Performance Testing

A Control Cycle Approach to Managing Reserve Risk

2010 CLRS Concurrent Session ST-7
Stephen Lowe and Yi Jing

September 21, 2010

TOWERS WATSON

Today's agenda

- Defining the problem
- Performance testing and the actuarial control cycle
- Case studies — real-world results

- This presentation is based on the paper

  Claim Reserving: Performance Testing and the Control Cycle
  - by Yi Jing, Joseph Lebens, and Stephen Lowe
  - Published in *Variance* (2009 V3 I2), available at www.variancejournal.org

Defining the problem

Questions for the reserving actuary

- How do you know that the methods you are currently using are the "best"?
  - What evidence supports your selection of methods?
  - What are the right weights for combining the results of the methods?
  - How do you decide when to change methods?
  - What is the confidence range around estimates from each method?
  - How do you evaluate the cost/benefit of developing new input data sources or implementing more complex methods?
- How do you measure and manage reserve risk?
  - How do you avoid overconfidence in the work of "unseasoned" actuaries?

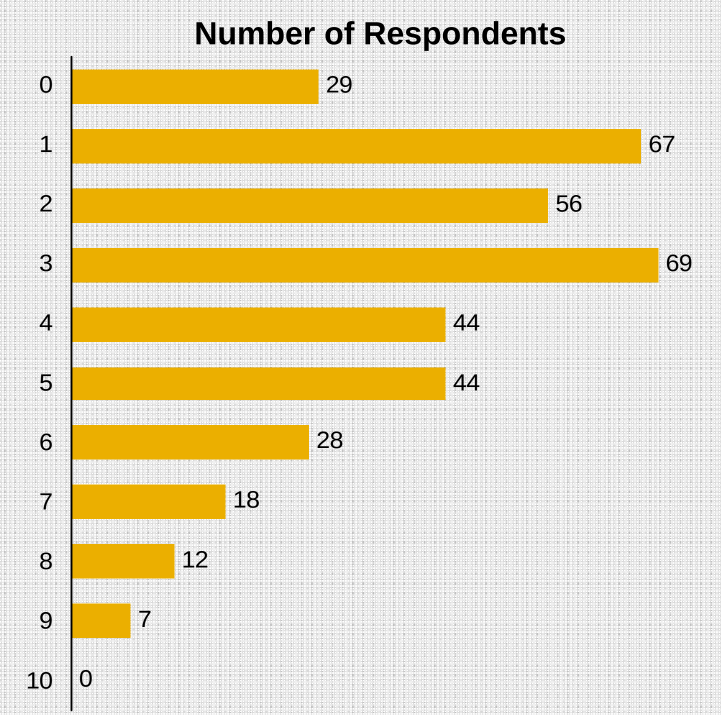
The results of our research illustrate the prevalence of actuarial overconfidence

2004 Confidence Quiz in *Emphasis*

The Quiz

- Objective: To test respondents understanding of the **limits** of their knowledge
- Respondents were asked to answer ten questions related to their general knowledge of the global property/casualty industry
- For each answer, respondents were asked to provide a range that offered a 90% confidence interval that they would answer correctly
- Ideally (i.e., if "well calibrated"), respondents should have gotten nine out of ten questions correct

Raw Scores of Respondents



Note: Based on 374 respondents as of 4/5/04. Profile of respondents: 86% work in P/C industry; 73% are actuaries.

Performance testing

Embedding reserve risk management

Reserves are forecasts!

- An actuarial method is used to produce a forecast of future claim payments
- An actuarial method consists of
  - An algorithm
  - A data set
  - A set of intervention points

$$\hat{L}_m^{(t)} = m(a, d, p)$$

$$L^{(t)} = m(a, d, p) + \varepsilon_m$$

- The actuary must
  1. Choose a finite set of methods $\{m_1, m_2, \ldots m_n\}$ from the universe M
  2. Choose a set of weights $\{w_1, w_2, \ldots w_n\}$ to combine the results of each method together
- Performance testing, via a formal control cycle, can help the actuary make these choices in a rigorous manner

Formally testing alternative methods yields some interesting and counterintuitive results

- Sometimes projecting case reserves is the best method
- Methods that use claim counts and averages outperform
- Methods that formally adjust for changing claim settlement rates or changing case reserve adequacy can produce better estimates
- The degree of correlation between methods is an important consideration in selecting methods, and weights used to combine them
- Hindsight errors are larger than those predicted by some stochastic methods

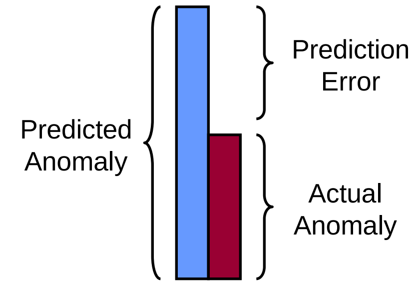
An aside: Case outstanding development

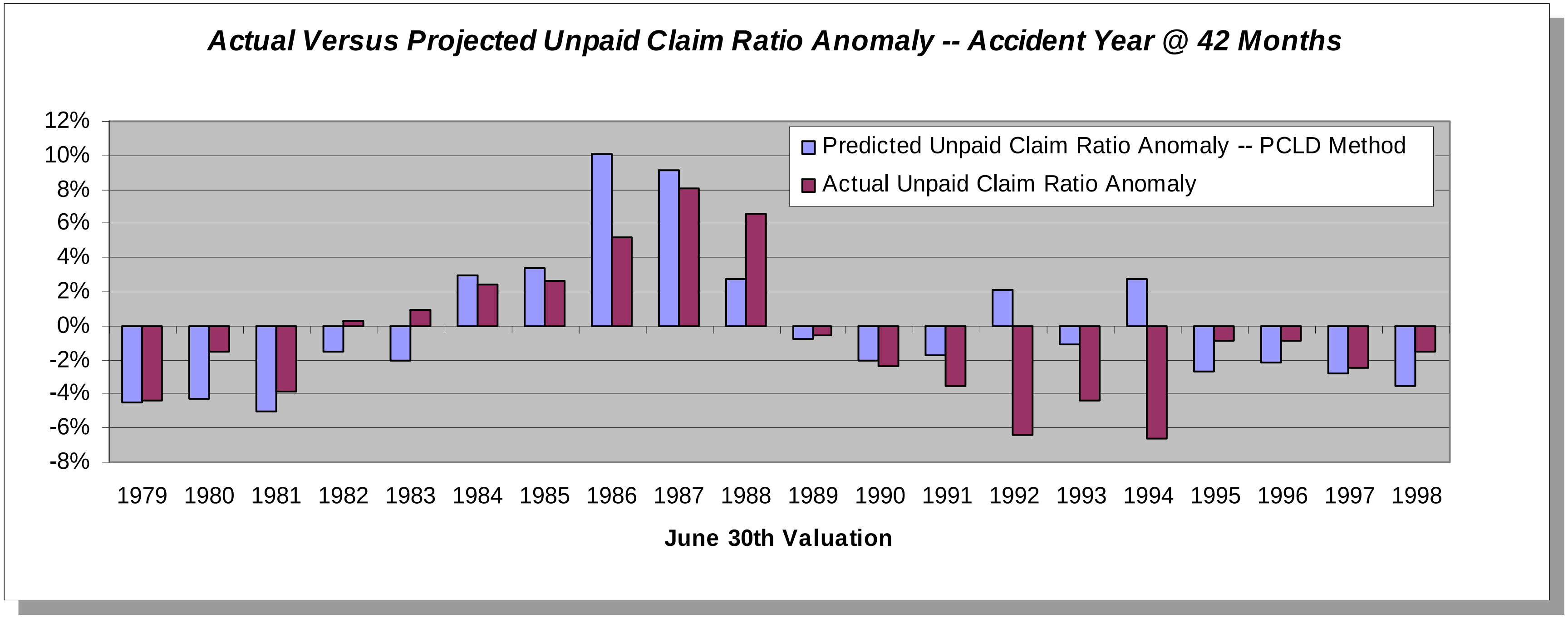
- Case reserve development factors inferred from selected paid and reported development factors

	12-24	24-36	36-48	48-60	60-72	72-84	84-96	96-108	108-120
Paid ATA Developmen Factors	4.000	2.000	1.650	1.350	1.180	1.080	1.030	1.010	1.000
Cumulative Development Factors	23.625	5.906	2.953	1.790	1.326	1.124	1.040	1.010	1.000
Percent Unpaid	95.8%	83.1%	66.1%	44.1%	24.6%	11.0%	3.9%	1.0%	0.0%
Reported ATA Developmen Factors	1.960	1.380	1.240	1.150	1.070	1.024	1.009	1.003	1.000
Cumulative Development Factors	4.277	2.182	1.581	1.275	1.109	1.036	1.012	1.003	1.000
Percent Unreported	76.6%	54.2%	36.8%	21.6%	9.8%	3.5%	1.2%	0.3%	0.0%
Percent in Case Reserves	19.1%	28.9%	29.4%	22.5%	14.8%	7.5%	2.7%	0.7%	0.0%
Case Reserve Development Factor	5.001	2.875	2.251	1.957	1.665	1.468	1.443	1.433	nm

Performance testing yields a formal measure of skill

- The skill of a method is measured by: $Skill_m = 1 - mse_m / msa$
  - mse = mean squared error
  - msa = mean squared anomaly
- Skill is the proportion of variance “explained” by the method



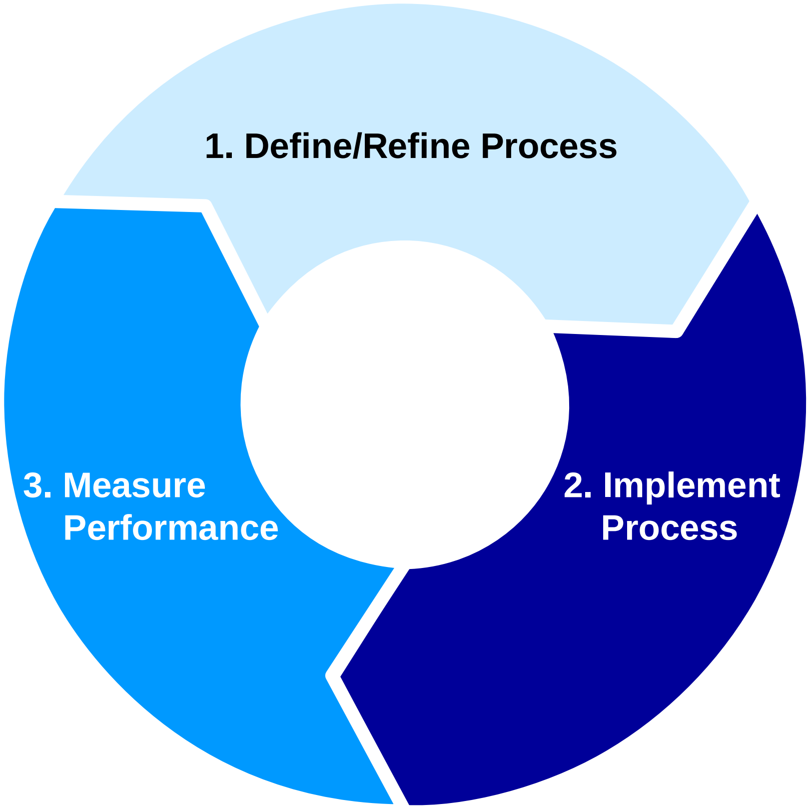


Performance testing of reserving methods can be part of an institutionalized control cycle

The Actuarial Control Cycle for the Reserving Process
Embedding Reserve Risk Management

Formal Performance Testing

- Are the current methods appropriate? Would changes to methods improve estimation skill?
- Are the data and other input accurate and sufficient? Would improvements or expansion of data improve estimation skill?
- Are there opportunities to improve process flow?
- Are emerging estimation errors within tolerances?



Reserving Process Elements

- Data used
- Actuarial methods employed
- Operational input
- Judgments and intervention points
- Process flow and timeline
- Quality assurance process

Case studies

Real-world examples (not in the published paper)

Case Studies

- US Personal Auto Liability
  1. Skill of chain-ladder methods
  2. Selecting optimal weighting between methods
  3. Validating a stochastic reserving model
- Selecting development factors
  4. US Personal Auto Liability
  5. US Other Liability Occurrence

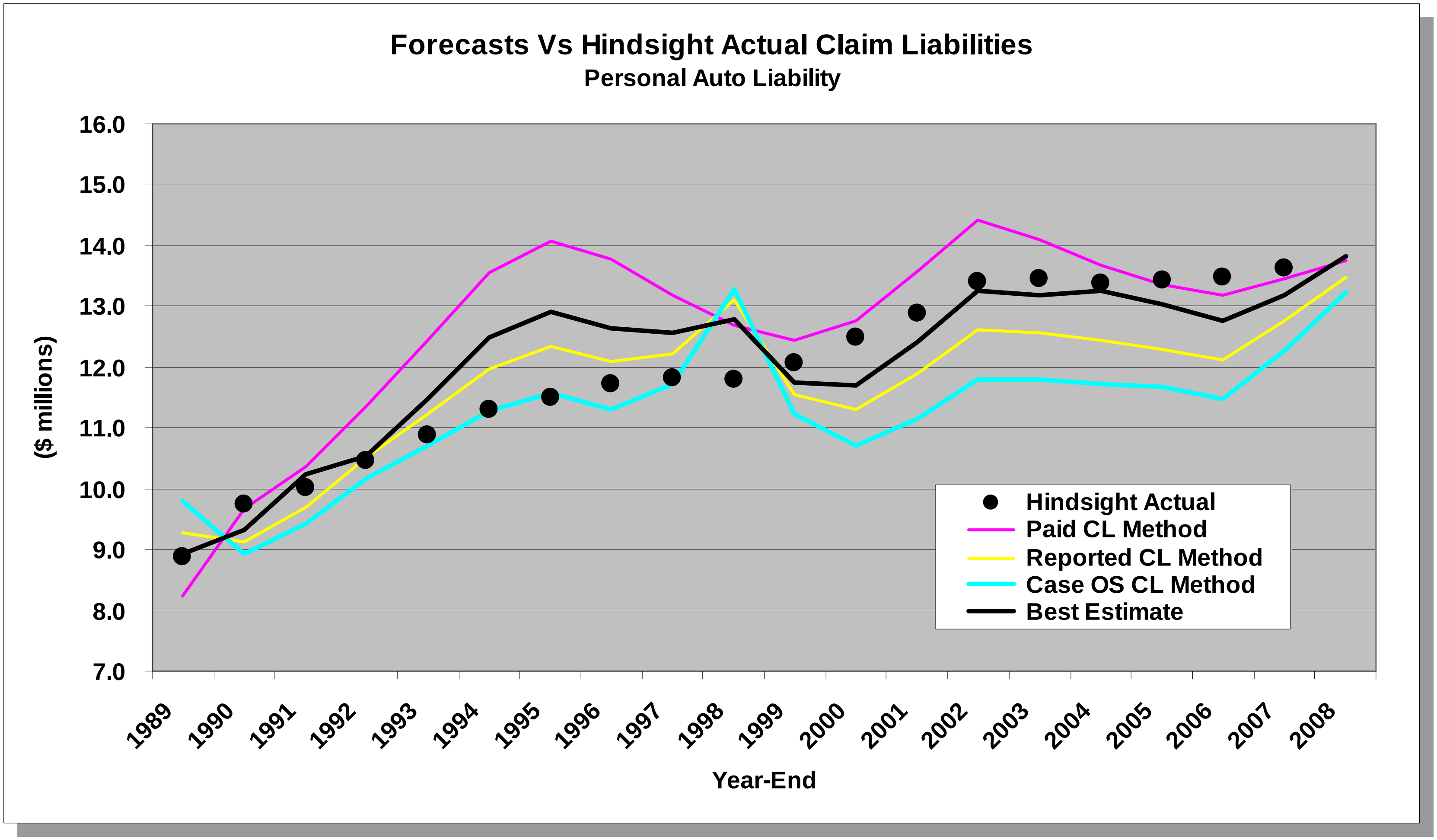
State Farm – Personal Auto Liability – Schedule P Data

Paid Claim Development Data (in $ millions)

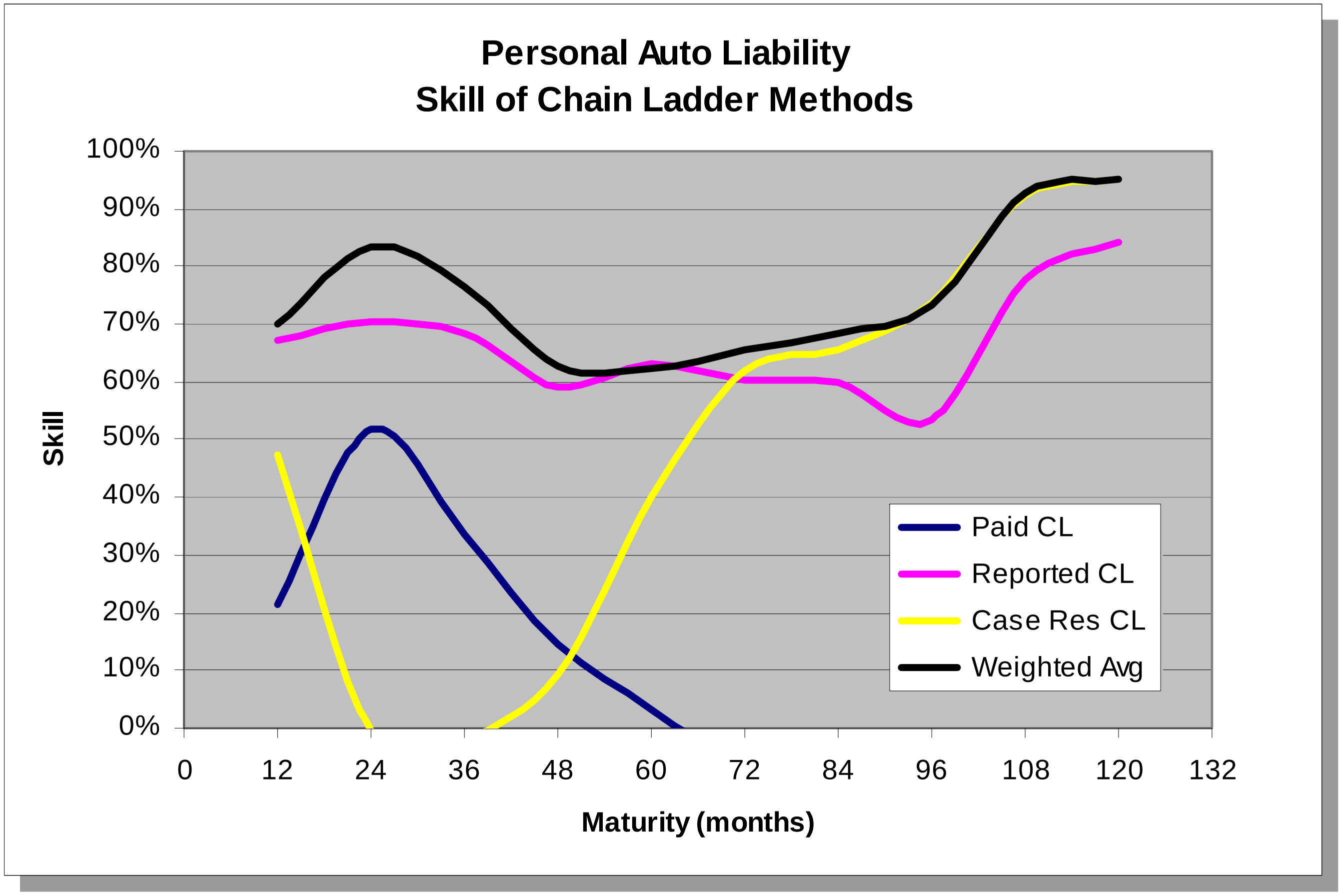
Accident Year	1	2	3	4	5	6	7	8	9	10	11+
1983	1,240	978	424	220	110	61	32	20	11	7	15
1984	1,437	1,164	523	269	143	80	44	27	15	8	18
1985	1,647	1,384	618	355	184	92	54	27	13	8	13
1986										8	13
1987										9	13
1988										7	9
1989										10	17
1990										12	19
1991										9	24
1992										10	18
1993										13	27
1994										14	42
1995										15	33
1996										14	35
1997										16	41
1998										17	36
1999										22	
2000											
2001											
2002											
2003											
2004	5,234	3,215	1,385	876	485						
2005	5,168	3,171	1,433	863							
2006	5,174	3,213	1,453								
2007	5,365	3,421									
2008	5,465										

- Historical estimates were made at nineteen prior year-ends, and compared with actual run-off to measure skill
- Four methods were tested
  - Paid chain ladder ($-weighted latest eight)
  - Reported chain ladder ($-weighted latest eight)
  - Case outstanding chain ladder (inferred case development derived from payment and reporting patterns)
  - Weighted average of above three methods, using indicated optimal weights

Summary of performance test results over nineteen-year hindsight test period



Observed skill varies by method and maturity



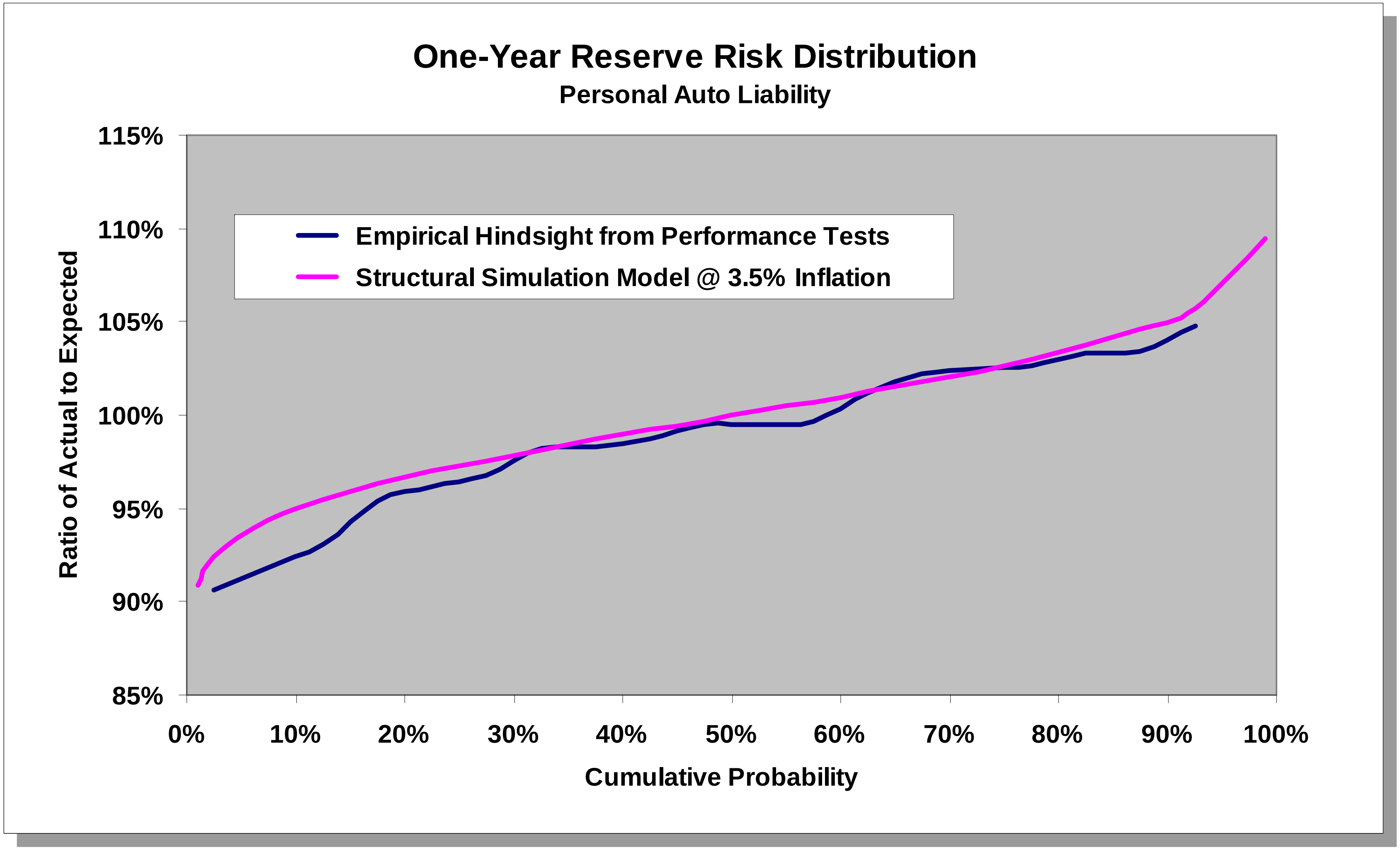
- Note that skill can be negative (e.g., paid method at 36 months), implying that the method induces volatility rather than explaining it

Indicated optimal weights by maturity reflect variances and correlations of errors

@ 24 months	Paid	Reported	Case OS		Std Dev		Weights
Paid CL	100%	33%	-6%		1.84%		.321
Reported CL		100%	92%		1.44%		.679
Case OS CL			100%		2.65%		.000

@ 84 months	Paid	Reported	Case OS		Std Dev		Weights
Paid CL	100%	85%	28%		.23%		.000
Reported CL		100%	74%		.13%		.349
Case OS CL			100%		.12%		.651

Results can be used to validate stochastic reserving models



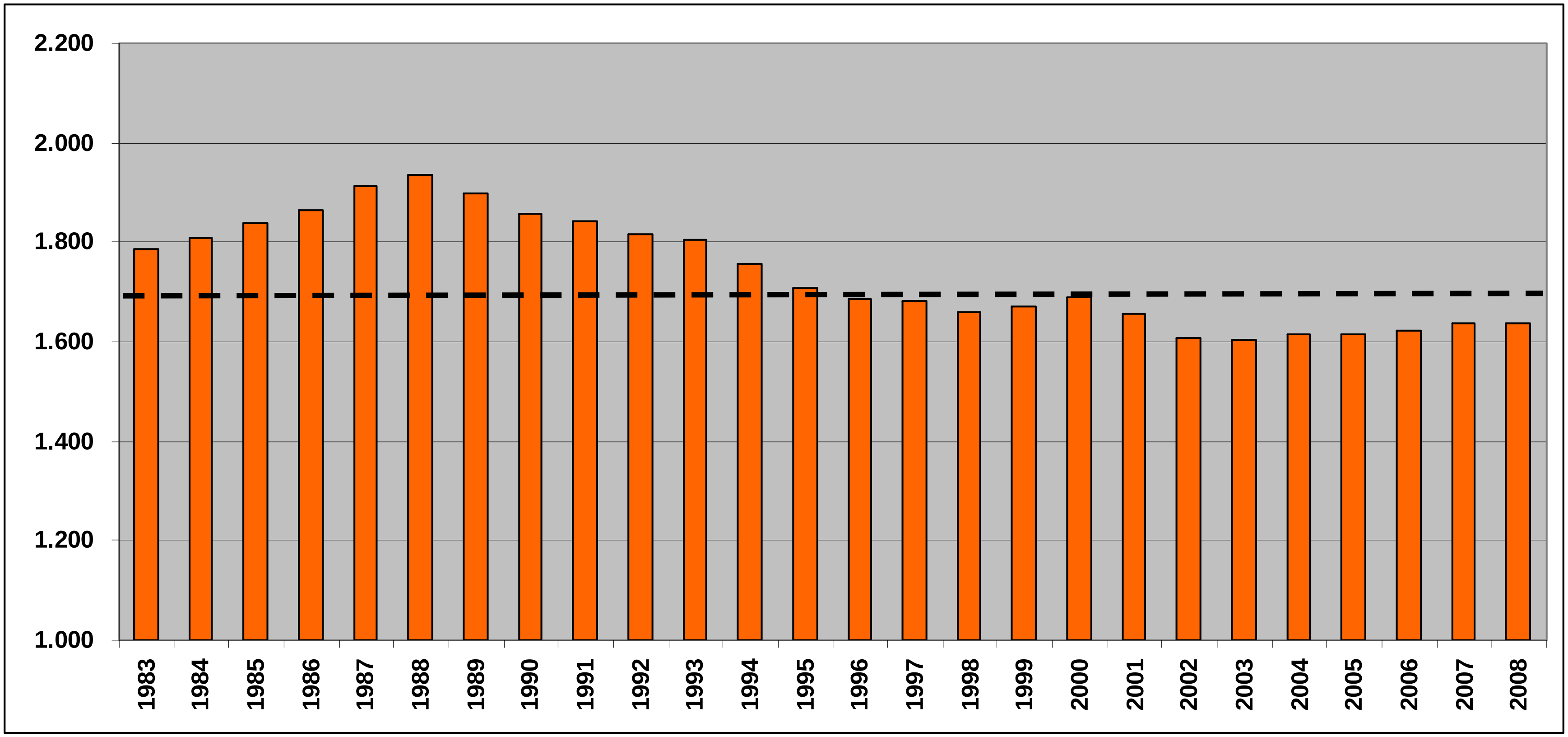
Methods for selecting age to age (ATA) loss development factors

Simple Average	$\overline{ATA} = \frac{\sum_{i=1}^{n} ATA_i}{n}$
Maximum Likelihood	$MLE = e^{\hat{\mu}+\hat{\sigma}^2/2}$, *assume lognormal distribution* $\hat{\mu} = \frac{\sum_{i=1}^{n} \log(ATA_i)}{n}$ $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (\log(ATA_i) - \hat{\mu})^2}{n}$
Volume-Weighted Average	$\overline{ATA} = \frac{\sum_{i=1}^{n} L_{2,i}}{\sum_{i=1}^{n} L_{1,i}}$
Latest Observation	$\overline{ATA} = ATA_n$

Which method is best?
What is the best value of n?

US Personal Auto Liability

- Actual paid ATA development factors from 12 to 24 months

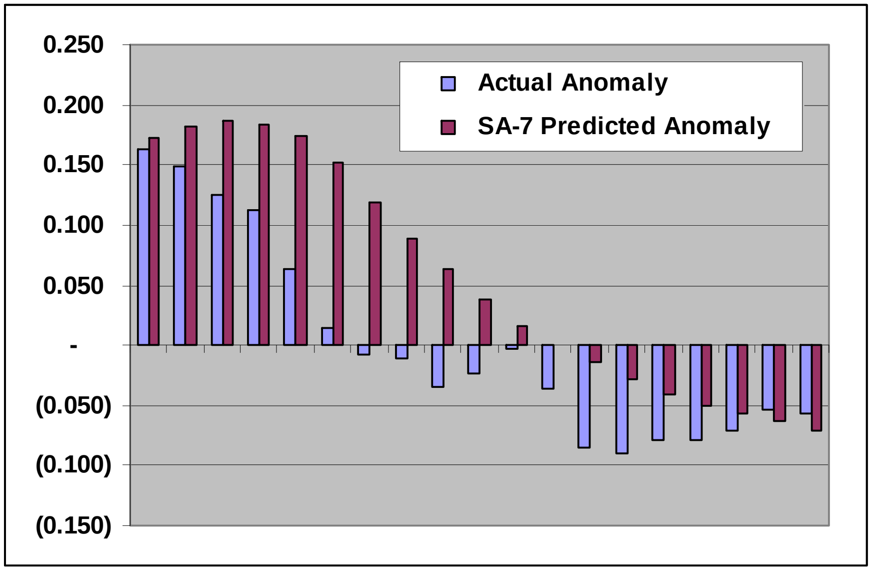


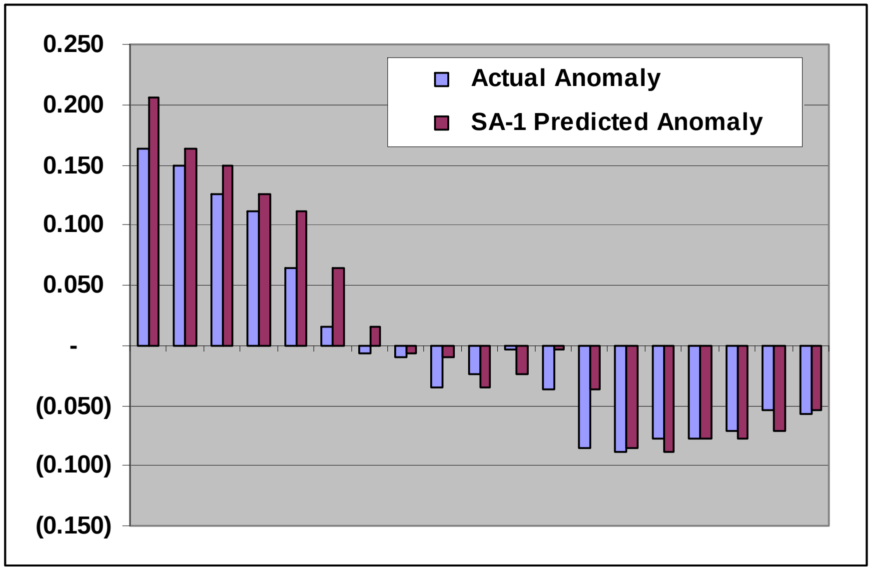
Sample calculations of ATA factor predictive skill

Accident	12 to 24 ATA Factors			Anomalies			Errors	
Year	Actual	WA-7	MLE-7	Actual	WA-7	MLE-7	WA-7	MLE-7
1990	1.856	1.876	1.864	0.164	0.184	0.172	0.020	0.008
1991	1.841	1.879	1.874	0.149	0.187	0.182	0.038	0.033
1992	1.818	1.878	1.879	0.126	0.186	0.187	0.060	0.061
1993	1.804	1.870	1.875	0.112	0.178	0.183	0.066	0.071
1994	1.756	1.858	1.866	0.064	0.166	0.174	0.102	0.110
1995	1.707	1.834	1.844	0.015	0.142	0.152	0.127	0.137
1996	1.685	1.802	1.811	(0.007)	0.110	0.119	0.117	0.126
1997	1.682	1.773	1.781	(0.010)	0.081	0.089	0.091	0.099
1998	1.658	1.749	1.756	(0.034)	0.057	0.064	0.091	0.098
1999	1.669	1.726	1.730	(0.023)	0.034	0.038	0.057	0.061
2000	1.689	1.707	1.708	(0.003)	0.015	0.016	0.018	0.019
2001	1.656	1.692	1.692	(0.036)	(0.000)	(0.000)	0.036	0.036
2002	1.607	1.677	1.678	(0.085)	(0.015)	(0.014)	0.070	0.071
2003	1.603	1.661	1.664	(0.089)	(0.031)	(0.028)	0.058	0.061
2004	1.614	1.649	1.652	(0.078)	(0.043)	(0.040)	0.035	0.038
2005	1.614	1.640	1.642	(0.078)	(0.052)	(0.050)	0.026	0.028
2006	1.621	1.634	1.636	(0.071)	(0.058)	(0.056)	0.013	0.015
2007	1.638	1.628	1.629	(0.054)	(0.064)	(0.063)	(0.010)	(0.009)
2008	1.636	1.622	1.622	(0.056)	(0.070)	(0.070)	(0.014)	(0.014)
Average =	1.692	1.745	1.748					
	Bias =	3.1%	3.3%					
			MSA =	0.006		**MSE =**	0.004	0.005
						Skill =	33.5%	24.7%

Lack of volatility, coupled with trend in ATA factors, causes long-term averages to have low skill

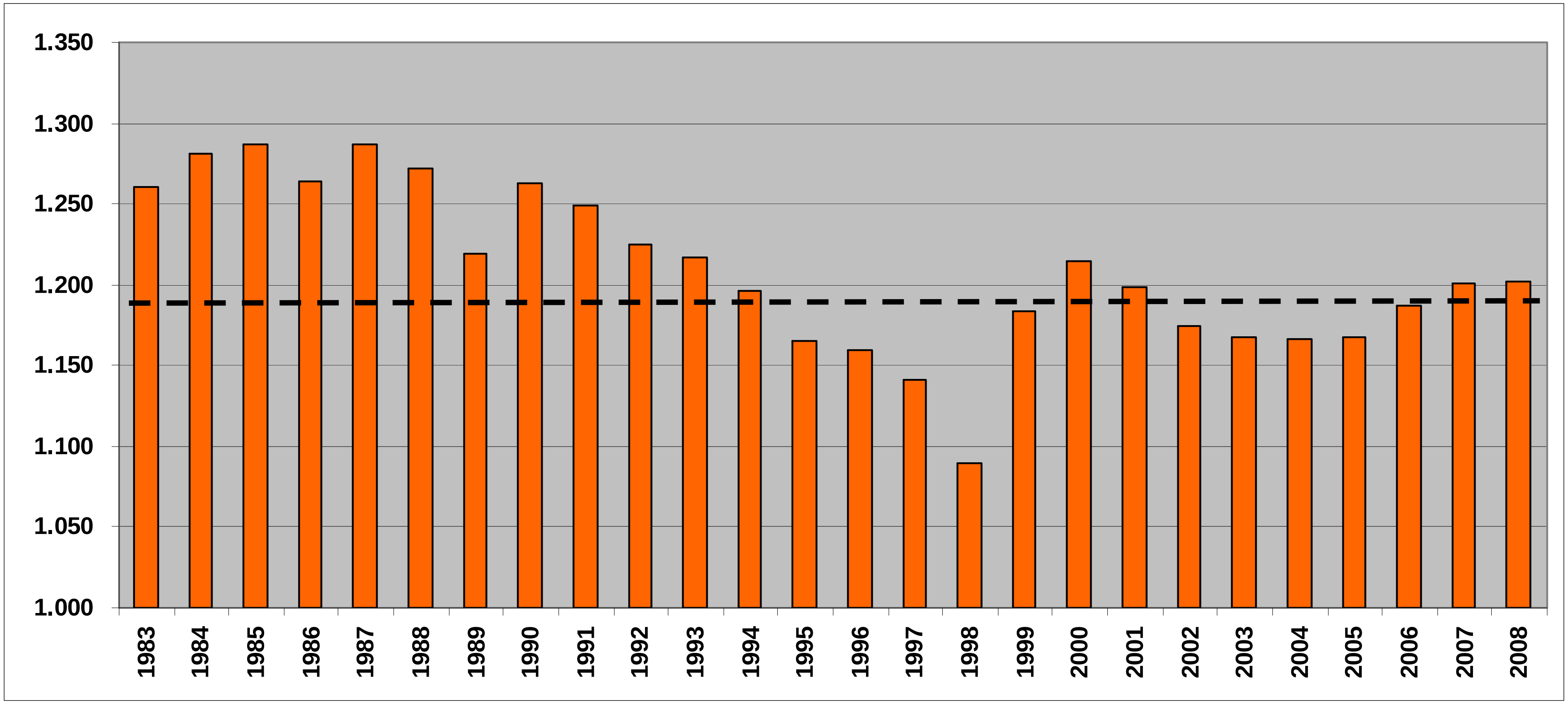
- Simple average of latest 7 factors is slow to respond to trend in factors
  - Predictive skill of simple average of latest 7 is 24%; very poor fit to pattern of anomalies
  - Predictive skill of simply using latest 1 observation is 90%; most of variation is explained





US Personal Auto Liability

- Actual reported ATA development factors from 12 to 24 months



Summary of measured skill for ATA selection methods

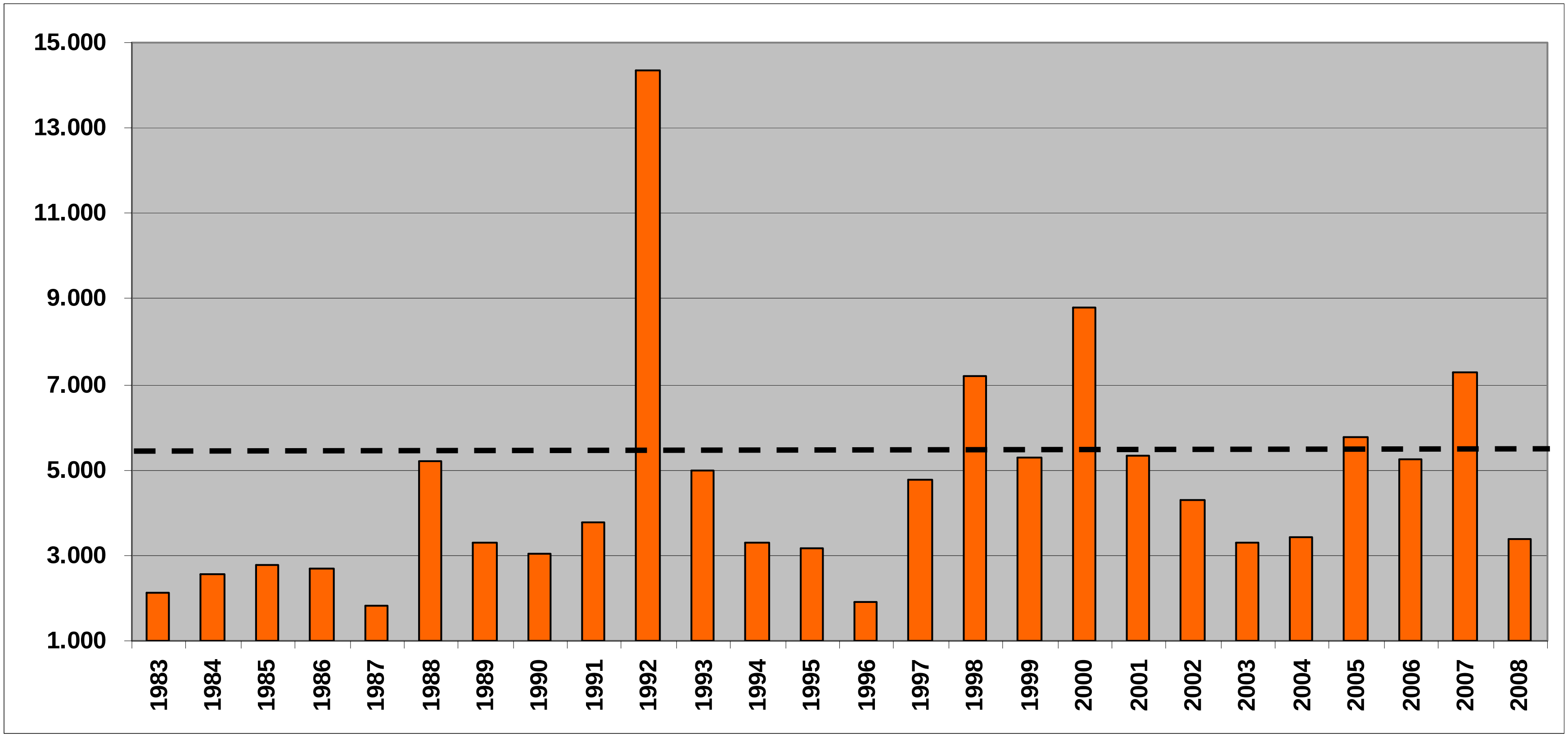
US Personal Auto Liability
12 to 24 months paid and reported development factors

ATA Selection Method	Paid Skill	Reported Skill
Simple Average – Latest 1	89.5%	33.4%
Simple Average – Latest 2	79.8%	20.9%
Simple Average – Latest 3	70.2%	12.5%
Simple Average – Latest 7	24.4%	-25.8%
Weighted Average – Latest 7	33.5%	-15.6%
Maximum Likelihood – Latest 7	24.7%	-25.8%

When using 7 observations, weighted average has *highest* skill, better than MLE

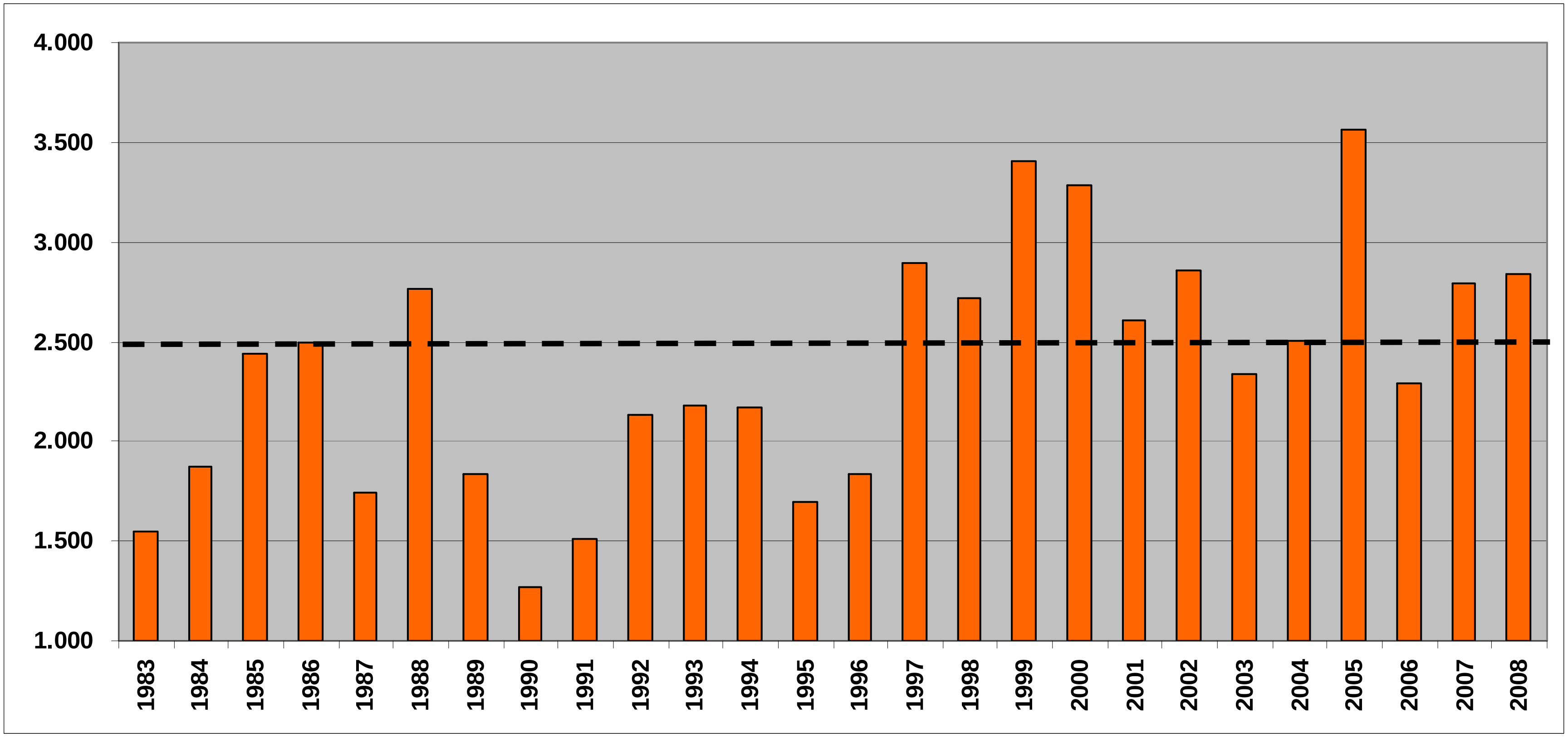
US Other Liability Occurrence

- Actual paid ATA development factors from 12 to 24 months



US Other Liability Occurrence

- Actual reported ATA development factors from 12 to 24 months



Summary of measured skill for ATA selection methods

US Other Liability Occurrence
12 to 24 months paid and reported development factors

ATA Selection Method	Paid Skill	Reported Skill
Simple Average – Latest 3	-63.7%	14.7%
Simple Average – Latest 6	-37.7%	14.0%
Simple Average – Latest 7	-29.9%	11.2%
Weighted Average – Latest 7	-35.2%	9.9%
Maximum Likelihood – Latest 7	-28.8%	11.1%

When using 7 observations, weighted average has *lowest* skill, MLE about the same as simple average

Conclusion and Q&A

Good reasons to do performance testing

1. Opportunity to improve accuracy of estimates
2. Formal rationale for selected actuarial methods
3. Input to development of reserve ranges
4. Cost / benefit of enhancements to data and systems
5. Supports Solvency II / Economic Capital
   - Embeds reserve risk management
   - Empirical validation of stochastic reserve risk models
6. Manage actuarial overconfidence