# Internet Appendix for

# "Rewriting History"*

Alexander Ljungqvist
*Stern School of Business*
*New York University*
and *CEPR*

Christopher Malloy
*Harvard Business School*

Felicia Marston
*McIntire School of Commerce*
*University of Virginia*

September 10, 2008

* Citation format: Ljungqvist, Alexander, Christopher Malloy, and Felicia Marston, 2009, Internet Appendix for "Rewriting History," *Journal of Finance* 64, 1935-1960, http://www.afajof.org/IA/2009. Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the authors of the article.

Ljungqvist, Malloy, and Marston (2009) document widespread changes to the historical I/B/E/S analyst stock recommendations database. This document contains supplementary tests and results.

*A. Patterns in Popular Conditioning Variables*

Recommendations affected by the changes to the I/B/E/S recommendations history appear to cluster according to three popular conditioning variables: the analyst's reputation, the brokerage firm's size and status, and the boldness of the recommendation. We measure analyst reputation using all-star status, as designated in the Oct. issue of *Institutional Investor* magazine preceding the recommendation in question. We divide brokerage firms into the 12 (generally large) firms sanctioned under the Global Settlement and all other firms, and we code a recommendation as bold if it was one notch or more above or below consensus (=mean recommendation) computed over the prior three months (requiring at least three outstanding recommendations).

In Table IA.I, we compare the frequency of these conditioning variables in the universe of historical recommendations and in the set of changed recommendations. We compare each tape to the next tape as well as to the adjusted 2007 tape.

We find that all-stars are significantly overrepresented among changed recommendations on the 2000 and 2001 tapes, while changed recommendations on the 2002 through 2004 tapes come disproportionately from unrated analysts. Relative to the adjusted 2007 tape, recommendations by unrated analysts are significantly more likely to need correction on every tape except the 2001 tape. Thus, tests comparing all-stars to unrated analysts may yield different results depending on which tape is used. Sanctioned banks are overrepresented among affected recommendations on the 2000 and 2001 tapes, and underrepresented on all later tapes. Relative to the adjusted 2007 tape, sanctioned banks are associated with a significantly lower need for corrections on every tape except the 2001 tape. Finally, bold recommendations are significantly overrepresented among affected records on all tapes. They are also consistently and significantly more likely to be subject to corrections on the adjusted 2007 tape.

*B. Further Evidence on the Effect of the Changes on Trading Signals*

To investigate some of the results in Ljungqvist, Malloy, and Marston's (2009) Tables IV to VI in greater depth, we employ a series of additional tests. First, we explore subsets of Table IV, Panels B and C by removing movements to and from one category at a time. For example, to determine if the rating category 4 (strong sell) is driving the results, we re-run the analysis for "All upgrades" and "All downgrades" by removing any upgrades/downgrades that involve movements to and from the rating category 4. We repeat this analysis for movements to and from each individual rating category (1-5). Panel A of Table IA.II reports summary statistics from this analysis.

The results indicate that while removing upgrades to and from category 4 does decrease the difference in event returns between the two tapes for "All upgrades," the remaining difference is still large economically and statistically significant (0.426%, $t$=8.67). Removing movements to and from categories 3 and 5 has a similar, but somewhat smaller dampening effect, while removing movements to and from categories 1 and 2 has the opposite effect (i.e., it increases the difference in event returns between the two tapes). Overall, the results in Table IV, Panel B do not appear to be driven by any particular rating category.

When we repeat this analysis for downgrades (Panel B of Table IA.II), we find similar results: although removing movements to and from category 4 has the biggest impact on the differential between the two tapes, it does not drive the differences between the two tapes. The differences in event returns for downgrades are still large and significant across the two tapes, regardless of which category we choose to remove.

To recap, the differences in event returns for both upgrades and downgrades remain economically large and statistically significant across the two tapes, regardless of which category we choose to remove. However, removing category 4 does have a nontrivial impact on these results. To understand why this is the case, we also examine some observations by hand. In doing so, we discover that the reason category 4 has a nontrivial impact in this particular case is that a large number of recommendations issued by eight quantitative research groups were erroneously included in the database around this time, and are captured on the 2004 tape. The quantitative research groups produce algorithmic recommendations constrained to be symmetrically distributed; hence, they include more category 4 and 5 recommendations than usual. According to Thomson, these recommendations were not supposed to be viewable by its clients yet were added to the database sometime between 2003 and 2004. They were subsequently deleted at some point between 2004 and 2005. This suggests that deletions from the 2007 tape are the main drivers of the changes we document in Table IV, Panels B and C.

To demonstrate this more formally, we take a different approach that seeks to trace back the impact of each of the four types of data changes that we document in Tables I and II (i.e., the additions, deletions, alterations, and anonymizations) on the results in Tables IV to VI. Note that since upgrades and downgrades are computed at the broker level, anonymizations will have no impact on the Table IV results. To measure the impact of each of the other three types of changes individually, we create three hypothetical 2004 tapes, each of which simulates what the 2004 tape would have looked like if one type of change had *not* occurred. We then use each of these new hypothetical tapes as the 2004 tape, recalculate all the trading signals, and re-compute the 2004 to 2007 differences reported in Table IV, Panels A and B. Specifically, we compute the 2004 tape: a) with "no deletions," meaning that we omit on the 2004 tape those records that are later deleted on the 2007 file, b) with "no alterations," meaning that we re-instate altered records to original values on the 2004 tape as we have done on the 2007 tape, and c) with "no additions," meaning that we take those records that were added on the 2007 tape and add them back to the 2004 tape. Note that we do not make any changes to the 2007 tape, so the 2007 event returns stay the same across all comparisons. We report these results in Table IA.III.

Consistent with our explanation above, this table demonstrates that the differences reported in Table IV, Panels B and C for the 2004 to 2007 comparison are primarily due to the impact of deletions from the 2007 tape (of which the quantitative research groups' recommendations are the primary subset; the remainder involve the removal of certain brokerage firm histories by request of the broker as of 2007). Once these deletions are removed from the 2004 snapshot, the event returns are not statistically different between the 2004 and 2007 tapes.

This raises the question of whether all of the results are driven by recommendations that are subsequently removed on the 2007 tape. To test this conjecture, we randomly pick three additional pairwise comparisons for the tests in Tables IV to VI, and perform the same in-depth decompositions as above for these new cases. (Assembling the hypothetical tapes and running the analysis is very time-consuming, which is why we choose three random comparisons.) For Table IV, we choose the 2000 tape, for Table V we choose the 2003 tape, and for Table VI we choose the 2002 tape, each of which is compared to the 2007 tape.

As with the 2004-2007 results reported in Table IV, the differences in 3-day event returns between the 2000 and 2007 tape are significant (although smaller in magnitude). However, unlike the 2004 to 2007 differences, which are driven mainly by deletions from the 2007 tape, as demonstrated above, the 2000 to 2007 differences are driven primarily by additions to the 2007 tape (i.e., records that are not on the 2000 tape but that have been added back on the 2007 tape). Only when these 2007 additions are added back to the 2000 tape does the spread in event returns between the two tapes become insignificant. In magnitude, only 1% of the difference in upgrade event returns between the 2000 and 2007 tapes is due to deletions (21% for downgrades), while 7% of the difference is due to alterations (3% for downgrades), and 85% of the difference is due to additions (76% for downgrades).[1] Thus, the *addition* of records to the 2007 tape appears to be the primary cause of the significant differences in event returns between the 2000 and 2007 tapes.

In Ljungqvist, Malloy, and Marston's (2009) Table V tests, when we conduct similar breakdowns for the 2003 vs. 2007 comparison, we find that most of the difference in the pre-2001 DGTW spread (=-0.020 from the table) is due to alterations: only when we remove the alterations from the 2003 tape does this spread become insignificant. In magnitude, approximately 12% of the difference in the spread is due to deletions, 53% is due to alterations, and 22% is due to additions.

In Ljungqvist, Malloy, and Marston's (2009) Table VI tests, we need to compute four hypothetical 2002 tapes rather than three, because anonymizations can impact individual analyst-level persistence (even though they cannot impact the measures of consensus recommendations from Table V, nor the upgrade/downgrade measures from Table IV). We find that most of the difference in the persistence spread is again due to additions: only when we add the 2007 additions back to the 2002 tape does this spread become insignificant. In magnitude, approximately 21% of the difference in the spread is due to deletions, 9% is due to alterations, 36% is due to additions, and 1% is due to anonymizations.

In summary, across all tests, differences between the various pairwise comparisons appear to be caused by different combinations of the four types of data changes.

*C. Further Evidence on the Persistence in Analyst Stock-picking*

Ljungqvist, Malloy, and Marston (2009) find evidence of persistence in individual analysts' stock-picking performance on every I/B/E/S tape, but the extent of persistence varies markedly across tapes. In Table IA.IV, we show that this result is even more pronounced if we filter on analysts' all-star status (defined as in Section I.C). A common modification to the persistence trading strategy is to buy on recommendations by all-star analysts who are also in quintile 5 and to sell on recommendations by non-all star analysts ranked in quintile 1. This assumes asymmetry in persistence among all-stars: they are likely to repeat good past performance but not poor past performance. Imposing this screen increases the differences in persistence spreads across the tapes. For example, we find a difference between the 2001 and 2007 tapes of 82.0 basis points over five trading days, an increase of 25.3% relative to the amount of persistence found on the 2007 tape. Similarly large differences exist between the 2002 and 2007 tapes (66.3 basis points, a 21.1% increase relative to 2007) and between the 2003 and 2007 tapes (36.6 basis

[1] Note that these numbers do not have to add up to 100% since the removal of one category at a time does not constitute a complete decomposition (since removing two categories at a time can result in additional differences as well).

points, a 12.1% increase relative to 2007).

## REFERENCES

Ljungqvist, Alexander, Christopher Malloy, and Felicia Marston, 2009, Rewriting History, *Journal of Finance* 64, 1935-1960.

## Table IA.I
## Patterns in Popular Conditioning Variables

The table documents patterns in the changes to the I/B/E/S historical recommendations database analyzed in Ljungqvist, Malloy, and Marston (2009). We examine year-over-year changes to the database by comparing data from adjacent annual downloads. We focus on the period for which each pair of downloads has overlapping coverage (that is, we ignore recommendations from the later tape that are dated after the cutoff date of the earlier tape.) The cutoff dates of our tapes are 7/20/00 ("2000 tape"), 1/24/02 ("2001 tape"), 7/18/02 ("2002 tape"), 3/20/03 ("2003 tape"), 3/18/04 ("2004 tape"), 12/15/05 ("2005 tape"), and 9/20/07 ("2007 tape"). According to Thomson, the 2007 tape contains data purged of all data errors we have identified, except that it continues to include broker-requested retrospective changes to recommendation scales. We also compare the 2000 through 2005 tapes to the 2007 tape, after reversing the broker-requested retrospective changes to recommendation scales. This adjusted version of the 2007 tape corresponds to the "as-was" historical recommendations database, which Thomson intends to make available to researchers in response to our investigation. In Ljungqvist, Malloy, and Marston (2009), we document four types of changes to the I/B/E/S recommendations data. We define an *alteration* as a broker/ticker/date triad that appears on both tapes but for which the recommendation on one tape is different than on the other tape. A *deletion* is a broker/ticker/date triad that appears on the earlier tape but not on the later tape to which it is compared. An *addition* is a broker/ticker/date triad that appears on the later comparison tape but not on the earlier tape. *Anonymizations* refer to cases where the analyst associated with a broker/ticker/date triad is identified by name on the earlier tape but is anonymous on the later tape. In this table, we compare the frequency of three popular conditioning variables in the universe of historical recommendations and in the set of recommendations subject to ex post changes (due to alterations, deletions, additions, or anonymizations). The three variables of interest condition on whether the analyst has all-star status (the top three rated analysts in each sector, as designated in the Oct. issue of *Institutional Investor* magazine preceding the recommendation in question), whether the brokerage firm is among the 12 firms sanctioned under the Global Settlement, and whether the recommendation was "bold," where bold is an indicator equaling one if the recommendation was one notch or more above or below consensus (=mean recommendation) computed over the prior three months (requiring at least three outstanding recommendations). We test for differences in fractions using standard two-sample *F*-tests of equal proportions. The tests compare the universe to the set of changed recommendations. Statistically significant differences at the 5% level are indicated with *.

| | **All-star analysts** | | | **Global Settlement banks** | | | **Bold recommendations** | | |
|---|---|---|---|---|---|---|---|---|---|
| | Share of recom-mendations universe | Share of changed recommendations | | Share of recom-mendations universe | Share of changed recommendations | | Share of recom-mendations universe | Share of changed recommendations | |
| | | relative to next tape | relative to 2007 tape | | relative to next tape | relative to 2007 tape | | relative to next tape | relative to 2007 tape |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| 2000 tape | 15.3% | 19.6%* | 9.1%* | 24.3% | 37.6%* | 10.4%* | 30.6% | 31.8%* | 35.9%* |
| 2001 tape | 13.3% | 23.1%* | 16.3%* | 23.2% | 44.2%* | 28.7%* | 31.0% | 33.1%* | 35.6%* |
| 2002 tape | 15.0% | 10.5%* | 8.6%* | 26.9% | 16.0%* | 12.4%* | 30.4% | 33.6%* | 37.3%* |
| 2003 tape | 15.0% | 6.4%* | 9.2%* | 28.8% | 13.7%* | 19.9%* | 31.6% | 39.6%* | 36.2%* |
| 2004 tape | 11.5% | 0.4%* | 1.4%* | 23.8% | 1.7%* | 6.6%* | 32.5% | 36.3%* | 35.6%* |
| 2005 tape | 13.9% | 14.0% | 4.0%* | 30.0% | 12.7%* | 19.1%* | 32.3% | 30.5%* | 33.7%* |

## Table IA.II
## Effect of Changes on the Abnormal Returns to Upgrades and Downgrades, Removing One Rating Category at a Time

This table compares the event returns to upgrades and downgrades for the 2004 and 2007 I/B/E/S tapes, and complements Table IV in Ljungqvist, Malloy, and Marston (2009). Panels A and B report differences in the three-day event-time returns between the 2004 and 2007 tapes for upgrades and downgrades, respectively. The first column includes all upgrades (downgrades), while the subsequent columns exclude all upgrades (downgrades) to and from a particular rating category (e.g., I/B/E/S recommendation code 2 = "buy"). *ERet04* and *ERet07* are the three-day raw event returns, calculated as the geometrically cumulated return for the day before, day of, and day after the recommendation, using data from the 2004 and 2007 tapes, respectively. *DiffEret* then equals the average difference between *ERet04* and *ERet07*. Analogously, we compute the three-day excess event return as the raw stock return less the appropriate size-decile return of the CRSP NYSE/Amex/NASDAQ index (not shown for brevity) and report *DiffEXret*, the average difference between the three-day excess return samples. *t*-statistics are in parentheses, and 5% statistical significance is indicated with *.

Panel A: Three-day upgrade event returns (in %): 2004 versus 2007 tapes

| | All Upgrades | All upgrades except to and from category: | | | | |
|---|---|---|---|---|---|---|
| | | *1* | *2* | *3* | *4* | *5* |
| *ERet07* | 3.02* | 2.99* | 2.86* | 3.04* | 3.02* | 3.07* |
| | (82.91) | (55.41) | (44.36) | (46.40) | (82.01) | (82.17) |
| *ERet04* | 2.30* | 1.87* | 1.93* | 2.47* | 2.60* | 2.28* |
| | (78.47) | (50.52) | (41.65) | (43.46) | (79.25) | (78.61) |
| *DiffEret* | 0.72* | 1.11* | 0.94* | 0.57* | 0.43* | 0.65* |
| | (15.37) | (17.48) | (12.02) | (6.62) | (8.67) | (13.65) |
| *DiffXret* | 0.72* | 1.12* | 1.00* | 0.60* | 0.42* | 0.66* |
| | (15.63) | (17.49) | (12.57) | (6.95) | (8.61) | (13.82) |

Panel B: Three-day downgrade event returns (in %): 2004 versus 2007 tapes

| | All downgrades | All downgrades except to and from category: | | | | |
|---|---|---|---|---|---|---|
| | | *1* | *2* | *3* | *4* | *5* |
| *ERet07* | -4.72* | -4.78* | -4.99* | -4.14* | -4.73* | -4.76* |
| | (-103.34) | (-70.97) | (-55.32) | (-54.39) | (-103.84) | (-102.03) |
| *ERet04* | -3.79* | -3.30* | -3.59* | -3.38* | -4.20* | -3.93* |
| | (-99.21) | (-65.91) | (-53.16) | (-50.67) | (-101.90) | (-98.92) |
| *DiffEret* | -0.93* | -1.49* | -1.41* | -0.76* | -0.54* | -0.83* |
| | (-15.66) | (-18.08) | (-12.73) | (-7.50) | (-9.75) | (-13.57) |
| *DiffXret* | -0.89* | -1.42* | -1.45* | -0.71* | -0.50* | -0.79* |
| | (-14.74) | (-16.85) | (-12.57) | (-7.00) | (-8.09) | (-12.67) |

## Table IA.III
## Effect of Changes on the Abnormal Returns to Upgrades and Downgrades, Removing One Type of Data Change at a Time

This table compares the event returns to upgrades and downgrades for the 2004 and 2007 I/B/E/S tapes, and complements Table IV in Ljungqvist, Malloy, and Marston (2009). The table reports differences in the three-day event-time returns between the 2004 and 2007 tapes for upgrades and downgrades. To measure the impact of each of the three types of data changes (deletions, alterations, additions) individually, we create three hypothetical 2004 tapes, each of which simulates what the 2004 tape would have looked like if one type of data change had *not* occurred. We then use each of these new hypothetical tapes as our 2004 tape, recalculate all the trading signals, and re-compute the 2004 to 2007 differences reported in Table IV, Panels A and B. Specifically, we compute the 2004 tape: a) with "no deletions," meaning that we omit on the 2004 tape those records that are later deleted on the 2007 file, b) with "no alterations," meaning that we re-instate altered records to original values on the 2004 tape as we have done on the 2007 tape, and c) with "no additions," meaning that we take those records that were added on the 2007 tape and add them back to the 2004 tape. Note that we do not make any changes to the 2007 tape, so the 2007 event returns stay the same across all comparisons. *ERet04* and *ERet07* are the three-day raw event returns, calculated as the geometrically cumulated return for the day before, day of, and day after the recommendation, using data from the 2004 and 2007 tapes, respectively. *DiffEret* then equals the average difference between *ERet04* and *ERet07*. Analogously, we compute the three-day excess event return as the raw stock return less the appropriate size-decile return of the CRSP NYSE/Amex/NASDAQ index (not shown for brevity) and report *DiffEXret*, the average difference between the three-day excess return samples. *t*-statistics are in parentheses, and 5% statistical significance is indicated with *.

| Three-day upgrade event returns (in %): 2004 hypothetical tapes versus 2007 tape | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | All upgrades | | | | All downgrades | | | |
| | All changes | No deletions | No alterations | No additions | All changes | No deletions | No alterations | No additions |
| *ERet07* | 3.02* | 3.02* | 3.02* | 3.02* | -4.72* | -4.72* | -4.72* | -4.72* |
| | (82.91) | (82.91) | (82.91) | (82.91) | (-103.34) | (-103.3) | (-103.34) | (-103.34) |
| *ERet04* | 2.30* | 3.06* | 2.30* | 2.30* | -3.79* | -4.78* | -3.77* | -3.80* |
| | (78.47) | (82.50) | (78.60) | (78.84) | (-99.21) | (-102.7) | (-99.16) | (-99.95) |
| *DiffEret* | 0.72* | -0.048 | 0.72* | 0.71* | -0.93* | 0.069 | -0.95* | -0.92* |
| | (15.37) | (0.93) | (15.61) | (15.39) | (-15.66) | (0.71) | (-16.50) | (-15.65) |
| *DiffXret* | 0.72* | -0.028 | 0.73* | 0.72* | -0.89* | 0.054 | -0.91* | -0.89* |
| | (15.63) | (0.61) | (15.75) | (15.61) | (-14.74) | (0.81) | (-15.21) | (14.70) |

## Table IA.IV
## Effect of Changes on Persistence in Individual Analyst Performance

The table reports tests of persistence in individual analysts' stock-picking skills. The table complements Table VI in Ljungqvist, Malloy, and Marston (2009). Tests are performed separately on the 2000, 2001, 2002, 2003, 2004, 2005, and 2007 tapes. For each analyst, we compute the average five-day DGTW-adjusted return of all upgrades and downgrades issued by that analyst over the previous six months; in doing so, we assume that we buy on upgrades and sell on downgrades. We then rank analysts into quintiles in January and July of each year based on their average five-day DGTW-adjusted return over the prior six months. Next we compute a "persistence spread" equal to the difference between the average five-day DGTW-adjusted return of analysts in the highest quintile (Q5) minus the average five-day DGTW-adjusted return of analysts in the lowest quintile (Q1), in each case computed over the following six months. The five-day return is the geometrically cumulated DGTW-adjusted return for the two trading days before through the two trading days after the recommendation. Daily DGTW characteristic-adjusted returns are defined as raw returns minus the returns on a value-weighted portfolio of all CRSP firms in the same size, (industry-adjusted) market-to-book, and one-year momentum quintile. We report persistence spreads for each I/B/E/S tape from 2000 through 2005 (shown in column (1)) and for the 2007 tape (shown in column (2)). Note that each tape is compared over its full available sample period to the 2007 tape, so the estimates for the 2007 tape shown in column (2) are different for each comparison tape. In column (3), we report differences between each tape and the 2007 tape. The results in this table are computed identically to those in Table VI in Ljungqvist, Malloy, and Marston (2009), except that we impose an additional all-star filter: we restrict quintile 5 to be the subset of quintile 5 analysts who are also all-star analysts (as designated in the preceding Oct. issue of *Institutional Investor* magazine), and we restrict quintile 1 to be the subset of quintile 1 analysts who are *not* also all-star analysts. *t*-statistics are shown in parentheses, and 5% statistical significance is indicated with *.

| Average five-day event returns (in %) from persistence quintiles with all-star screens included | | | |
|---|---|---|---|
| | Persistence spread (Q5-Q1) (1) | Persistence spread (Q5-Q1) from 2007 tape (2) | Difference in persistence spreads, 2007-200X (3) |
| 2000 tape | 3.049* | 3.158* | 0.109 |
| | (4.61) | (7.60) | (0.22) |
| 2001 tape | 4.059* | 3.239* | -0.820* |
| | (9.32) | (8.74) | (-2.80) |
| 2002 tape | 3.811* | 3.149* | -0.663* |
| | (9.11) | (8.72) | (-2.57) |
| 2003 tape | 3.404* | 3.038* | -0.366 |
| | (10.01) | (8.46) | (-1.94) |
| 2004 tape | 3.131* | 2.964* | -0.168 |
| | (9.32) | (9.03) | (-1.07) |
| 2005 tape | 2.991* | 2.897* | -0.094 |
| | (10.43) | (9.94) | (-0.82) |