How Confident is the Information Ratio?

Christopher Lewis | September 2022

Read Time: 20 mins

Key Takeaways

1. Information Ratios are directly linked to the statistics concept of confidence levels.
2. For the same Information Ratio, a longer track record provides greater confidence that the expected, long-run relative performance of a manager (or model) is positive.
3. Over longer time horizons, the magnitude of Information Ratio needed to reach a particular confidence level or peer ranking (e.g. top quartile, top decile, etc.) tends to shrink.

Summary

The Information Ratio metric of investment performance is closely related to the broader, multi-disciplinary concept of statistical confidence. The former scales a manger's relative performance within a past data sample, but a very simple adjustment can translate that measure into a hypothesis test of whether expected relative returns (i.e. in the long run) are positive. Using Information Ratio in this context is more forward-looking, and offers a method to compare managers (or models) with different lengths of data history. All else being equal, a longer track record behind the same Information Ratio provides greater confidence that the past performance was not a fluke.

Background

Information Ratio ("IR") has become a widely adopted performance metric in the investment management community. Dividing a manager's average past relative return by the volatility of those relative returns provides a useful perspective on efficiency. Any deviation from a passive benchmark incurs some amount of tracking error, so Information Ratio essentially measures the "bang for the buck" of active return per unit of active risk.

However, sorting managers (or models, backtests, etc.) by their IRs and selecting the highest one would be naïve for a few reasons. First, the inherent volatility of asset returns means that attractive-looking performance can be caused by temporary, random luck. Even the top-ranked IR might not have any repeatable skill behind it. Secondly, the same Information Ratio number can convey different levels of confidence: if two managers have the same IR but one has a longer track record, the more tenured manager presents more (statistical) confidence that their expected relative return is positive.

The second point is the subject of this paper. It seems obvious that longer track records are desirable, but how exactly does one weigh the tradeoff between past performance and track record length? From the standpoint of statistical confidence, would it be preferable to hire a manager with an IR of 0.75 and five years of live history, or one with a 0.433 IR but 15 years of history behind it?

Translation to Confidence Levels

Information Ratios to t-statistics

Fortunately, the definition of Information Ratio is already very similar to a metric that can calculate this tradeoff: the t-statistic (Goodwin, 1998). A t-statistic (often shortened as just "t-stat") measures how far away observed data is from a base case scenario which would be unremarkable.

The adjustment from IR to t-statistic is quite simple: multiply the observed, annual IR by the square root of the years of data behind it (the appendix shows the proof for this). Returning to the two theoretical managers introduced above, both have a t-statistic of 1.68, which indicates the same level of confidence that their previous outperformance did not come from lucky streaks.

$$\textit{Example manager t-stats: } IR\sqrt{N} = 0.433\sqrt{15} = 0.75\sqrt{5} = 1.68$$

Naturally there would be other, deeper due diligence questions to ask too (e.g. after 15 years, is the one manager nearing retirement?), but this adjustment from IR to t-statistic provides a simple way to appropriately compare Information Ratios for different time horizons.

t-statistics to Confidence Levels

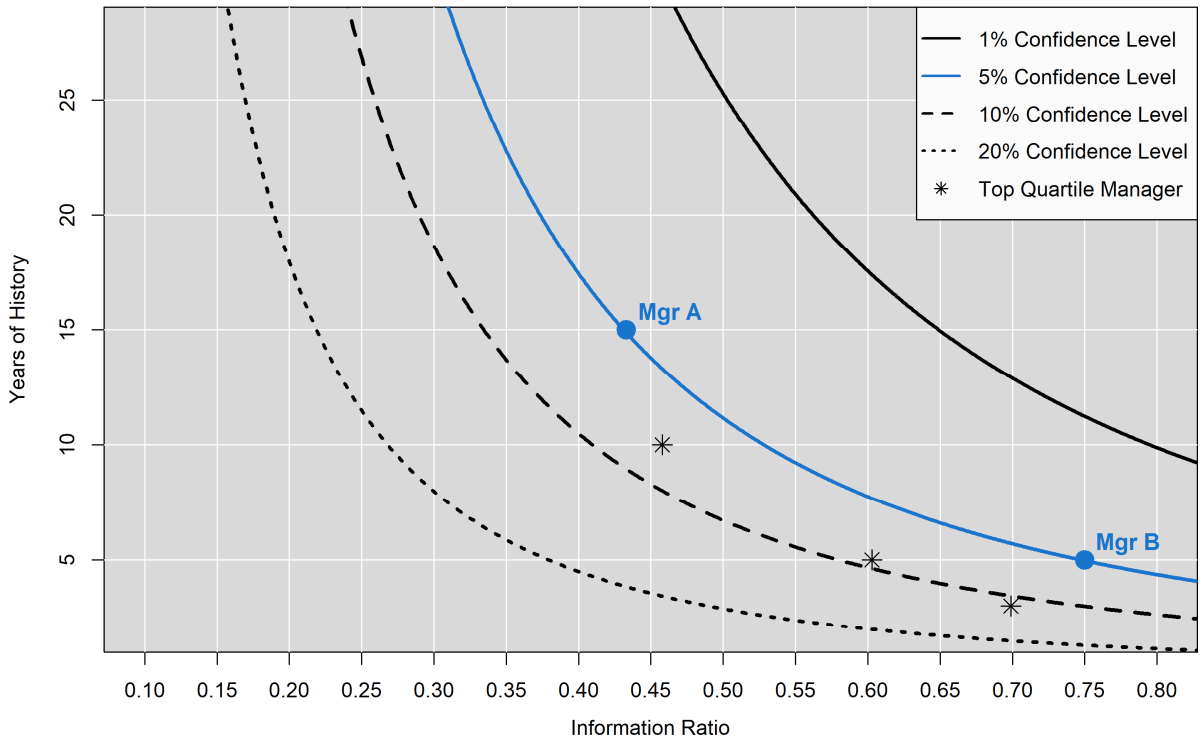
The superiority of larger t-statistics can be viewed as probabilities too[1]. A t-statistic of 1.68 corresponds to a 5% confidence level[2], which means that if 100 truly unskilled managers were studied together, about five of them should post a t-stat at least that high. Confidence levels are also called significance levels or Type I error rates.

Figure 1 graphs the relationship that IR and track record length have with confidence levels. The curved lines depict four distinct confidence levels, with stronger confidence (i.e. smaller % levels) moving from the bottom left corner towards the upper right. The two theoretical managers are marked in Figure 1, along with empirical quartile IRs observed for domestic, Value equity managers.

[1] A true statistician might bemoan this phrasing, as a population is not a random variable. Either the manger has skill or not; sampling data does not change that (unknowable) reality. In practice, this loose vernacular makes decision making clearer.

[2] One-tailed tests are used here, although an argument can be made that repeatable underperformance could be valuable too.

Figure 1: Confidence Levels as a Function of Information Ratio and Track Record



Source: eVestment and WEDGE Capital Management. Top quartiles span all products in the eVestment-defined U.S. Large Cap, Mid Cap, SMID Cap, and Small Cap Value equity universes that have monthly, gross-of-fee Information Ratios available for Q2 2022 versus their Russell Value benchmarks.

Empirical Data

A subtle feature in Figure 1 is that the number of years on the y-axis gets rather large. Most marketing materials and due diligence questionnaires do not lean on 25-year return metrics, if they are available at all. Databases of manager returns seem to take a similarly recent view: the eVestment database of institutional managers, for example, only stores IR metrics for up to ten-year time horizons.

Despite this time horizon myopia, the eVestment database is analyzed in Table 1. IRs for all domestic, Value equity products were downloaded at 3, 5, and 10-year time horizons. The results are interesting on multiple fronts. First, longer time horizons have fewer products to study. Such population shrinkage is almost certainly a symptom of survivorship bias: products with poor results are more likely to close or re-brand themselves, and disappear from a database, over longer time periods. Secondly, but at odds with the survivorship bias effect, the average and top quartile IRs decline as the time horizon lengthens.

This second finding aligns with the confidence framework described already. In Figure 1, the top quartile IRs (marked by asterisks) trace out a similar curve as the 10% confidence level. This result is meaningful because if no skill existed across the industry of active managers, the 25th percentile (i.e.

top quartile threshold) should follow a 25% confidence level curve. That the industry has posted IR results beyond that hurdle of randomness is an indication of collective skill (at least prior to fees).

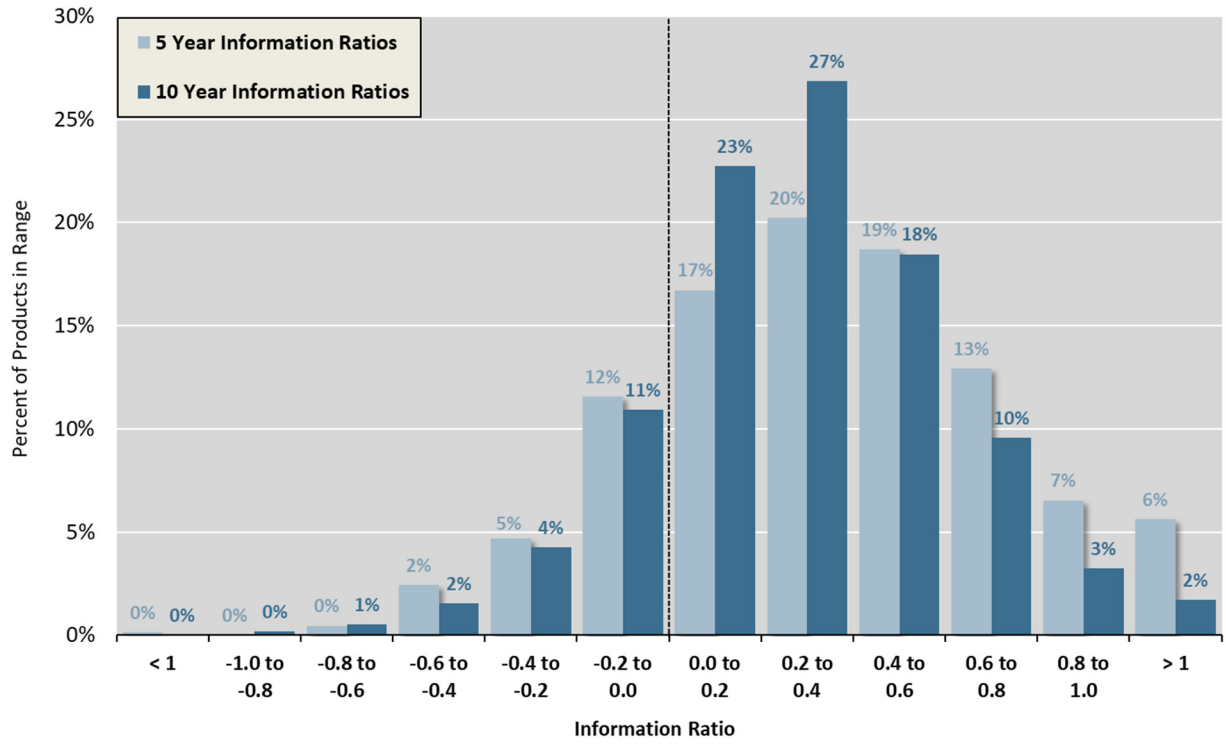
Table 1: Value Manager Information Ratios at 3, 5, and 10-Year Horizons

	Value Manager Statistics			% Change	
	3 Year	5 Year	10 Year	3 to 5Y	5 to 10Y
# of Products	697	658	585	-5.6%	-11.1%
Average IR	0.40	0.35	0.27	-13.5%	-21.2%
Std Deviation of IRs	0.50	0.41	0.33	-18.0%	-19.7%
Top Quartile IR	0.70	0.60	0.46	-13.7%	-24.0%
Median IR	0.37	0.33	0.28	-9.0%	-16.5%

Source: eVestment and WEDGE Capital Management. Presented data spans all products in the eVestment-defined U.S. Large Cap, Mid Cap, SMID Cap, and Small Cap Value equity universes that have monthly, gross-of-fee Information Ratios available for Q2 2022 versus their Russell Value benchmarks.

Figure 2 provides a more granular histogram of the 5 and 10-year manager data only. The same condensing effect is visible: IRs above ~0.6 or so become less frequent at a 10-year horizon than they are for 5 years, while the middle ranges of IRs increase in frequency.

Figure 2: Value Manager Information Ratios at 5 and 10-Year Horizons



Source: eVestment and WEDGE Capital Management. Presented data spans all products in the eVestment-defined U.S. Large Cap, Mid Cap, SMID Cap, and Small Cap Value equity universes that have monthly, gross-of-fee Information Ratios available for Q2 2022 versus their Russell Value benchmarks.

Practical Considerations

Past vs. Future

An important distinction for applying statistical methods to investing is that evolution can happen much more rapidly than in other scientific fields. Market participants are constantly trying to adapt and uncover new sources of alpha, whereas other scientific studies operate in a more static environment. Drug trials, for example, do not have to worry about the human genome dramatically changing in the next few years. The investment landscape can shift that rapidly though, so it is worth emphasizing the boilerplate compliance warning that past performance does not guarantee future results.

Technically, when studying past returns the "confidence" that is identified in a t-statistic only applies if the same set of circumstances from the dataset will occur again in the future. In the strictest sense this is impossible: the future always changes to some degree. However, many macroeconomic variables mean-revert over time, and cycles and manias (both high and low) have a stubborn way of repeating themselves. So, while the language of "proving manager skill" at a "X% level of confidence" may be too strong for reality, this type of confidence analysis can still be useful in a decision-making process.

Survivorship Bias

As mentioned earlier, a pervasive issue in studying databases of manager returns is that the worst performers tend to disappear. On the one hand, this issue can bias the pool of survivors to look better than the real-time landscape was. On the other hand, survivorship bias sets the bar unfairly high for individual manager comparisons. The "unfairness" arises because the surviving managers competed against – and likely outperformed – those who left the database, but no longer get credit for winning that battle in peer rankings. It is quite possible that over a long time horizon, a below-median ranked manager could have outperformed through the individual years comprising that longer horizon.

Multiple Testing

Data mining is also a concern for this type of confidence-based analysis. As more individual t-stats are calculated, the interpretation of confidence attached to each of them becomes weaker. In the classic example, if 100 monkeys built portfolios by throwing darts at a wall, five of them should end up passing a 5% confidence test. Calculating a t-statistic for all hundred monkeys would find those five lucky ones and "hire" them. Backtested investment performance should be especially suspicious in this regard, as there is typically no accounting of all the different permutations that were attempted before the final, attractive result is shown to outsiders. This backtesting concern extends to both quantitative products and individual factors: truly, there is no substitute for a live, out-of-sample track record.

There are methods that can correct for the multiple testing problem. A few papers on the topic are listed in the references section (see Benjamini and Hochberg, 1995; and Novy-Marx, 2015).

Gross or Net of Fees?

All analysis in this paper is conducted on a gross-of-fee basis. The same numerical work can be performed net-of-fees too, although there are a few reasons why gross is actually preferable, at least to start out with. Deducting the same, fixed fee each period lowers the numerator of IR but does not change the denominator (a uniform fee has no variance). Thus, a net-of-fee IR measure can distort the efficiency interpretation of active return per unit of active risk, especially over the longer timeframes where gross IRs tend to shrink towards a narrower range of outcomes. Fee changes or performance-based fee structures can be even more distortive, as they alter the IR without having anything to do with investment activity. Even the net calculation itself can introduce unwanted ambiguity, as there are different approaches managers can use to calculate net-of-fee returns.

Most important of all though, there is a natural remedy for the manager who scores well on gross IR and confidence but poorly on net results: demand lower fees! Finding a fix for weak gross-of-fee performance is much more difficult. For this reason, studies of the active management industry are often done in two stages. First, look to see if and where managers can add value before fees, and then consider whether their fee structures and sharing of alpha is reasonable. Berk and Green (2004) and Fama and French (2010) are two classic academic works that make this gross versus net distinction.

Concluding Remarks

Information Ratios have a natural connection to the statistics concept of confidence levels. For the same IR, a longer track record provides greater confidence that past outperformance did not come from luck. Empirical analysis looking across managers finds the same phenomenon: over longer time horizons, the IR threshold it takes to reach a particular peer ranking tends to shrink, while the confidence levels associated with those same breakpoints tend to improve. Ultimately, confidence estimates are not a panacea to solve manager allocation or model selection decisions, but they can be a useful tool, especially for comparing performance across different track record lengths.

Disclaimer

independent third-party sources and may be linked to third-party sources. WEDGE Capital Management L.L.P. "WEDGE" believes that the sources from which such information has been obtained are reliable; however, it cannot guarantee the accuracy of such information and has not independently verified the accuracy or completeness of such information.

This paper expresses the views of the author as of the date indicated and such views are subject to change without notice. WEDGE has no duty or obligation to update the information contained herein. Further, WEDGE makes no representation and it should not be assumed that past investment performance is an indication of future results. Further, wherever there is a potential for profit, there is also the probability of loss.

References

Akepanidtaworn, K., Di Mascio, R., Imas, A., and Schmidt, L. (2019). Selling Fast and Buying Slow: Heuristics and Trading Performance of Institutional Investors. *NBER Working Paper Series,* No. 29076. Available at: https://www.nber.org/papers/w29076.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society,* 57 (1), 289-300.

Berk, J. and Green, R. (2004). Mutual Fund Flows and Performance in Rational Markets. *Journal of Political Economy,* 112 (6), 1269-1295.

Cremers, M. and Pareek, A. (2016). Patient Capital Outperformance: The Investment Skill of High Active Share Managers Who Trade Infrequently. *Journal of Financial Economics,* 122 (2), 288-306.

Fama, E. and French, K. (2010). Luck versus Skill in the Cross-Section of Mutual Fund Returns. *Journal of Finance,* 65 (5), 1915-1947.

Gerakos, J., Linnainmaa, J., and Morse, A. (2016). Asset Managers: Institutional Performance and Smart Betas. *NBER Working Paper Series,* No. 22982. Available at: https://www.nber.org/papers/w22982.pdf.

Goodwin, T. (1998). The Information Ratio. *Financial Analysts Journal,* 54 (4), 34-43.

Novy-Marx, R. (2015). Backtesting Strategies based on Multiple Signals. *NBER Working Paper Series,* No. 21329. Available at: https://www.nber.org/system/files/working_papers/w21329/w21329.pdf.

Appendix

Details of Information Ratio Translation

Table 2 provides complete details for the translation between an Information Ratio and a hypothesis test of whether expected outperformance (i.e. the long-run numerator of Information Ratio) is positive. The operator E stands for expected value (in this case just a simple average) of what follows in parentheses, σ signifies a calculation of standard deviation, and N is the number of years of live history behind the Information Ratio being studied. r_p and r_b stand for product and benchmark returns, respectively, with $r_{p,t}$.and $r_{b,t}$ signifying specific returns at time t.

Table 2: Translation of Information Ratio to t-statistic and 5% Confidence Level

Null Hypothesis: manager has no skill; expected returns are indistinguishable or worse than the benchmark.	$E(r_p - r_b) \leq 0$
Alternative Hypothesis: the manager has skill; expected returns are higher than the benchmark.	$E(r_p - r_b) > 0$
Information Ratio: sample, average relative returns divided by standard deviation of those relative returns.	$\frac{E(r_{p,t} - r_{b,t})}{\sigma(r_{p,t} - r_{b,t})}$
Sample t-statistic: how far from the null hypothesis (no skill) the sample returns are, with N years of data.	$\frac{E(r_{p,t} - r_{b,t}) - 0}{\sigma(r_{p,t} - r_{b,t})/\sqrt{N}} = IR\sqrt{N}$
Years Required for 5% Confidence: a t-statistic greater than or equal to ~1.68 implies a 5% likelihood that the null case is true (i.e. no skill) but the data sample just got lucky.	$IR\sqrt{N} > t^*$ $N > \frac{1.68^2}{IR^2};\ \forall\ IR > 0$

The final line shown in Table 2, $N > \frac{1.68^2}{IR^2}$, is what graphs the 5% confidence curve in Figure 1. The other curves are created by substituting in the appropriate t-statistic for their respective confidence levels[3].

[3] In graphing Figure 1, all t-statistics use 60 degrees of freedom, which is reverse-engineered to set the t-statistic for 2.5% confidence to equal 2.00 (using a one-tailed test). A commonly used heuristic in research is to lean on t-stats above 2 to signify a meaningful result, as 2.00 corresponds to 5% for a two-tailed test. A more precise approach would calculate the degrees of freedom independently for each situation, but for sufficiently long time horizons will not make much of a difference.

Information Ratio Populations by Manager Style

The population of eVestment manager Information Ratios is compared across Value, Core, and Growth styles in Table 3. All three styles have roughly similar population sizes, but their IR distributions are quite different. Perhaps the most meaningful difference is the decline in average and median IRs moving from Value to Core and from Core to Growth. These population midpoints are a simple indication of whether there is manager skill on average in that style segment. However, the volatility of IRs across managers is noticeably wider for the Growth style. The net effect of these two competing forces leaves the top quartile IR for Growth managers roughly in line with that of Value and Core managers, especially over the five and ten year horizons.

Table 3: Style-Separated Information Ratios at 3, 5, and 10-Year Horizons

	Value Managers			Core Managers			Growth Managers		
	3 Year	5 Year	10 Year	3 Year	5 Year	10 Year	3 Year	5 Year	10 Year
# of Products	697	658	585	625	591	464	558	533	455
Average IR	0.40	0.35	0.27	0.20	0.16	0.14	-0.01	0.13	0.07
Std Deviation of IRs	0.50	0.41	0.33	0.52	0.47	0.39	0.72	0.62	0.41
Top Quartile IR	0.70	0.60	0.46	0.54	0.50	0.40	0.49	0.61	0.41
Median IR	0.37	0.33	0.28	0.18	0.17	0.14	0.03	0.11	0.05

Source: eVestment and WEDGE Capital Management. Presented data spans all products in the eVestment-defined U.S. Large Cap, Mid Cap, SMID Cap, and Small Cap equity universes for each of the Value, Core, and Growth styles. Only products with monthly, gross-of-fee Information Ratios available for Q2 2022 versus their size and style-appropriate Russell benchmarks are included.

Table 4 breaks up the pool of Value managers specifically into Large, Mid, and Small-Cap size categories[4]. All three size ranges show positive average IRs. The Large Cap Value managers have done better over the past three years, but over a ten-year horizon the results tilt in favor of the Small Cap Value managers.

Despite the minor differences in Table 4, just realizing that average IRs have been positive and relatively similar across market cap ranges carries some practical value, as it undermines a common refrain in asset allocation: that the Large Cap equity space is too competitive, and impossible to outperform in. Even for the worst of the three time horizons studied, a 0.25 average IR for Large Cap Value managers suggests that reasonable fee levels[5] can be supported across the industry. The concern of survivorship

[4] SMID managers are dropped for this analysis because of their structural overlap with Mid and Small caps.

[5] For example, a 4% tracking error and 0.25 average IR would imply an even 1% of alpha/fees available to share, on average.

bias is still lurking, but a simple estimation[6] suggests that bias is not strong enough to overturn the result of positive average IRs.

These observations from the eVestment database align with academic literature that finds institutional money managers outperform, on average (see Cremers and Pareek, 2016; Gerakos, et al. 2016; and Akepanidtaworn, et al. 2020). This conclusion does not contradict the mathematical identity that the average investor must hold the market portfolio, and therefore cannot outperform, since other market participants – such as mutual funds and retail investors, for example – can underperform to maintain that equilibrium.

Table 4: Size-Separated Value Information Ratios at 3, 5, and 10-Year Horizons

	Large Cap Value			Mid Cap Value			Small Cap Value		
	3 Year	5 Year	10 Year	3 Year	5 Year	10 Year	3 Year	5 Year	10 Year
# of Products	331	317	284	80	74	67	207	196	177
Average IR	0.51	0.43	0.25	0.39	0.28	0.29	0.28	0.27	0.31
Std Deviation of IRs	0.51	0.44	0.34	0.49	0.41	0.42	0.46	0.33	0.27
Top Quartile IR	0.79	0.70	0.44	0.70	0.48	0.47	0.59	0.46	0.47
Median IR	0.49	0.43	0.26	0.38	0.28	0.29	0.24	0.27	0.31

Source: eVestment and WEDGE Capital Management. Presented data spans all products in the eVestment-defined U.S. Large Cap, Mid Cap, and Small Cap Value equity universes that have monthly, gross-of-fee Information Ratios available for Q2 2022 versus their Russell Value benchmarks.

[6] For the five and ten year horizons, take the decrease in manager count versus the three year horizon, and assign all of those "missing" managers an Information Ratio equal to the bottom-decile point for that time horizon (which is negative in all cases). Then, calculate a new average IR across both the visible managers and the assumed exits that were assigned bottom-decile IRs. This procedure would lower the average IRs shown in Table 4 by a range of 0.02 to 0.08 only.