entities in stories in a way analogous to the way we think of individual characters.

But, contrast this with a disconnected sub-population. Take a sub-portfolio of (homogenous) insureds in an insurer's book of business. There, it may well be sensible to think only in terms of representative individuals, because the composing individuals do not feel connected to each other: in their minds, they do not constitute a group.

3.9 The Place of the Mathematical Language

One clear cut distinctive feature of many economic narratives is their use of the mathematical language. In that way, they contrast themselves pretty clearly from most other forms of narratives (plays, novels, movies, etc.) where the mathematical language may be invoked incidentally, often through a character that is using mathematics. By contrast, in their published versions, economic

narratives may look more like a mathematical theory (with assumptions, lemmas and theorems) than a text like *A Midsummer Night's Dream*.

Hopefully, by now, it should be apparent that the mathematical language enters economic narratives as a secondary feature. The audience of the story is supposed to be able to take the abstractions expressed in mathematics and convert them back to characters with aspirations, views, etc., that are acting in an environment that they may be able to change.

In effect, we are saying that, even when they are expressed mathematically, economic models draw upon the capacity of its readers and users to be able to follow a story.

The remaining of section 3.9 is mainly addressed to a public that is familiar with economic theoretical work and can be safely skipped at a first reading. Example of readers that may find the remaining of the present section useful are those that are aware of the mathematical formulation of the lemons model of Akerlof or the underlying game-theoretic Industrial Organization works underlying the Feldblum article.

Now, even when the economic story is making serious allowance to the natural language narrative form, mathematics is often invoked in at least two sub-arguments: (1) in figuring out what would be best to do from the point of view of the characters and (2) in determining how the story would need to turn out so that no character would wish to change the course of action they are pursuing (when the story enters an equilibrium). These two uses of mathematics within economic narratives require specific attention.

3.9.1 Optimizing Behavior

The use of mathematics to determine the best possible reaction of a character given its motivation and beliefs requires specific attention because, in economic narratives, that use of mathematics is often associated with the two following sub-arguments: (1) the actual process used by the character to figure out what to do next is going to give exactly the same answer as the mathematically determined optimal course of action and, therefore, (2) there is no need to examine exactly how people actually select their course of action. In fact, the work that we exposed above suggests that both assumptions are problematic. Regarding (2), our work suggests that persuasive narratives allow the audience to understand why the characters are taking the actions they are taking: the audience needs the inside view. This is in direct contradiction with a proposal to not examine the way people choose a course of action. More importantly, if it is the case that people actually do end

up choosing actions that match what a mathematically constructed solution to an optimization problem would say they should do, the work done above suggests that this very fact requires an explanation in terms of the motivation of the characters, their understanding of the environment they are in, etc. This suggests that it may be worthwhile for economists to, also maybe, attempt to solve the problem at hand from the point of view of the characters in a narrative the same as the characters would tackle the problem. So doing, they would address both points.

3.9.2 Equilibrium Selection

Mathematics is often used in economic narratives to identify the conditions under which no characters would be motivated to change their course of action: thus, inducing an equilibrium. A commonly encountered problem is that there may be quite many possible 'final situations' in the story where the 'no-incentive-to-change' condition may be met. How, then, does the economist choose which of those final situations could actually be an outcome of the story? The work above suggests that it is in 'going through the story', many times, under many relevant initial situations, that the question may be best resolved. In effect, we are suggesting that equilibria of the story may be selected[16] using the narrative causality criterion presented in section 3.5.

4. BASIC APPLICATIONS

Given the large emphasis on mathematics in actuarial training and given that a lot of efforts of mathematics arose out of problems of the natural sciences, it is common for actuaries to be indoctrinated in the natural sciences approach to deduction and causality. Two characteristics of argumentation in the natural sciences come to mind here. One, argumentation in the natural sciences often uses arguments by covering laws, or the Deductive Nomological model. It can be expressed as:

1. Laws: $L_1, L_2, \ldots, L_n$
2. Initial Conditions: $C_1, C_2, \ldots, C_n$
3. Therefore, E (the *explanandum* or phenomenon of interest) (Reiss 2012, 32)

Two, argumentation in the natural sciences often invokes the following version of causality:

(a) X is universally associated with Y;

---

[16] Note that equilibrium is not the only possible outcome of a process. For example, as in Feldblum's model, equilibrium never arises. When that happens, mathematical structures are often convenient to characterize the outcome: does it lead to periodic behavior, does it have an attractor, is there a time scale where the solution stabilizes? Even if an equilibrium occurs, it is not necessarily obvious that only one type of equilibrium will arise for very close parameters of the problem: in effect, is the model subject to bifurcations?

(b) Y follows X in time;
(c) X and Y are spatio-temporally contiguous (there are no time-wise or space-wise gaps in between X and Y). (Reiss 2012, 104)

With the work that has been done so far, it should be apparent that economic narratives do not function in these ways. Given that we have argued that the narrative as whole serves as the argument, it should be clear that no more than bits and pieces of the narrative can employ the Deductive Nomological model of argumentation; and, often, it will be used to allow the audience to keep track of the effects of efficient causes in the story. Also, given that we have argued that narrative causality is at work in stories, we can see that a different notion of causality is at work. (1) In stories, there is enough place left for contingency and luck for effects not to be universally associated with their causes. (2) In stories, things like simultaneous and, even, reverse causation are common. Take, for example, the case of the story that rationalizes the pricing cycle. In that case, the insurers are taking actions based on their beliefs about other insurers, and their own action induces the other insurers to act in the predicted way. Here, the phenomenon of self-confirming beliefs is an example of contemporary causation. Reverse causation arises naturally when actions are taken in a forward-looking context: e.g., a character does an action at one point in time because of a belief about what will happen in the future: the future drives the past. (3) Action at a distance is quite common in narratives: in effect, the narrative is constructed around the propagation of a cause to its effects. Next, we will inquire about the effects of these inappropriate ways of arguing on actuarial work.

4.1 The Past Is Not Necessarily Representative of the Future

> (...) [N]on-autonomous relations are not lawlike; they do not represent the underlying causal ordering. (...) [C]ausal ordering is a property of models that is invariant with respect to interventions within the model and structural equations are equations that correspond to the specified possibilities of intervention. (Hoover, Econometrics as Observation: The Lucas Critique and the Nature of Econometric Inference 2008, 301-302)

Here is a common mistake made by actuaries when they do not use the right concept of causality in their work: they wrongfully assume that past trends will carry forward in the future. Take, for example, loss trends. It is not an uncommon actuarial assumption to use the same loss trends for both the past as for the future. Even when they are not the same, future trends are not uncommonly selected as a carry-forward of recent trends. Now, imagine that those trends are quite high: that is, loss costs are increasing quite fast. If we were to then imagine a narrative that incorporates insureds, insurers and governments, it is not unreasonable to imagine a scenario where the insurers feel pressure to charge the loss cost trends in the premium, where insureds then pressure the government for premium reduction, and the government may have to intervene to set up measures

to contain the loss cost trend. In effect, in the story, the feedback is making it such that the past loss cost trend is unlikely to continue unabated. In effect, we are saying that regular relations are not necessarily causal and, if they are not causal, it is quite possible for the regularity to disappear quite rapidly. This suggests that, when proper causes have not been identified, it is prudent to set up regular and high frequency monitoring to validate the regularities that are being exploited in practice are still there. It is also means that, for medium- or long-term forecasting, the identification of the appropriate causal mechanisms is an integral part of a successful forecasting process.

4.2 The Nose of the Donkey Does Not Cause Its Tail

> There was a man who sat each day looking out through a narrow vertical opening where a single board had been removed from a wooden fence. Each day a wild ass of the desert passed outside the fence and across the narrow opening — first the nose, then the head, the forelegs, the long brown back, the hindlegs, and lastly the tail. One day the man leaped to his feet with a light of discovery in his eyes and he shouted for all who could hear him: "It is obvious! The nose causes the tail!"
> Stories of the Hidden Wisdom from the Oral History of Rakis (Herbert 1984, 359)

Another common actuarial mistake is to (sometimes implicitly) assume that what comes first in a regular succession must be the cause of a phenomenon. Take the following example. Imagine a homogeneous subportfolio of insureds that are relatively insensitive to rate increases. It could be that this is a class of insureds that values dearly their time and thus attach a high perceived cost to shopping for insurance: they do not shop much for insurance. The effect of their reduced willingness to shop will be to (1) increase their retention and (2) increase their profitability. Now, assume (price) elasticity modeling is done: by examining the impact on retention of a change in price. What the modeling actuary will find is that increased retention and increased profitability, both observed first, leads to lower price elasticity (observed second because it is a result of the price change). But, the true causal channel was exactly the opposite: reduced price elasticity leads to increased retention and increased profitability. Thus, it is not necessarily the case that what comes first in the observation of a regular succession is the cause of what follows.

5. ADVANCED APPLICATIONS

Finally, we can now address more involved applications of the work done above. These applications bear on the future of the training of actuaries and they bear on the future of actuarial practice.

5.1 Model Calibration and Story Telling

The first (advanced) application of our work relates to the way statistical work should be thought and conducted within the actuarial profession. One way to put the point is to say that actuarial statistical practice needs to become more like econometrics and less like statistics. That is, it is a consequence of our work that calibration (even probabilistically- or statistically-minded calibration) of models supporting policy-making needs to heavily relate to the narrative that support them. Why? Because, then, the modeling actuary is likely to fall prey to the two fatal flaws that were mentioned in section 4: inappropriately assuming that the past will carry forward in the future and inappropriately (potentially implicitly) assuming that what is observed first is the cause of what comes after.

Another way to cast the point is as follows. One of the difficulties with purely statistical strains of reasoning (e.g., predictive modeling) is that it remains agnostic about the nature of the underlying mechanisms at work. And, because of this agnostic inclination, the efforts of the modeler are then re-directed towards what is possible for the modeler: for example, the modeler then focuses on in-sample quality of fit statistics, out-of-sample model performance, information criteria, etc. Again, because of the agnosticism about causes, little if any efforts are put into understanding how come the observed results have come about. And, if actuarial science were statistics, this would be no great loss. However, actuaries do no work in a vacuum: actuaries need to <u>convince</u> stakeholders of the insurer that the measures that they are proposing based on the estimation/calibration work is justified, sensible, prudent, etc. Thus, actuaries need to be able to weave together their statistical work and the economic theorizing work. And, their education should reflect that requirement and not be excessively oriented towards predictive modeling.

Also, as was also mentioned in the introduction, actuaries that go through the process of theorizing about the results they obtain may be less likely to fall prey to over-fitting of the data: because they may question some of the raw results they obtain and drop some relationships or some variables from their models when the fitted relationship does not appear sensible based on their theorizing.

A (disciplined) process of storytelling about the observed relationships in the data that can be quite useful is highlighting incoherent observations coming from the model: when the actuary begins to rationalize the observations emanating from the model, some relationships may begin to appear contradictory, thus providing an opportunity to selectively revisit the model.

5.2 Understanding Model Blind Spots

Here, we intend to walk the readers through two examples of understanding model blind spots: (1) we will go through the example of three potential sources of inelasticity to insurance prices and how they lead to radically different strategies and (2) we will go through an Enterprise Risk Management exercise to understand how model blind spots imply that much caution is necessary in the use of model results. Our hope here is to highlight how (economic) theorizing can be crucial for the modeling actuary in the interpretation and application of the found results.

Let us start with the price elasticity exercise. Suppose a modeling actuary goes through an exercise like that explained in (Beyond the Cost Model: Understanding Price Elasticity and Its Applications 2013). Then, suppose that the modeling actuary identifies that the age with insurer, that is the number of years since the original inception date of the policy, is a critical variable in driving down the price elasticity of demand (for the insurance provided by the insurer). An almost immediate application of that result in premium optimization would be to attempt to increase rates (potentially moderately) for the segment of insureds that have been with the insurer a long time. Now, with further theorizing, the modeling actuary could come up with, at least, three big working hypotheses for why that segment of insureds is so price inelastic: (1) the insureds in the segment are price inelastic because the insurer has properly and better identified a large cost differential that allows the insurer to price so attractively in the segment that a mild variation in price does not lead to a material fluctuation in the buying ratio, (2) the insureds in that segment experience large search costs (e.g., they value greatly the time spent on researching a competitive offer, they loathe going through the 20 questions round with brokers and insurers to receive alternative quotes, they fear that the potentially negative perception that could arise by having many parties accessing information like credit scores, etc.), or (3) the insureds in that segment value loyalty (maybe in the hope that the length and the strength of the relationship with the insurer may induce a more understanding attitude on the part of the insurer in the event of a claim). These differing working hypotheses lead to different 'side-predictions' about the appropriate course of action to take in the segment. Start with the example of large search costs. This could lead the actuary to see if the rating algorithm could not be better integrated with a web on-line tool or if the algorithm could be simplified without materially affecting rate adequacy. If it were a matter of loyalty, the insurer should be seeking to strengthen the relationship with the insured as much as possible: this could lead to some concerted efforts with agents and/or brokers to meet or communicate more actively with the client. If the effect was only due to price, then all of the extra efforts (of the options laid out above) could well be wasted and should not be attempted at all. Thus, the identification of price elasticity leaves the

actuary with a puzzle leading to further questions and modeling[17]; but, it also means that prudence is the better part of valor in the use and application of model results.

Now, let us examine an Enterprise Risk Management application. Take the framework as laid out in (Actuarial Geometry 2006). There Steve Mildenhall is attempting to build a framework for (insurer insurance operations) risk assessment that constitutes an improvement on some work of the prior generation that followed too closely the finance literature without taking into consideration the particular nature of insurance risk: e.g., non-transferrable contracts, no way to take a multiple or fractional position on a contract, etc. Our efforts thus far, though, do allow us to identify the material model blind spot of the enterprise as he lays it out: there is no theory of prices underlying the work. There is no theory of valuation. There is no theory of strategic interactions. And, here is one place where this lack of taking into account of strategic interaction becomes important. Go back to the Feldblum story about the rationality of the pricing (and reserving) cycle that we mentioned above. In that case, fluctuations in the industry Loss Ratio (thus of most insurers) are induced by strategic interactions; and, those Loss Ratio fluctuations are leading to material insurance risk, even to the point of being a material source of insolvency risk. In this case, the model blind spot can be identified: leading to prudence in model application. That being said, as opposed to the prior example, here the remedy to the blind spots of the model are not easy to implement; but, at least, the actuary can go through an exercise in imagination of thinking about which ways the model results could be biased because of the model blind spots.

5.3 An Alternative View of a Good Rating System

Let us look back at our story on the actuarial view of adverse selection. We concluded that that story was warning insurers that rate segmentation was crucial to their continued solvency. But, is that the final word on the story? What does the moral of the story become when we revisit section 3.1; what happens when we try to better understand what the insureds want from the underwriting process? Sure, insureds want available coverage. Sure, they want it cheap. Sure, they want to be able to make sense of why the price is so high. Yet, people now constantly feel pressed for time. There is a definite sense that we are bombarded with information, that we must constantly be available to

---

[17] For example, if loyalty was at the core of the observed price inelasticity, then increasing the number of (positive) contacts with the insured should lead to increased buying due to a price elasticity that decreases even further; while, if search costs was at the heart of the observed price inelasticity, increased communication (especially about further offers) may make the insured more price elastic as the insured may be force to experience some of the negative effects of those search costs. Under a loyalty causal channel, increased communication can be valuable to the insured; under a search cost causal channel, the insured values minimizing shopping activities and would prefer to be left alone as much as possible. Thus, different causal channels lead to different side-predictions that can be tested.

examine and process information, ... Time feels like it is speeding up (Birnbaum and al. 2012). And, in that context, it is no wonder that insureds put a premium on not having to answer 20 questions about every minutia that an actuary may believe could be relevant to assess insurance risk. So, what can actuaries do to address that aspiration? For one, actuaries can look to answer some of the questions that they would like answered on their own. How? By peering through public records. For example, works on building have to be recorded with municipal entities. Another example: why ask an insured for neighboring exposure and why not instead do a Google search of the insured's address. The idea is for insurers to make sure that, when they go to the insured to answer a question, it is because <u>only</u> the insured could have answered the question. For two, actuaries can examine more closely their predictive modeling output and reflect on whether or not the extra variables and very many variables that they would like to use for ratemaking <u>materially</u> contribute to (either) the insurance risk (or, the assessment of the price elasticity) of the insured. What do we mean by materially? We say that a factor is material if one's decision would change had the factor changed. So, when an extra piece of information would affect the premium by a 'pocket money' amount, is it really necessary to have the variable in the rating algorithm? In effect, we are saying that <u>simplicity</u> is also valuable; but, more importantly, we are saying that rating algorithm simplicity is becoming more valuable in this time when people feel pressed for time.

5.4 Claims Are Not Just Numbers

> The Rockfeller Corporation studied why customers defect and found the following: (...)
> 68% The customer believes you <u>do not care</u> about him or her. (Baker 2006, 163)

Another insight that can be gleaned from the work above is that a good modeler of human content must constantly make the effort to connect back with the underlying <u>human</u> reality that is being modeled. How tempting for actuaries to produce reports about Loss Ratios, retention, closing ratio, etc. examining the influence of variable *X, Y, Z* and writing a quick comment to a superior about the found relationships. Now, try the following. Think back to a time you needed to work with your insurer. You just suffered a fire. You just got into a car accident. Your possessions just got stolen. ... Now, think about how time felt right there and then. I would wager that, in some ways, time slowed down: in the sense that many concerns just took a back seat to that event. At the same time, time may have felt to speed up because, before you know it, you needed to be rushed to a hospital, or ... The point being that immediate priorities take over routine. And, this is a time of profound human vulnerability. It is a time when being treated like a person, with dignity, is profoundly valued. Actuaries need to remember that persons are at the root of the numbers they work with. And, maybe, a good way for actuaries to be reminded of that message is to get involved

in the administration of claims. Why not get actuaries involved in the modeling of the claims hotline queue? Why not get actuaries involved in attempting to understand when a claim is likely to get big enough to warrant a more senior adjuster to assist in the settlement process? In effect, why not let actuaries use their acquired skills in modeling to assist the people that will be working directly with the insureds in their time of need?

6. CONCLUSION

We would like to finish this essay with a bit of a more general reflection. We have argued that causation in the natural world is dramatically different from that in the human world. But, maybe, this is not so completely true. Recently, Ilya Prigogine has argued that the natural sciences need to make room for 'dramatic changes of behavior', even in deterministic systems. More generally, he was advocating a re-thinking of the laws of mechanics (whether classical or quantum) in the terms of evolution of probabilities. (Prigogine 1994, 51) Now, it appears the way that narratives function is not so completely different: in the sense that the events of a story are connected by <u>probability</u> and not certainty.[18]

Moreover, we do believe that mathematics may help in the formalizing of the modeling scheme described above. We suspect that logic could be expanded to account for an evaluation of whether the events in a story are appropriately connected. That may require logicians to formalize the notion of coherence. Perhaps that will require a probabilistic and fuzzy intentional time-tensed dynamic logic.

Acknowledgment

4. REFERENCES

[1]   Akerlof, George A. "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism." *Quarterly Journal of Economics* 84, no. 3 (August 1970): 488-500.
[2]   Angrist, Joshua D., and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton, New Jersey: Princeton University Press, 2009.
[3]   Backhouse, Roger E. *The Puzzle of Modern Economics: Science or Ideology.* New York, NY: Cambridge University Press, 2010.
[4]   Baker, Ronald J. *Pricing on Purpose: Creating and Capturing Value.* Hoboken, New Jersey: John Wiley & Sons, 2006.
[5]   Birnbaum, Jean, and et al. *Où est passé le temps.* Paris: Éditions Gallimard, 2012.
[6]   Buchanan, James M., and Viktor J. Vanberg. "The Market as a Creative Process." In *The Philosophy of Economics: An Anthology*, by Daniel M. Hausman, 378-398. New York, NY: Cambridge University Press, 2008.

---

[18] See (The Market as a Creative Process 2008) for another take for why modeling in natural and human sciences may be becoming more alike.

[7]   Bühlmann, Hans. "The Actuary: the Role and Limitations of the Profession Since the Mid- 19th Century." *ASTIN Bulletin – The Journal of the International Actuarial Association* 27, no. 2 (November 1997): 165-171.

[8]   Carr, David, Charles Taylor, and Paul Ricoeur. "Discussion: Ricoeur on Narrative." In *On Paul Ricoeur*, by David Wood, 160-187. New York, NY: Routledge, 1991.

[9]   Chavas, Jean-Paul. *Risk Analysis in Theory and Practice*. San Diego: Elsevier Academic Press, 2004.

[10]   Dowling, William C. *Ricoeur on Time and Narrative*. Notre Dame, Indiana: University of Notre Dame Press, 2011.

[11]   Feldblum, Sholom. "Underwriting Cycles and Business Strategy." Edited by Casualty Actuarial Society. *Forum*, Spring 1990.

[12]   Frees, Edward W. *Regression Modeling with Actuarial and Financial Applications*. New York, NY: Cambridge University Press, 2010.

[13]   Friedman, Milton. "The Methodology of Positive Economics." In *The Philosophy of Economics: An Anthology*, by Daniel M. Hausman, 145-178. New York, NY: Cambridge University Press, 2008.

[14]   Guven, Serhat, and Michael McPhail. "Beyond the Cost Model: Understanding Price Elasticity and Its Applications." *Casualty Actuarial Society e-Forum*, Spring 2013: 29.

[15]   Herbert, Frank. *Heretics of Dune*. New York, New York: The Berkley Publishing Group, 1984.

[16]   Hoover, Kevin D. "Econometrics as Observation: The Lucas Critique and the Nature of Econometric Inference." In *The Philosophy of Economics: An Anthology*, by Daniel M. Hausman, 297-314. New York, NY: Cambridge University Press, 2008.

[17]   Hsiao, Cheng. *Analysis of Panel Data*. 2nd. New York, NY: Cambridge University Press, 2003.

[18]   Klugman, Stuart A., Harry H. Panjer, and Gordon E. Willmot. *Loss Models: From Data to Decisions*. New York, NY: John Wiley and Sons, 1998.

[19]   Knight, Frank. "Economics and Human Action." In *The Philosophy of Economics: An Anthology*, by Daniel M. Hausman, 100-107. New York, NY: Cambridge University Press, 2008.

[20]   Mailath, George J., and Larry Samuelson. *Repeated Games and Reputations*. Oxford: Oxford University Press, 2006.

[21]   Mildenhall, Stephen P. "Actuarial Geometry." 2006.

[22]   Prigogine, Ilya. *Les lois du chaos*. Paris: Flammarion, 1994.

[23]   Reiss, Julian. *The Philosophy of Economics*. New York, NY: Routledge, 2012.

[24]   Ricoeur, Paul. *Temps et récit(Tome I: L'intrigue et le récit historique)*. Paris: Éditions du Seuil, 1983.

[25]   —. *Temps et récit(Tome II: La configuration dans le récit de fiction)*. Paris: Éditions du Seuil, 1984.

[26]   —. *Temps et récit(Tome III: Le temps raconté)*. Paris: Éditions du Seuil, 1985.

[27]   Robbins, Lionel. "The Nature and Significance of Economic Science." In *The Philosophy of Economics: An Anthology*, by Daniel M. Hausman, 73-99. New York, NY: Cambridge University Press, 2008.

[28]   Samuelson, Larry. *Evolutionary Games and Equilibrium Selection*. Cambridge, Massachusetts: The MIT Press, 1997.

[29]   Simon, Herbert. "A Behavioral Model of Rational Choice." *Quarterly Journal of Economics* 69, no. 1 (February 1955): 99-118.

[30]   Sugden, Robert. "Credible Worlds, Capacities and Mechanisms." *Erkenn* 70 (2009): 3-27.

[31]   Sugden, Robert. "Credible Worlds: The Status of Theoretical Model in Economics." In *The Philosophy of Economics: An Anthology*, by Daniel M. Hausman, 476-510. New York, NY: Cambridge University Press, 2008.

[32]   Veblen, Thorstein. "The Limitations of Marginal Utility." In *The Philosophy of Economics: An Anthology*, by Daniel M. Hausman, 129-142. New York, NY: Cambridge University Press, 2008.

[33]   Werner, Geoff, and Claudine Modlin. *Basic Ratemaking*. Casualty Actuarial Society, 2010.

[34]   Wikipedia. *Four Causes*. http://en.wikipedia.org/wiki/Efficient_cause#Efficient_cause (accessed 02 23, 2013).

[35]   —. *George Akerlof*. http://en.wikipedia.org/wiki/George_Akerlof (accessed 03 03, 2013).

[36]   —. *Principle of sufficient reason*. http://en.wikipedia.org/wiki/Sufficient_reason (accessed 02 23, 2013).

Biography of the Author

Marc-André Desrosiers is a project manager for the Research and Development team at Intact Financial Corporation. He has completed a Bachelor degree in philosophy and one in actuarial science at Concordia University, Montréal. He also completed his MBA at University of Calgary. He is interested in all aspects of ratemaking.