





















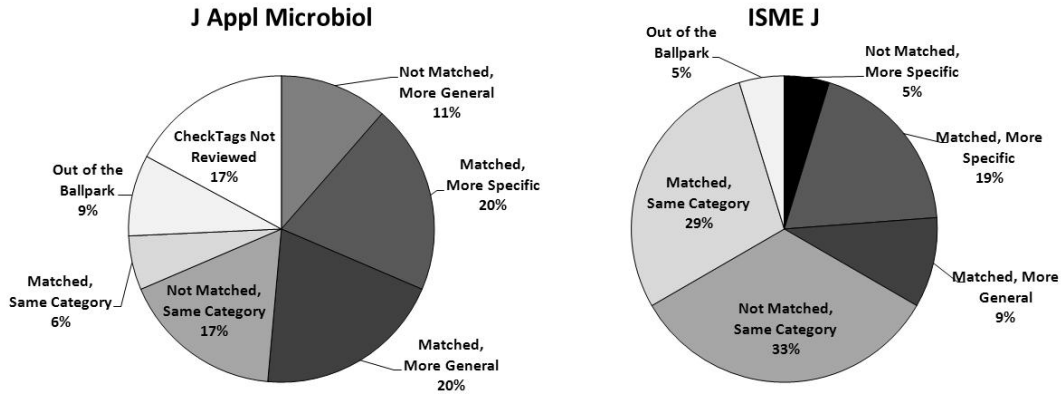






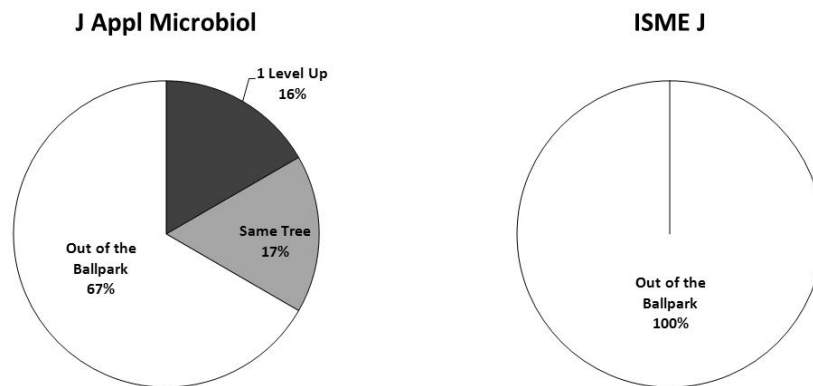


The reasons why MeSH terms were deleted were varied making it difficult to generalize from this small sample size. *Matched, More Specific* and *Not Matched, Same Category* were more prominent, suggesting MeSH terms were deleted and more specific MeSH terms were added because the original MTIFL-indexed terms were too general (see Table 5 and Figure 5).



**Figure 5.** LHC-generated evaluation data (in percent) for MTIFL indexed MeSH terms changed by indexers.

LHC also generated evaluative data for the reasons why IM terms were deleted by comparing human added to deleted MTIFL-indexed IM terms (see Figure 6). Because of our small sample size, not many IM terms could be evaluated (six from *J Appl Microbiol* and one from *ISME J*). For *J Appl Microbiol*, 67% (or 4 out of the 6) IM terms fit the category of *Out of the Ballpark*. The other two IM terms for *J Appl Microbiol* were found to be either in the *Same Tree* or *One Level Up* when compared to the added IM term.



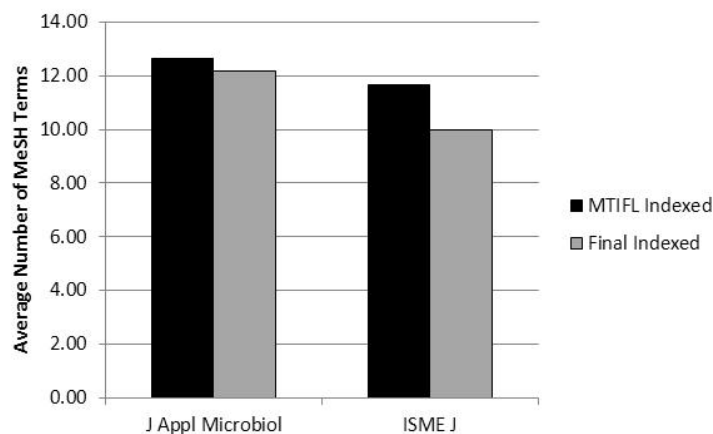
**Figure 6.** LHC-generated evaluation data (in percent) for IM terms changed by indexers.

## Senior Indexer Analysis

Two NLM Senior Indexers evaluated indexing on 12 MTIFL-indexed articles. First, the number of MeSH terms indexed by MTIFL, MeSH terms deleted and added by indexers, IM terms, critical MeSH terms and final number of indexed terms were calculated and averaged (see Table 6 and Figure 7). A Student's two-tailed, unpaired t-test determined there were no significant differences between the number of MeSH terms indexed by MTIFL compared to the final number of MeSH terms indexed after completion by human indexer (*J Appl Microbiol*  $t(10) = 0.2007$ ,  $p=0.8449$ , and *ISME J*  $t(10) = 0.7187$ ,  $p=0.4888$ ). Although MeSH term counts were not exactly similar, the data correspond to those statistics from LHC, as reported above. The discrepancies are most likely due to an experimenter counting error.

	MTIFL-Indexed MeSH	Deleted MeSH	Added MeSH	IM Terms	Critical MeSH	Final Indexing
<b>J Appl Microbiol</b>						
<b>Sum</b>	76	30	27	6	25	73
<b>Average</b>	12.67	5.00	4.50	1.00	4.17	12.17
<b>ISME J</b>						
<b>Sum</b>	70	21	11	2	8	60
<b>Average</b>	11.67	3.50	1.83	0.33	1.33	10.00

**Table 6.** Total and average number of MTIFL-indexed MeSH terms, MeSH terms added and deleted by indexers, IM terms, critical MeSH terms and final indexed MeSH terms for six *J Appl Microbiol* and 6 *ISME J* articles, as determined by the Senior Indexer Analysis.



**Figure 7.** Average number of MeSH terms for six *J Appl Microbiol* and six *ISME J* articles. MTIFL-indexed MeSH term counts (dark bar) and final number of MeSH terms indexed (light bar) are displayed.

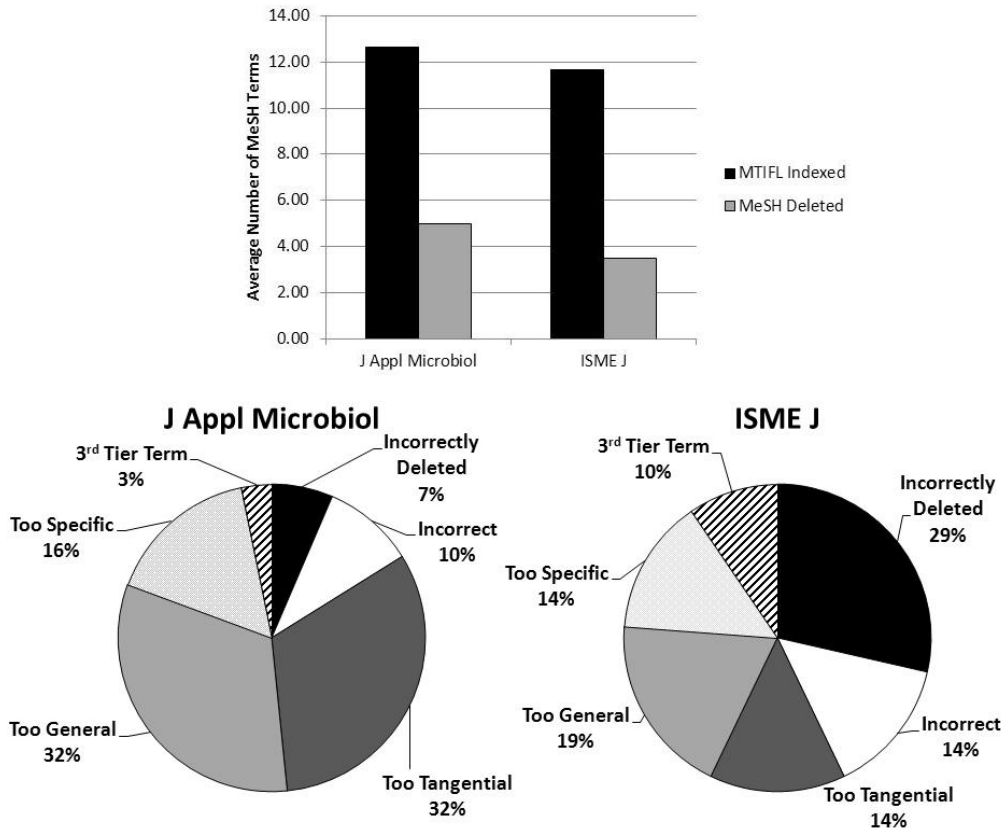
Next, the Senior Indexers evaluated MeSH terms that were deleted by the indexers. For the *J Appl Microbiol* articles, MTIFL originally indexed 76 MeSH terms and 30 were deleted by indexers (average 12.67 indexed, 5 deleted per article). These results were statistically significant, suggesting the indexers made significant changes to MTIFL indexing by deleting terms ( $t(10) = 3.2658, p=0.0085$ ). Similarly for *ISME J* articles, MTIFL originally indexed 70 MeSH terms and 21 were deleted by indexers (average 11.67 indexed and 3.50 deleted per article) and these were found to be significantly different ( $t(10) = 3.4426, p=0.0063$ ). Again, these results were similar to those described in the LHC statistics.

To evaluate why MTIFL-indexed MeSH terms were deleted, the Senior Indexers grouped reasons deleted into six categories:

- *Incorrect* – Wrong MeSH term
- *Too Tangential* – Inappropriate
- *Too General* – Higher up in the MeSH tree
- *Too Specific* – Lower down in the MeSH tree
- *3<sup>rd</sup> Tier Term* – Not the point, but can be used
- *Incorrectly Deleted* – Should not have been deleted

Figure 8 shows the categorical representation for the deleted terms in a percentage. For *J Appl Microbiol* the deleted terms primarily fell into the categories of *Too Tangential* (32%), or *Too General* (32%). For *ISME J*, deleted terms primarily fell into the categories of *Incorrectly Deleted* (29%), or *Too General* (19%). All other categories were also near equal in representation and with this small sample size only generalizations can be made about why MeSH terms were deleted.





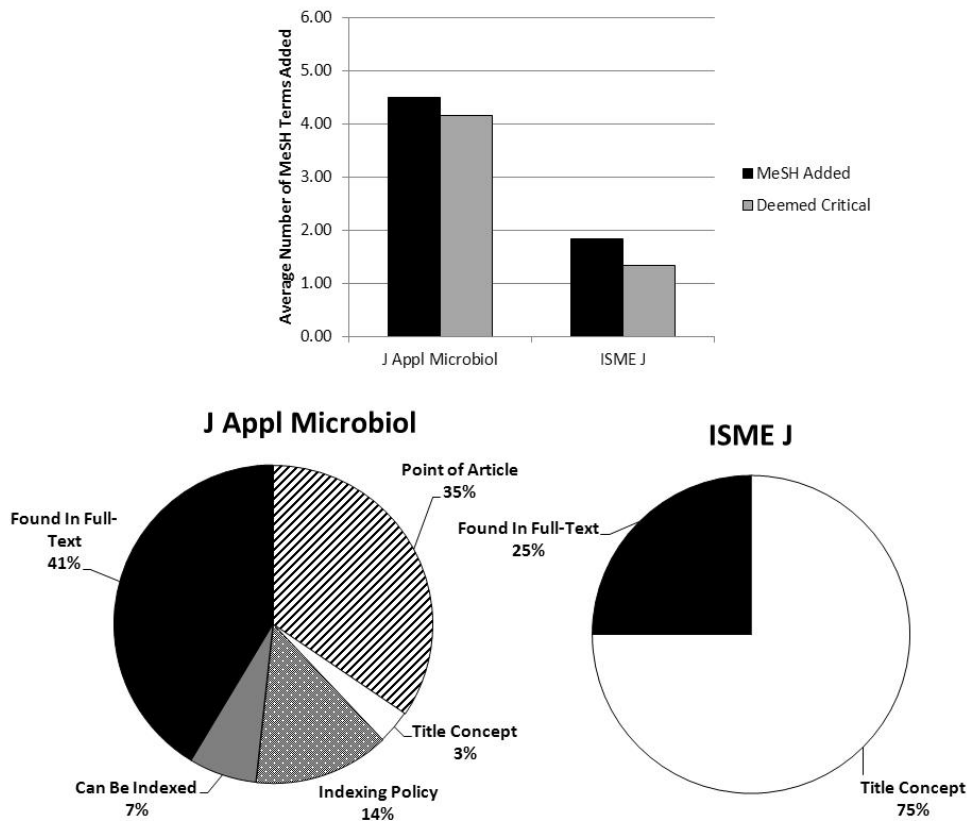
**Figure 8.** Top panel displays the average number of MeSH terms indexed by MTIFL (dark bars) and deleted by indexers (light bars) and the bottom panel displays categorical representation on why the MeSH terms were deleted by the indexers.

Next, the Senior Indexers evaluated MeSH terms that were added by the indexers. For the *J Appl Microbiol* articles 27 (average 4.5 per article) and for *ISME J* articles 11 (average 1.83 per article) MeSH terms were added by indexers. The number of MeSH terms added during the indexing process was not found to be a significant number when compared to the final number of MeSH terms; *J Appl Microbiol*,  $t(10) = 0.2806$ ,  $p=0.7847$ ; and *ISME J*,  $t(10)=0.9583$ ,  $p=0.3605$ . However of the 27 MeSH terms added for *J Appl Microbiol*, 25 (4.17 on average, or 93%), were deemed critical to the indexing of the article and for *ISME J*, 8 of the 11 added terms (1.33 on average, or 73%), were deemed to be critical (see Figure 9, top panel). These results show that even though the number of MeSH terms added was not significant in comparison to the total number of terms indexed, a significant number of the terms added were deemed to be critical to the indexing of the article. These results provide important evidence that the indexers are a vital component of the MTIFL-method of indexing, performing a necessary step in the indexing process.

To evaluate why indexers added additional MeSH terms, the Senior Indexers grouped reasons added into five categories:

- *Point of Article* – Necessary for indexing
- *Title Concept* – Necessary for indexing
- *Indexing Policy* – Coordination and subheadings
- *Can Be Indexed* – Not the point but can be used
- *Found in Full-Text* – Found in other than title and abstract

Figure 9 (bottom panel) shows the categorical representation for the deleted terms in a percentage. For *J Appl Microbiol* most (41%) of the MeSH terms were added because those concepts were *Found In Full-Text* of the article and another 35% of the MeSH terms were added because they were the *Point of Article*. For the *ISME J* articles evaluated, MeSH terms were found to be added because they were deemed as a *Title Concept* (75%) or because they were *Found in the Full-Text* (25%) of the article. Again, these results provide evidence that indexers serve a critical role in indexing of MTIFL-indexed articles. Many of the added MeSH terms come from scanning the entire full-text of the article.



**Figure 9.** Top panel displays the average number of MeSH terms added by indexers (dark bar) and deemed critical (light bars) by Senior Indexers and the bottom panels displays the categorical representation on why the MeSH terms were added by the indexers.

## Indexer Ratings

The final qualitative approach assessed looked to determine if a quick rating of MeSH and IM terms could be used to evaluate indexing. Four NLM indexers were given a semi-random set of articles and asked to rate how satisfied they were with the MeSH and IM terms indexed, ranging from *Strongly Agree* (1) to *Strongly Disagree* (5) (see Figure 1). For *J Appl Microbiol*, the MeSH and IM terms were rated slightly better (1 - *Strongly Agree* to 2 - *Somewhat Agree*) for MTIFL-indexed articles (1.58 and 1.54, respectively) compared to the human-indexed articles (2.29 and 2.38). These results were found to be significantly different;  $t(46) = 2.1704, p=0.0352$ , for MeSH term ratings, and  $t(46) = 2.4949, p=0.0163$ , for IM ratings (see Table 8).

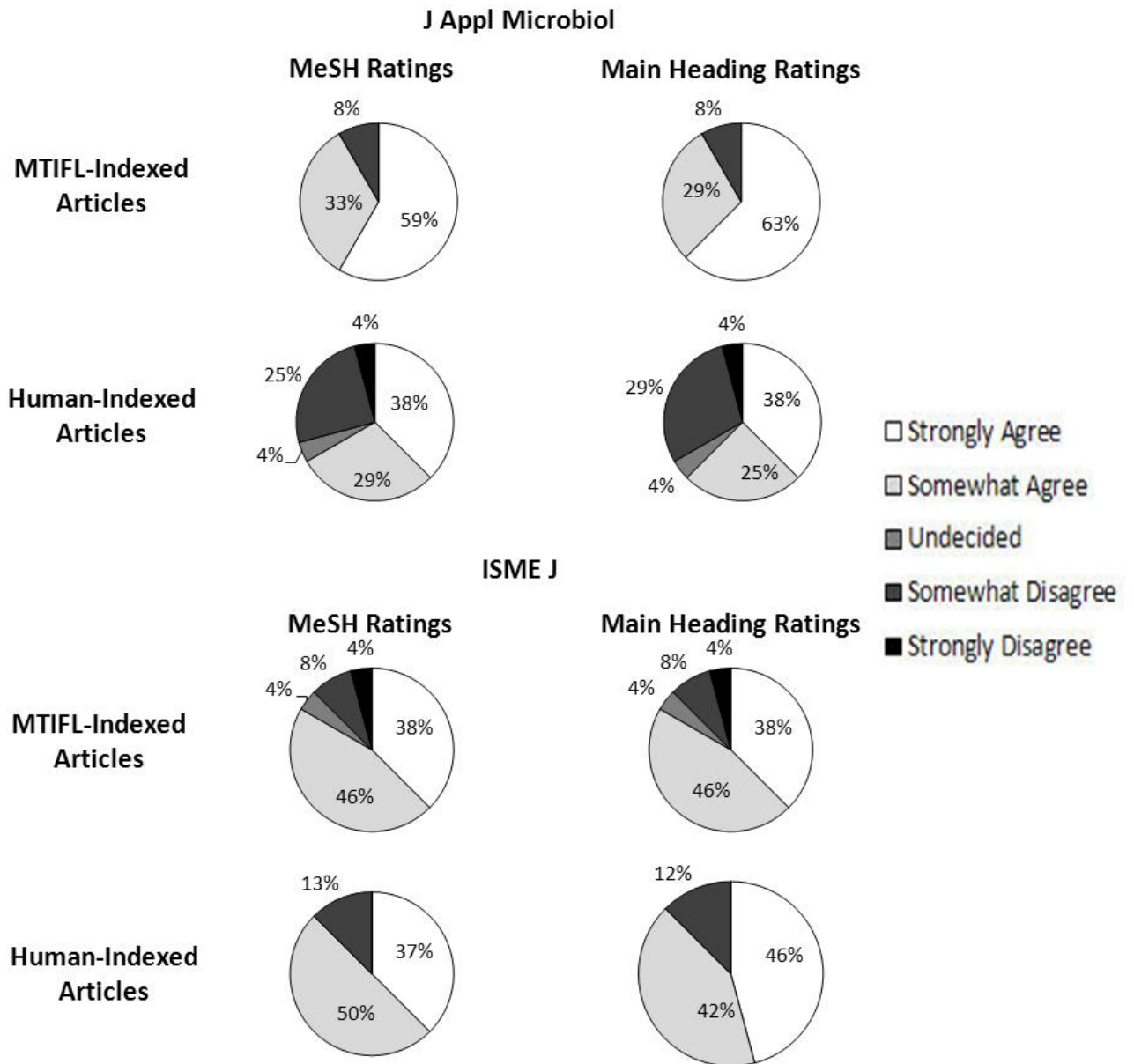
For *ISME J*, the MeSH and IM term ratings for the two different modes of indexing were not found to be as different, when rated by indexers. MTIFL indexing was rated 1.96 (MeSH) and 1.96 (IM), whereas human indexing was rated 1.88 (MeSH) and 1.79 (IM). No significant differences in ratings were found for MeSH term ratings  $t(46) = 0.2838, p=0.7778$ ; or IM ratings  $t(46) = 0.5599, p=0.5783$  (see Table 8). These results show that the MeSH and IM terms indexed by both the human indexers and the MTIFL-assisted methods were rated satisfactory in indexing the articles.

Journal	Question	Index Method	Avg Rating	t	df	p	sig
J Appl Microbiol	1 - MeSH	Human	2.29	2.1704	46	0.0352	Yes
	1 - MeSH	MTIFL	1.58				
	2 - IM	Human	2.38	2.4949	46	0.0163	Yes
	2 - IM	MTIFL	1.54				
ISME J	1 - MeSH	Human	1.88	0.2838	46	0.7778	ns
	1 - MeSH	MTIFL	1.96				
	2 - IM	Human	1.79	0.5599	46	0.5783	ns
	2 - IM	MTIFL	1.96				

**Table 8.** Average MeSH and IM term ratings and t-test results for the 12 *J Appl Microbiol* (6 human and 6 MTIFL-indexed) and 12 *ISMEJ* (6 human and 6 MTIFL-indexed) articles, as rated by four NLM Indexers.

Figure 10 displays the percent distribution of ratings for both journals, and for both methods of indexing. Overall, the indexers rated the MeSH and IM terms (for both human and MTIFL-assisted indexing methods) as *Strongly Agree* to *Somewhat Agree* more often. Percentage-wise, both methods of indexing were rated on the positive side of the continuum, and represented greater than 60% of the ratings. Interestingly, for the *J Appl Microbiol* human-indexed articles, indexers *Strongly Disagreed* with MeSH and IM term indexing approximately 4% of the time. For the *ISMEJ* MTIFL-indexed articles indexers *Strongly Disagreed* with MeSH and IM term indexing about 4% of the time. However, because both methods, on all measures, rated more on the positive side of the continuum, we can say the indexing of these articles, regardless of the method of indexing, was deemed as sufficient. In fact, for the MTIFL and human methods of indexing, *ISME J* indexing was rated as more satisfactory approximately 80% of the time. For the MTIFL method on *J Appl Microbiol*, the indexing was rated over 90% as positive.

After completion of the Indexer Rating Survey it was discovered that the SCR terms were accidentally omitted from the articles needing to be rated. The SCR terms were then pulled for the articles (6 of the 24 articles had SCR terms) and the indexers were asked if they would change their original ratings based on the new information about indexed SCR terms. Few changes were made to the original ratings for the indexed MeSH and/or IM terms, therefore, the final results still stand as presented above.



**Figure 10.** MeSH (charts on left) and IM (charts on right) term rating data for *J Appl Microbiol* (top panel) and *ISME J* (lower panel) for MTIFL-indexed (top portion of each journal panel) and human-indexed (bottom portion of each journal panel) articles.

## Discussion

We tested three different approaches to determine a method that could be suggested for the evaluation of MTIFL indexing. While the sample sizes used in these tests were too small to complete an evaluation of MTIFL, the sample sizes were appropriate for an assessment of the approach. That is, generalizations about the conclusions that could be generated from these methods can be made based on these initial assessments. Overall, none of the three approaches tested were found to be appropriate because each approach had positive and negative aspects.

### Evaluation of the Lister Hill Center Statistics Method

The statistics generated by LHC can be used to monitor MTIFL-indexed MeSH and IM terms and can be used to provide continuous feedback to the MTI program, creating a better tool that can be used for indexing assistance. These statistics can be generated fairly quickly and are relatively easy to understand and interpret. However, the statistics do not give information about the quality of the final indexing. The quantitative results are based on the human indexer making changes during the indexing process, which may or may not reflect good indexing outcomes. However, as a method for MTIFL evaluation, this quantitative method supplies high quality quantitative data such as counts and explanations for MeSH terms that are being added and/or deleted from the MTIFL indexing and has potential if used in combination with other methods of evaluation.

### Evaluation of the Senior Indexer Analysis Method

Similar to the sample size utilized in the LHC statistical analyses, the sample size (n=6 articles for each of the two journals, N=12 articles total) was on the smaller side. However, we were able to see statistical significance in a number of analyses, so generalizations about the usability of this method for the evaluation of the MTIFL indexing could be made. First, the biggest disadvantage for this method was the labor-intensive procedure. The method utilizes the expertise of one or more senior indexers (i.e., for consistent data), who are well versed with the policies and rules of NLM indexing and who are knowledgeable about the biomedical field being indexed. Because of this, taking senior indexer time and effort away from regular duties to analyze completed indexed articles may not always be a practical or feasible option. In addition to the senior indexers doing the critical analysis of the MeSH terms, another person versed in statistical methods and evaluation is needed to compile, analyze and interpret the data. For this pilot study, two Senior Indexers took the time to analyze 12 articles, and the Associate Fellow analyzed and interpreted the data. Thus, this method may not always be a feasible option to which the Index Section can commit time, effort and money.

Despite all of the constraints with this method, some of the evaluative data resulting from the critical analysis was beneficial. This approach provided excellent in-depth analysis for MTIFL indexing (i.e., details about applying MeSH terms throughout the entire indexing process). For example, even though the number of MeSH terms added/deleted compared to the final number of MeSH terms indexed for a particular article was not always statistically significant, the number

of critical MeSH terms (as deemed by the in-depth Senior Indexer Evaluation approach) was found to be significant for the indexing quality of the article. Additionally, the details from the categories about why MeSH terms were added or deleted (e.g., *Too Tangential*, *3<sup>rd</sup> Tier Term*, *Too Specific/General*) can lead to better filters and triggers for LHC staff to incorporate into the software for better automated assistance. Therefore, this approach, although labor-intensive, was found to be extremely beneficial to the evaluation of the MTIFL indexing.

### Evaluation of the Indexer Rating Method

The method of rating indexed MeSH and IM terms, as conducted by NLM indexers, holds some promise as a method for evaluating indexing. The survey rating model takes minimal indexers' effort (although, indexed citations will need to be rated by multiple indexers so reliable rating data can be generated). This method might be good for quick and continuous indexing evaluation. The evaluative results are not rich with details about the MeSH and IM terms indexed, but the method could supply information on whether an article is indexed well enough, or needs further attention. Any method of indexing (traditional or MTIFL) could be evaluated via this rating method. Mostly, this approach has promise for expansion with automated methods for sampling, collecting, and analysis. If this method were combined with other methods of evaluation, potentially powerful analyses for the evaluation of indexing could be completed.

## Conclusions

After completion of the testing and assessing of these quantitative and qualitative approaches, it is now clear what the ideal model for the evaluation of MTIFL should:

- Supply both quantitative and qualitative data.
- Lend itself to straightforward data collection.
- Permit easy analysis.
- Be feasible.

The workload of the indexers in the Index Section is already extensive; an evaluation model for MTIFL must be practical (i.e., integrated with the indexers workflow) if it is to be justifiable (i.e., efforts, costs and time) and truly useful.

One option that could be pursued for indexing evaluation is a combined approach of two or more of the tested methods. As indicated, the statistics from the LHC provided quick and efficient information about MTIFL indexing, including details about the MeSH terms that were ultimately deleted or added by the indexers. These data provided not only quick information about the indexing process of the MTIFL journals, but also about the level of confidence by continuously checking and evaluating the F-scores. Combining the LHC statistics with the Indexer Ratings

could give a more complete picture of the indexing process. This type of continuous, random rating could lead to a system of flagging journals that may need special attention for review.

In pursuit of the goal for access to biomedical information, the National Library of Medicine has been at the forefront of the collection, organization and dissemination of the world's biomedical literature. In order to maintain its' high-level of integrity, new approaches to the evaluation of MTIFL indexing need to be considered. The pilot studies reported here have proved to be a good starting point into the assessment of possible evaluation models that can ensure continued excellence in MEDLINE indexing.



## **Acknowledgements**

I would like to thank my project sponsors, Rebecca Stanger and Olga Printseva, for submitting this project and for allowing me to take liberties to expand beyond the original proposed project idea. I want to thank them for their assistance and advice throughout the project. I send many thanks to Jim Mork for statistics and assistance with data interpretation and to Carol Fisher, Mary Hantzes, Eduardo Tello and Janice Ward who assisted in the rating of MeSH terms. And a great big thank you to Lou Knecht and Rebecca Stanger for taking the time to edit and help revise the project presentation and report.

I am also grateful to Dr. Donald A.B. Lindberg, Betsy Humphreys and the National Library of Medicine for continued support of the Associate Fellowship Program. And I would like to thank Sheldon Kotzin, Joyce Backus for support throughout the fellowship year.

Finally, I would like to thank Lou Knecht and Torri Kellough for their mentorship and Kathel Dunn for her guidance and support. To my fellow 2011-2012 Associate Fellows, Bethany Harris, Jessi Van Der Volgen and Michele Mason-Coles, for comradeship throughout the Fellowship year. And last but not least, a big thank you to my husband, Ryan Spaulding.

## References

1. Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM Indexing Initiative's Medical Text Indexer. *Stud Health Technol Inform.* 2004;107(Pt 1):268-72.
2. Aronson AR, Mork JG, Lang F-M, Rogers WJ, Jimeno-Yepes AJ, Sticco JC. The NLM Indexing Initiative: Current status and role in improving access to biomedical information: Report to the Board of Scientific Counselors, National Library of Medicine, Bethesda, MD, April 5, 2012.
3. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp.* 2001:17-21.
4. Aronson AR, Mork JG, Lang FM, Rogers WJ, Neveol A. NLM Medical Text Indexer: A tool for automatic and assisted indexing. US National Library of Medicine, LHCNBC, Communications LHNCfB; 2008. Report No.: LHCNBC-TR-2008-002.
5. Aronson AR, Bodenreider O, Demner-Fushman D, Fung KW, Lee VK, Mork JG, Neveol A, Peters L, Rogers WJ. From indexing the biomedical literature to coding clinical text: Experience with MTI and machine learning approaches. *Proceedings of the ACL '2007 Workshop "BioNLP"*. 2007:105-12.

## Appendix A.

Raw data for statistics used in journal analyses.

Indexing	#Articles	#MeSH	#IM	#SCR	Avg MeSH	MeSH Stdev	Avg IM	IM Stdev
<b>Arch Microbiol</b>								
Indexer (Issue 1)	7	85	32	30	12.14	5.27	4.57	0.79
Indexer (Issue 2)	5	63	22	30	12.60	4.88	4.40	1.14
Indexer (Issue 3)	7	75	28	34	10.71	4.07	4.00	1.00
MTIFL Method (Issue 1)	7	84	26	34	12.00	1.91	3.71	1.11
MTIFL Method (Issue 2)	7	60	25	26	8.57	3.51	3.57	1.51
MTIFL Method (Issue 3)	8	98	34	44	12.25	3.11	4.25	0.71
<b>Can J Microbiol</b>								
Indexer (Issue 1)	14	200	66	63	14.29	2.89	4.71	1.54
Indexer (Issue 2)	16	153	70	48	9.56	3.24	4.38	1.20
Indexer (Issue 3)	9	108	34	26	12.00	3.67	3.78	0.67
MTIFL Method (Issue 1)	13	149	48	74	11.46	3.60	3.69	0.75
MTIFL Method (Issue 2)	10	133	43	27	9.57	3.20	4.57	0.67
MTIFL Method (Issue 3)	7	67	32	12	11.92	2.51	3.85	1.27
<b>ISME J</b>								
Indexer (Issue 1)	10	132	42	41	13.20	5.37	4.20	1.14
Indexer (Issue 2)	10	129	48	28	12.90	4.09	4.80	1.69
Indexer (Issue 3)	9	77	34	15	8.56	3.00	3.78	0.83
MTIFL Method (Issue 1)	16	146	68	38	9.13	2.47	4.25	1.61
MTIFL Method (Issue 2)	14	131	53	38	9.36	1.45	3.79	1.19
MTIFL Method (Issue 3)	12	112	42	31	9.33	2.10	3.50	1.17
<b>J Appl Microbiol</b>								
Indexer (Issue 1)	36	441	164	127	12.25	3.06	4.56	1.34
Indexer (Issue 2)	37	485	173	146	13.11	3.30	4.68	1.47
Indexer (Issue 3)	37	462	152	143	12.49	3.32	4.11	1.20
MTIFL Method (Issue 1)	27	344	115	101	12.74	2.96	4.26	1.02
MTIFL Method (Issue 2)	24	283	104	109	11.79	2.40	4.33	0.87
MTIFL Method (Issue 3)	25	282	110	82	11.28	3.14	4.40	1.00

# Appendix B.

## A description of MTI statistics.

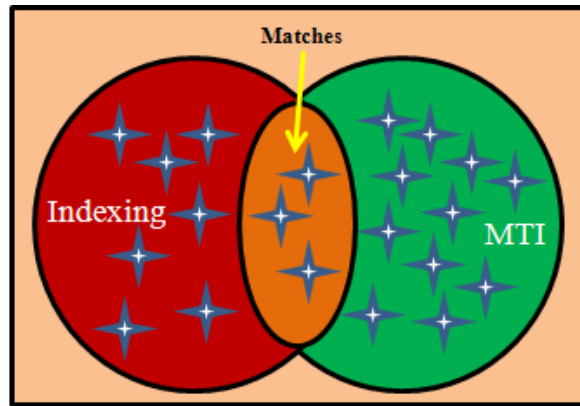


Figure 1 - Example for detailing Recall, Precision, and F2 calculations

The following table illustrates how to compute the Recall, Precision, and F<sub>2</sub> measure the way we do for the MTI statistics using the information in *Figure 1*.

<b>Number of Indexing MHs</b>	10
<b>Number of MTI Recommendations</b>	14
<b>Number of Matches</b>	3
<b>Recall</b>	$\frac{\text{Number Matched}}{\text{Number Indexing MHs}} = \frac{3}{10} = 0.3 \text{ (30.00\%)}$
<b>Precision</b>	$\frac{\text{Number Matched}}{\text{Number MTI Recommendations}} = \frac{3}{14} = .2143 \text{ (21.43\%)}$
<b>F<sub>2</sub> Measure</b>	$\frac{5 \cdot \text{Precision} \cdot \text{Recall}}{(4 \cdot \text{Precision}) + \text{Recall}} = \frac{5 \cdot .2143 \cdot .3}{(4 \cdot .2143) + .3} = .2778 \text{ (27.78\%)}$