A MEDLINE Indexing Experiment Using Terms Suggested by MTI

A Report

National Library of Medicine
Indexing Initiative Project

Recent advances in the quality of indexing achieved by the Medical Text Indexer, MTI and the work towards integrating that system with DCMS have suggested the need to perform an evaluation of that indexing by MEDLINE indexers. MTI is a project of the NLM's Indexing Initiative. The experiment described here constitutes an assessment of whether indexing terms suggested by MTI would facilitate the work of the indexers. We sometimes refer to this list of subject headings found by MTI as MTI results.

This experiment was conducted by the core team of the NLM's Indexing Initiative with the assistance of volunteer indexers from the Bibliographic Services Division (BSD), and technical support from the DCMS development team in the Office of Computer and Communications Systems. DCMS is the Citation Capture and Maintenance System used daily by the MEDLINE indexers.

This document is a report on the experiment that presents the information present in the questionnaires and indexing. The following sections describe the design of the experiment, how the experiment was conducted with some analysis of the questionnaires and of the indexing itself.

1.0 Experimental Design

We measured the usefulness of the MTI suggested indexing terms by presenting such terms to indexers actually performing indexing. This section describes the main components of the design of the experiment.

Scope of the experiment. For the initial experiment, we followed the model of the 1997 evaluation and kept the number of articles to be indexed and the number of indexers participating in the experiment relatively modest.

Indexer participation. Because the experiment does add to the effort of indexing, experienced indexers who have already volunteered to support the Indexing Initiative were used. Since the advisory design of the experiment minimizes the intellectual effort or distraction of participation, it should be possible to add other indexers including contractors in later, expanded experiments.

Journal selection. In order to keep the experiment as simple as possible, each indexer participating in the experiment indexed all articles from a single issue of a journal on a subject within his/her area of expertise. The indexers were asked to select the journal issue used in the experiment. A balance between clinical and pre-clinical articles was attempted.

MTI Processing. The articles from the selected journal issues were retrieved from MEDLINE, the title and abstract were processed by MTI. The MTI processing used the default parameters and did not use any of the recently introduced tweaks.

Indexing and Evaluation. Using the normal, production version of DCMS, the indexers were instructed to open the Related Records pane for each of the articles in their assigned journal. The indexers then completed their indexing while selecting the subject headings they saw as appropriate for the article from the Indexing Initiative window. At the completion of the indexing for the article the indexer was presented with a window that asked them to evaluate some suggested subject headings that they had not chosen. At the completion of all the articles in the assigned journal, each indexer was asked to complete a short survey on the general utility of the suggested subject headings.

Analysis. The last step in the experiment consists of both a technical analysis comparing the MTI-suggested indexing terms with those actually chosen by the indexers and also an informal analysis of the completed surveys. The technical analysis provides an objective evaluation measure. The informal analysis attempts to assess user satisfaction with the experimental indexing process and identify areas for improvement of MTI performance. The analysis of rejected terms provides a very specific opportunity for improving the utility of MTI results.

2.0 Conducting the Experiment

As a pre-test for the system we ran a trial of the objective analysis on the MTI results and the indexer selected subject headings from a previous issue of some journals that were recently indexed by some of our volunteer indexers. This analysis was used to select the parameters and tweaks to use in the experiment.

Following a kick-off meeting on March 14, the experiment was conducted from March 19 through April11, 2002. Instructions provided to the volunteer indexers at the kick-off meeting can be found in Appendix A.This section provides some details on the experimental procedure outlined above.

2.1 Indexers and Journals

The volunteer indexers are the following:

- David Cissel (Technical Information Specialist, BSD)
- Monika Devay (Technical Information Specialist, BSD)
- Mary Hantzes (Technical Information Specialist, BSD)
- Min chi Huang (Technical Information Specialist, BSD)
- Esther Lawrence (Chemist, BSD)
- Rebecca Stanger (Technical Information Specialist, BSD)
- Joe Thomas (Technical Information Specialist, BSD)
- Dorothy Trinh (Technical Information Specialist, BSD)
- Janice Ward (Technical Information Specialist, BSD)
- Melissa Yorks (Technical Information Specialist, BSD)

The indexers range in experience from less than two to more than twenty years.

The selected ten journals have been grouped by the type of articles that they usually contain. The particular issue is indicated by the volume and issue number.

- Pre-Clinical

  Biochemical and Biophysical Research Communications 291(3)

  The Journal of General Virology 83(Pt 3)

  European Journal of Pharmacology 435(2-3)

  Nature 415(6875)

  Science 295(5561)

- Clinical

  International Journal of Gynaecology and Obstetrics 76(2)

  Clinical Genetics 60(6)

  Pediatrics 109(3)

  Archive of Surgery 137(3)

  Japanese Circulation Journal 65(12)

2.2 MTI Processing

The MTI used the default selection of parameters. A selection of post-processing enhancements (tweaks) to MTI were considered but not used for creation of the suggested indexing lists.

The tuning of MTI parameters and tweak selection was conducted prior to the experimental processing by the indexing and precision evaluation of 200 articles from clinical and pre-clinical journals from 2000. (This same test collection was used to select the default parameters for MTI.) This process considered whether a selected ranking score threshold or a term count limit would be used to determine which terms identified by

MTI were presented to indexers in the experiment.We compared tweak 2 and default parameters for suggestion lists of various lengths by calculating precision, recall, and *F* measure. After review of these statistics we selected the list of the top 25 terms from default processing. These settings emphasized recall while maintaining reasonable precision. (α for the *F* measure was 0.2). This selection was not made on the absolute highest *F* measure score, because the differences between the resulting scores were not statistically significant. After a comparison of tweak 2 and the default output, the BSD liaison also selected the longer default output.

2.3 The Evaluation System

With the goals of minimizing the impact of the evaluation on indexing production and realistically approximating production use of the MTI results, the experimental system is an extension of the production DCMS. DCMS, the Citation Capture and Maintenance System, is the primary tool used by MEDLINE indexers to add MeSH terms as subject heading and qualifiers to a MEDLINE record.

In 1997 when indexers were presented with suggestions as already selected terms, they worked to correct the indexing much as they would revise a novice indexer. This placed an additional intellectual burden on the experimental indexers and did not allow the use of the suggested terms in an advisory way. Therefore, in this experiment, indexers were allowed to select a pane that contains a list of suggested subject headings from which they could choose those that looked promising.

This section describes the modifications made to DCMS to support this experiment, the details on the presentation of the suggested subject headings, and the follow up questionnaires.

2.3.1 DCMS Modifications

The Rel Record tab is one of the collection of tab panels appearing on the right side of the main indexing screen in DCMS. The default source of data for this pane was changed to MTI and the pane was labelled “Indexing Initiative” because it is the search method. The new panel presents the data described in the next section. When an indexer selects a subject heading it is added to the standard indexing pane, but not in the quick-edit mode. The indexers had normal access to the auto-completion term pane, and the other DCMS provided indexing tools. They were able to return to this Indexing Initiative pane at anytime while indexing the article.

In addition to this main interface change, DCMS also supports an interface with a Indexing Initiative created evaluation module. This interface includes the following interactions:

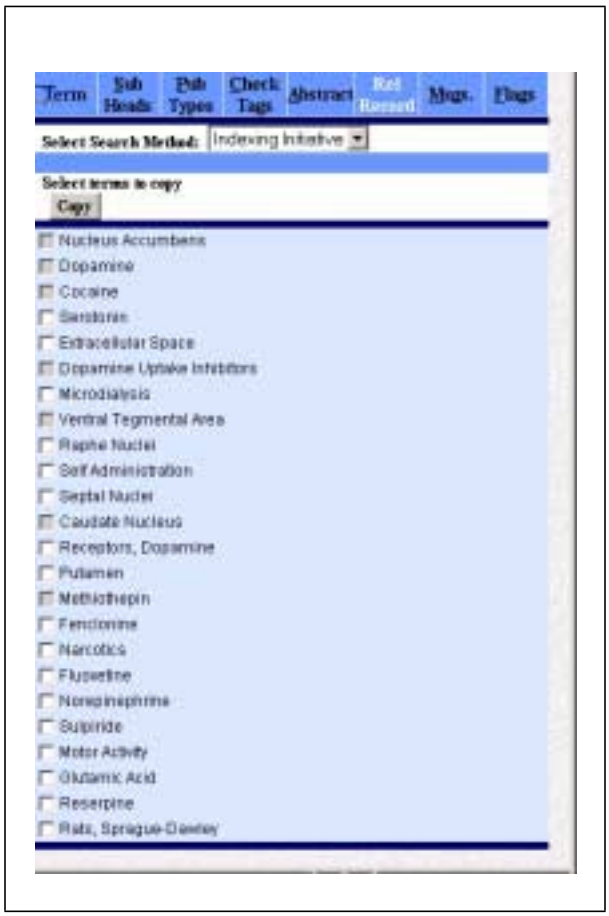
1. When the indexer selects the *Rel Record* tab, DCMS sends the PMID to the evaluation module at a predefined URL. The evaluation module returns the MTI results as a set of pipe delimited records that contain the MeSH terms (subject headings). DCMS presents this data in the Indexing Initiative pane as described below.
2. When the indexer completes an article, DCMS notifies the evaluation module and presents the Article Questionnaire form in a new browser window. The notification includes the list of terms selected from the Indexing Initiative pane and the full set of indexing records.

3. When all the articles for the journal have been indexed, DCMS notifies the evaluation module and presents of the Final Indexing Initiative Questionnaire form.(These forms are described in later sections.)

2.3.2 Subject Heading Presentation

The Figure 1 shows an example of the modified *Rel Record* tab in DCMS.

FIGURE 1. Sample of the MTI Suggested Terms Presentation



2.3.3 Article Questionnaire

The Article Questionnaire presented at the completion of each article has two parts: Some initial questions assess MTI's overall performance on the article. The second part is a reject questionnaire to find out why some of the suggested terms were not used.

Initial Questions. Introductory text will be at the top of each questionnaire followed by a request and a multiple choice question:

> If the list of suggested subject headings made you think of some new conceptual idea to use in your indexing, please explain here. (text box)

Did the list of suggested subject headings cover the purpose of the article?

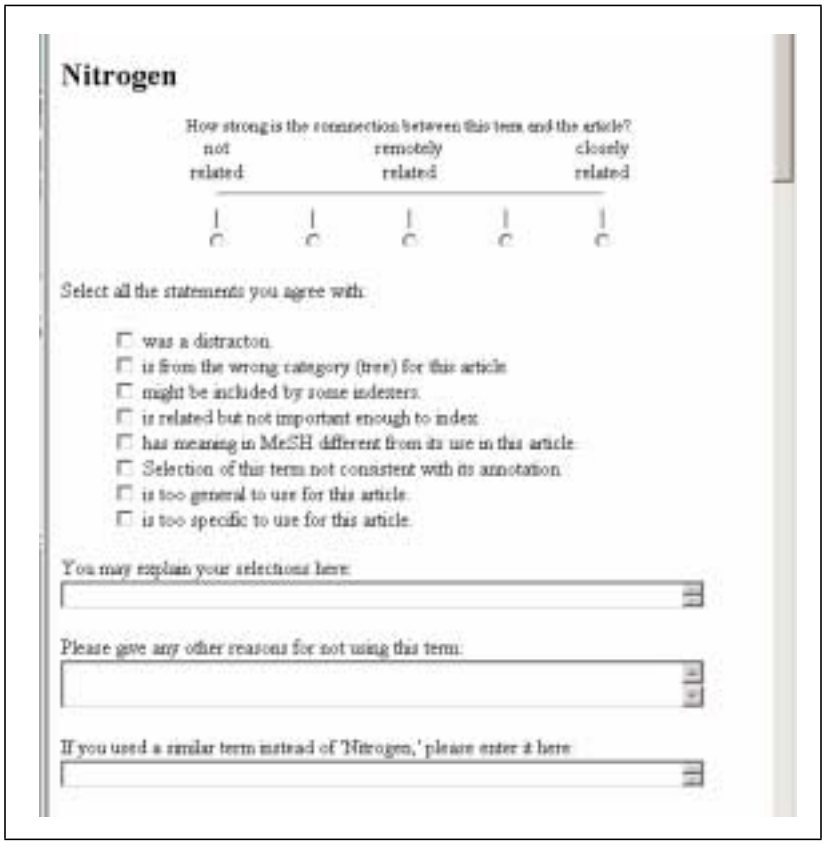
The choices for this question are “Yes,” “Partially,” and “No.”

Reject Questionnaire. When indexers select terms to use from the Indexing Initiative pane, we know that they found them useful. However, we know very little about why the other terms MTI selected were not useful. Therefore, at the completion of the indexing for an article the evaluation module constructed a questionnaire about the terms not selected (rejects).

Our inspiration for aspects of the rejected terms to ask the indexer about was the comments of the indexers in the previous experiment. Their concerns were focused on accuracy, specificity, and whether the term was related or not.

The evaluation module will compare the initial suggested list to the list of terms selected by the indexer and produce a section of the Article Questionnaire that asks about that term. Figure 2 shows a sample.

FIGURE 2. Sample Reject Form



2.3.4 Final Survey

The purpose of the final questionnaire is to assess the indexers' satisfaction with the MTI results. A series of questions were asked. Some questions were repeated from the 1997 experiment to facilitate comparison(1, 3, 12, 17, 18, 19). Some were answered by entering a number that represents a location on a eleven-point Likert scale:

0-Very Strongly Disagree, 5-Neither Disagree Not Agree, 10- Very Strongly Agree

The Likert scale questions are organized around several aspects of the MTI system and the DCMS interface used in this experiment. They address these categories of performance and usability:

- Judgemental Performance: Overall Quality (13 - 5),
  Term Quality (- 7),
  Completeness (16 - 8);
  Helpful (1 -)
- Opinion on Usability: Confidence (9 - 15),
  Ease of use (- 10)
  List size (11- 2),
  Explanation (- 3);
- Organizational Impact: Workload (4 -),
  Efficiency(6 - 12),
  Novice users(14 -).

Some statements express the quality in a positive way; some negatively. The first number listed in each pair is a positive statement and the second is negative.

Accompanying each statement and its Likert scale a Comments text box was provided.The individual statements and the final four open-ended questions are given in full below in the Results section. The full Final Indexing Initiative Questionnaire is in Appendix B.

3.0 Results

There was extensive data collected during this experiment. In this section we report and summarize how well MTI did with indexing the selected articles, and how well it supported the indexers. First we present an overview description of the information collected and then in subsequent sections report the findings.

3.1 Overview of Information Collected

There were four main collection points during the experiment. First the MTI indexing was stored in a database to support the supply of that indexing to DCMS and for the later comparison with the MEDLINE indexing. The second collection point was the completion of the indexing of an article. At that point we collected and stored the current main headings and check tags added by the indexer. Third, when the indexer completed the Article Questionnaire, we stored that data. The fourth and final collection

point during the experiment was at the completion of the last article in the journal. In addition to a questionnaire for that article, the indexers completed the Final Indexing Initiative Questionnaire and the responses were collected. (Due to web browser problems, two indexers had to respond a second time to the Final Questionnaire on paper. Their responses were added to the other online data.)

The final indexing data was also collected from MEDLINE through PubMed after the release of the indexing. Comparison of that indexing to the MTI indexing is reported in the Section 3.2, "Objective Measures," on page 10.

We present this information from the most general to the most specific. The Final Questionnaire ends with four open-ended questions.We present highlights and then these comments in their entirety. The first 16 statements of the Final Questionnaire were scored by the indexers to indicate their agreement or disagreement. These are summarized and the full responses are in Appendix B. The comments that accompanied the statement scores are available in Appendix C.

The Article Questionnaire asks the indexers for an overall evaluation of the suggested terms for that article. These data are presented in the Section 3.5, "Article Performance Evaluation," on page 18. The remaining sections of the Article Questionnaire evaluate terms suggested by MTI but rejected by the Indexer. These term evaluations are presented in Section 3.6, "Reject Evaluation," on page 20.

Table 1 shows the quantities of each type of information collected.

TABLE 1. Collected Data Summary

Data Category	n
Articles with indexing to compare:	273
Recommended Terms with feedback:	809
Overall Article evaluations:	161
MTI impressions:(Final Questionnaire)	10

Note that the number of articles evaluated are significantly less than the number with indexing. The responses from many of the article questionnaires were not collected due to technical problems during the experiment.

3.1.1 MTI Impressions

One indexer did not respond to the final four questions, but provided comments on the earlier statements. A couple others left one or two blank.

The three strongest themes of their responses were

- Concern about misleading terms.
  "I found myself considering terms that I ordinarily wouldn't have indexed."
- Terms were too general
  "[Did not like] Mostly the lack of specificity,"

- Input entry terms not recognized
  "Entry terms were not always included in the list..."

One positive theme was

- Saves typing time.
  "clicking on terms instead of typing them saved time"

3.1.2 Scored Statements

The indexers rarely agreed or disagreed strongly with the statements provided. They also were diverse in their opinions. They agreed enough to have significant opinions on two topics out of the eleven covered by the sixteen questions: List Size and Completeness.

List Size. With a average score 2.2 ($\sigma = 2.48$) This is about half way between *Strongly Disagree* and *Neither Disagree Nor Agree*, indicating they did not want a longer list, but preferred a shorter one.

11. I would rather see a longer list of suggested terms. (Average score 1.4, $\sigma = 1.9$)

2. For most articles the list of suggested subject heading was too long. (7.0, $\sigma = 2.72$)

Completeness: Although they did not interpret the pair of completeness questions as opposites, the indexer volunteers were in agreement on statement 8.

8. Important subject areas were sometimes missing from the list of suggested headings. Average score = 7.4 ($\sigma = 2.25$)

This result is tempered by the response to 16, which was more dispersed.

16. MTI coverage of significant topics was good. (Average score = 5.7, $\sigma = 2.37$)

Confidence. The indexers agreed most on their view of their confidence in the MTI results. They mildly disagreed, expressing a weakness of confidence in the accuracy of the subject heading recommended by MTI. with an average score of 3.9. ($\sigma = 1.37$, randge2-6)

Overall. The indexers were generally inclined to select a neutral position, They disagreed (assigned lower scores) more frequently to the positive statements and, to a lesser degree, agreed more often with negative statements.

3.1.3 Article Performance Evaluation

The overall performance of MTI on each article was assessed with one multiple choice question and one open-ended question. In general the indexers thought that the MTI terms usually covered the purpose of the article. On only 5% of the articles did the Indexers report that the list of suggested headings made them think of some new conceptual idea to use in their indexing.

3.1.4 Reject Evaluation

The reject evaluation included a score for the relatedness of the rejected term, selection of reasons for its rejection, comments on selections made, indexer provided reasons, and reporting of terms used in place of the suggested term.

The initial rating of the top five terms not used by the indexers (rejected) was nearly evenly split between the three labeled categories: *not related, remotely related* and *closely related*.

Table 2 shows how often the descriptive statements shown in Figure 2, "Sample Reject Form," on page 6, were applied to the reject terms. The significance of these evaluations is enhanced by the awareness that only the top five rejects were evaluated. Probably 15 to 20 lower ranked MTI suggested terms were not individually evaluated.

TABLE 2. Frequency of Statement Assignments

Statement	Percent	Statement	Percent
Distraction	16	MeSH Different	2.1
Wrong Category	4.5	Annotation	1.1
Another Might	5.2	Too General	26
Not Important	11	Too Specific	4

3.2 Objective Measures

The objective measure of MTI performance is how well its suggested terms covered the indexing of the volunteers. We will assess this with Precision and Recall as well as a simple count of terms from the list used for each article.

3.2.1 Indexing

There were 324 articles in the selected journals. Some articles were deleted from PubMed and others were considered out of scope and not indexed. The remaining 272 articles are the corpus for our comparison. Table 1 shows the totals for all the journals. The "Used Headings" refer to the headings that were suggested by MTI and appear in the MEDLINE indexing. The *Used Headings* numbers include all headings marked MH in MEDLINE. They are both IM and NIM and include checktags. For this comparison subheadings were not considered although they did sometimes appear in the MTI list of recommended terms.

TABLE 3. Indexing Totals

Category of Subject Headings	Count
Suggested Headings:	7216
Used Headings:	2099
MEDLINE Headings:	3851
Used IM Headings:	814
MEDLINE IM Headings:	1001

3.2.2 Precision and Recall

The degree to which the list of suggested subject headings match the terms chosen by the indexers can be estimated by computing the precision and recall of the one to the

other. These values for each article can be averaged over the entire journal issue and for the complete experiment.

The precision and recall ware calculated first for all the suggested terms compared to all the MEDLINE terms. The second set of values are calculated for only the MEDLINE Index Medicus terms: those that appear in MEDLINE with an asterisk. Since MTI does not designate which terms it thinks are IM, the precision value, which was based on the full list of suggested terms, is very low. The precision and recall values were calculated for each article and the averages for the entire experiment appear in Table 4.

TABLE 4. Indexing Precision and Recall

Statistic	Value	Standard Deviation
Precision:	0.29	0.12
Recall:	0.55	0.184
IM Precision:	0.11	0.058
IM Recall:	0.81	0.266

Precision is the ratio of the number correct suggested headings to the total suggested. Recall is the ratio of that same number of correct suggestions to the total number of headings in MEDLINE for that article.

An example may clarify this data. A sample article is the following:

> Aksnes G, Diseth TH, Helseth A, Edwin B, Stange M, Aafos G, Emblem R. Appendicostomy for antegrade enema: effects on somatic and psychosocial functioning in children with myelomeningocele. Pediatrics. 2002 Mar;109(3):484-9. PMID: 11875145

This article was index during the experiment 3/28/02 by one of the volunteer indexers. MTI suggested 27subject headings. Table 5 shows both lists. The indexer chose 18 terms with subheadings to index the article. The two sets of terms match on 11 headings(8 main headings and 3 checktags). This yields a precision (11/27) of 0.41, which is better than the average. There were 6 headings marked IM. Five of those IM terms were suggested by MTI. The precision for the IM terms is (5/27) 0.19. The recall for all terms is (11/18) 0.61. The recall for IM terms is then (5/6) 0.83. MTI recommended “Surgery” as a subheading that was correct, but not considered in the evaluation.

TABLE 5. Sample Article - PMID 11875145

MTI Suggested Headings	MEDLINE Indexing
Enema	*Enema/psychology
Meningomyelocele	Meningomyelocele/*complications/psychology/surgery
Fecal Incontinence	Fecal Incontinence/etiology/psychology/*surgery
Ostomy	
Appendix	Appendix/*surgery
Colostomy	Colostomy
Constipation	Constipation/etiology/surgery
Mental Health	
Social Adjustment	
Spinal Dysraphism	
Self Concept	*Self Concept
Urinary Incontinence	
Adaptation, Psychological	
Anus, Imperforate	
Stomas	Stomas
Epispadias	
Anus	
Mental Disorders	
Ileostomy	
Interpersonal Relations	
Hirschsprung Disease	
Sick Role	
Chronic Disease	
Child	Child
Human	Human
Adolescence	Adolescence
	Case-Control Studies *Enterostomy Female Laparoscopy Male Postoperative Complications

3.2.3 Availability of Terms

How many appropriate MeSH terms were provided for each article? The average number of all terms was 7.72.

How may IM terms were in the MTI list? The average number provided was 2.99.

3.3 MTI Impressions

These question were presented for indexer response at the end of the Final Indexing Initiative Questionnaire:

17. What did you like about the list of suggested subject headings?
18. What did you *not* like about the list of suggested subject headings?
19. Did using the MTI supplied terms change the way you indexed any of the articles? If yes, please list those terms and describe any changes.
20. Please share any suggestions you have for making the MTI provided subject headings more useful.

This section gives a summary and the full text of these responses.

3.3.1 Summary

Table 6 lists an idea and the number of times it was identified in the indexer responses to the four open-ended questions.

TABLE 6. Ideas from Indexers about MTI

Idea	N	Indexer Expressions
Concerned about misleading terms.	6	I found myself considering terms that I ordinarily wouldn't have indexed. The suggestions of nonspecific or incorrect headings could mislead an indexer.
Terms were too general	6	[Did not like] Mostly the lack of specificity, It included too many too general terms...
Input entry terms not recognized	4	Entry terms were not always included in the list. Also match the BXs/synonyms, so they don't appear to be missing.
Quick Edit	4	... let it function in quick edit.
Source of Terms	3	... MTI could indicate where their terms came from. ...why there were listed.

TABLE 6. Ideas from Indexers about MTI

Idea	N	Indexer Expressions
Saves typing time	3	clicking on terms instead of typing them saved time
Check Tags first	3	THE CHECKTAGS WERE NOT INCLUDED IN THE BEGINNING

There were other useful suggestion, that are reasonable to consider implementing.

- "Get rid of the HMs of chemtool terms?"
- "Make a link to file MeSH, indicate trees...[in display]"

Some were not enthusiastic; another acknowledged learning something new.

- "... it might be a good starting point for some terms, if you are CLUELESS...." (their ellipsis)
- "Yes, Mice, MRL lpr is a disease model for autoimmune diseases, which I did not know before."

3.3.2 Indexer Responses

17. What did you like about the list of suggested subject headings?

Would be easy to use (good format) and may save my typing time--if only the quality was better (and worth to use it)!!

Saved me keystrokes

I liked the fact that I could choose them as a group and copy in one step

Some were accurate, but so obviously required for indexing that one would not need the list (title headings)

I like prioritized list.

iT SEEMED TO FOLLOW THE LIST OF MAIN PURPOSE OF THE ARTICLE IN THE MAIN.

clicking on terms instead of typing them saved time.

Some of the time they were right "on target"!

(No response from 2 indexers.)

18. What did you not like about the list of suggested subject headings?

CONSISTENTLY TOO GENERAL--however, it might be a good starting point for some terms, if you are CLUELESS....

Needs some embellishments

It included too many too general terms and other terms that were not really related to that particular article. I found myself considering terms that I ordinarily wouldn't have indexed. Also I was tempted to index more overlapping terms

Mostly the lack of specificity, but also the inclusion of related but unnecessary headings. The suggestions of nonspecific or incorrect headings could mislead an indexer.

Missing important chemicals or subject headings

THE CHECKTAGS WERE NOT INCLUDED IN THE BEGINNING

All the wrong terms or terms that made me look throught the article wondering why there were listed. It was like revising rather than indexing.

Most of the time they were "way off"!

19. Did using the MTI supplied terms change the way you indexed any of the articles? If yes, please list those terms and describe any changes.

(Summary 4 no, 1 both ways, 2 yes to content, 1 yes to system use)

a couple?? times I added something after looking at list, but not absolutely essential stuff (so overall, the "help" was not worth it, in my opinion)

Yes, I usually work in quick edit and I found I was going back and forth a lot. I'm used to adding subheadings in quick edit and this was awkward for me because the terms transferred in the validated form so I had to go back to quick edit to add them and I found myselfsometimes forgetting to do that

no

Yes, Mice, MRL lpr is a disease model for autoimmune diseases, which I did not know before.

I would not let it change my way.

NO

No, although looking at one of the questionnaires reminded me of a term I had meant to add.

Yes, I did make some changes, unfortunately I cannot remember the particular terms.

20. Please share any suggestions you have for making the MTI provided subject headings more useful.

(There were nine responses on this request.)

Get rid of the HMs of chemtool terms? Also match the BXs/synonyms, so they don't appear to be missing IMPROVE the QUALITY of terms (without making term list longer) Perhaps ranking is unnecessary (once ranked list is initially generated), and present terms in alphabetical order?? (so I can easily find the terms I know I want, to save typing)

I have told you all of my ideas before. Make a link to file MeSH, indicate trees, let it function in quick edit. Show where the terms originate.

It would be helpful if a term from the list or a see reference added manually appeared as grayed out. It would also be helpful if you could transfer terms to quick edit. Als it would be helpful if MTI could indicate where their terms ccame from.

Many of the suggestions were from introductory material which was not part of the article. Elimination of nonspecific terms, for instance DIABETES in an article about TYPE II diabetes. Entry terms were not always included in the list, for instance, BIOLOGICAL OSCILLATORS is an entry term to BIOLOGICAL CLOCKS. The latter was on the list, but BIOLOGICAL OSCILLATORS was not, bud the article called itoscillators.

1. If MTI finds the specific terms, either chemicals or subject headings, general terms can be dropped out of list. 2. Suggested check tags can have a higher priority instead of in the bottom of thelist.

I dont have any suggestions today, but having a discussion with the articles and indexing in hand, may be useful.

SORRY, i AM NOT SURE.

Make the list of suggested terms shorter. Most of the terms I chose came from the top of the list. If it would find all the check tags, correctly, it would be a big help but it was really bad at this.

Please, list the TITLE-words, not the verbs, but the NOUNS!!

3.4 Scored Statements

3.4.1 The Scores

The following sections list the statements for which the indexers were asked whether they agreed or disagreed. The average scores are presented such that a score of 10 is the strongest agreement with a positive concept. (For negative statements their 10 - n score was used for averaging.) Therefore a score of 5 is essentially no judgement on the topic.

Judgemental Performance: 4.7

- Overall Quality (+ -) 4.7

5.4 13. The Indexing Initiative pane helps me to produce a high-quality result (a fully indexed article with the correct terms.)

3.8 5. I think the overall quality of the suggested heading lists is unacceptable.

- Term Quality (-) 4.7

5.4 7. The Indexing Initiative pane encourages me to use extraneous terms in my indexing.

- Completeness (+ -) 4.2

5.7 16. MTI coverage of significant topics was good. ($\sigma = 2.37$)

2.6 8. Important subject areas were sometimes missing from the list of suggested headings. ($\sigma = 2.25$)

- Helpful (+) 5.4

5.4 1. The Indexing Initiative pane was a helpful tool in indexing.

Opinion on Usability:

- Confidence (+ -) 4.6

3.9 9. I have confidence in the accuracy of the suggested headings from MTI.

5.3 15. I am apprehensive about accepting subject headings recommended by MTI.

- Ease of use (-) 3.1

3.1 10. I find the Indexing Initiative pane difficult to use.

- List size (+ -), 2.2 ($\sigma = 2.48$)

1.4 11. I would rather see a longer list of suggested terms.

3.0 2. For most articles the list of suggested subject heading was too long.

- Explanation (-) 4.4

4.4 3. MTI should give justification for its suggestions.

Organizational Impact:

- Workload (+),

3.7 4. Using MTI will decrease the workload of indexers.

- Efficiency(+ -) 4.8

4.1 6. Indexing can be performed faster using terms suggested by MTI.

5.4 12. Regular use of the Indexing Initiative pane would slow down indexing.

- Novice users (+) 5.7 ($\sigma = 2.87$)

5.7 14. Using the Indexing Initiative pane will improve the skills of an inexperienced indexer.

The statement on which the indexers most agreed (smallest standard deviation) was number 9.($\sigma = 1.37$) With a maximum of 6 and a minimum of 2, they only mildly disagree with having confidence in the accuracy of the MTI recommendations(4.6).

3.4.2 The Comments

The 87 comments that were given for explanation of scores assign to the statements were diverse but helpful. They can be found in their entirety in Appendix C. The listing there is organized by topic as the scores were reported above.

One particularly interesting comment to statement #8 concerning whether important subject areas were sometimes missing was,

> "especially concept terms which don't actually appear in the text (ie. structure activity relat, gene expression reg, protein binding, host parasite relationship, the like)"

This reminds us of the fundamental limitations of an automated system like MTI.

On whether the II pane was a helpful tool(#1) this comment was included: "The limiting factor in indexing is not the typing or inputting of terms, but rather the time it takes to decide what are the correct terms. "

Comment on statement 4: Using MTI would decrease the workload of indexers, "...if the headings were more in sync with indexing policy."

Comment on Efficiency (#6, 12): "I think initially it would [slow down indexing,] but I think after indexers get used to it, it would probably make it faster. I found it disrupted my process of indexing, but I think I could adapt to it"

3.5 Article Performance Evaluation

There were two questions asked to assess overall performance of MTI for each article that opened the Article Questionnaire. The first was multiple choice and the second open-ended.

3.5.1 Did the list of suggested subject headings cover the purpose of the article?

Table 7 shows the indexers answers to the question above. Three articles have no response.

TABLE 7. MTI Article Performance

n	Percent	Coverage
59	37	Yes
83	53	Partially
16	10	No

3.5.2 If the list of suggested subject headings made you think of some new conceptual idea to use in your indexing, please explain here.

There were only 8 (5%) responses to this question but they came from five different indexers. Three of them have provided two responses each. Their responses are shown in Table 8.

TABLE 8. Responses to New Idea Request

Article PMID	New Idea
11842249	gamma interferon, recombinant (>spec than existing gamma interferon
11842255	oligosaccharides, asparagine added
11818113	this was a "brief communication", only 1 page and 4 references, therefore indexed non-depth
11818116	I am wondering why you didnt have the term trophoblastic neoplasm. This was a difficult article to index.
11875576	I added c-myc genes. I was not going to at first, but changed my mind after the first questionnaire
11846742	I used anemia, hemolytic, hereditary, which is a synonym for anemia, hemolytic, congenital, but is not further subdivided in mesh to the more specific, nonspherocytic.
11767989	i DID NOT USE dATA cOLLECTION AT FIRST BECAUSE i DIDN'T REALIZE IT WAS X-REF TO SURVEYS
11767998	YES, TO COVER THE CONCEPT OF 'SUBTHRESHOLD STIMULATION' COSE FROM YOUR LIST INSTEAD OF MH TENS

3.5.3 Changes in indexing.

From the evaluation of the duplicate Article Questionnaires for the same article, we found that four indexers (one of them twice) changed their indexing to include a term appearing on the questionnaire after completing the questionnaire the first time. Here are those terms and the articles for which they were used.

TABLE 9. Terms Selected on Review

PubMed Id	MeSH Heading
11767989	Data Collection
11821019	Adenosine Triphosphate
11875574	Tymus Gland
11875576	Genes, myc
11846739	Mutation

3.6 Reject Evaluation

The indexers were asked to evaluate the top five suggested terms that they did not select (rejects). So this evaluation excludes the possibly 20 lower ranked suggested terms. This section presents the summary of the scores for the strength of the connection between the 809 rejected terms and the article. It also presents more details about that relatedness or lack of it. Several summary views of the statements selected to describe the terms precede the summaries of the comments that accompanied those selections.

3.6.1 How Connected to Article

The primary evaluation of a rejected term was the button selected by the indexer to indicate how strong the connection was. For our analysis we have assigned numbers (1-5) to each button. Table 10 shows the indexers selections. The indexer volunteers failed to select a value for 3% of the reject terms.

TABLE 10. How strong is the connection between this term and the article?

Meaning	Score	Percent	N
Not related	1	0.34	267
	2	0.07	58
Remotely related	3	0.16	127
	4	0.11	85
Closely related	5	0.32	249

On average the not selected terms were rated “Remotely related,” but as can bee seen from the table the selections are broadly distributed.

This distribution of these scores for the selected descriptive statements reveals some informal correlation. This information is shown in Table 11.

TABLE 11. Related Scores by Statement

Statement score	1	2	3	4	5
Distraction	78	8	17		3
Wrong Category	9	6	10	5	5
Another Might	1		7	14	19
Not Important	1	9	40	27	16

3.6.2 Descriptive Statements

All of the descriptive statements are provided as reasons for not choosing the term being evaluated. These descriptive statements come in three groups: There are four that

describe or reflect the nature of the relationship to the article. There are two that refer to MeSH and two more that deal with specificity.

How Related. Table 12 summarizes the reject evaluation by showing how often and in in which combinations the statements about the terms were selected. The statements are noted by abbreviated phrases, the full statements can be seen in Section 2, "Sample Reject Form," on page 6. The X's indicate which statements were selected for the count at the top.

TABLE 12. Statement Assignments to Rejected Terms

Statement	16%	3.9%	3.8%	11%	.4%	.4%	.1%	.1%	.9%	
count	129	32	31	85	3	3	1	1	7	Total
Distraction	X				X	X				135
Wrong Category		X			X		X	X		37
Another Might			X			X	X		X	42
Not Important				X				X	X	93

Table 11 above shows the distribution of Related scores for each. One or more of these statements were only select for 36% of the terms evaluated.

MeSH. The statement "has meaning in MeSH different from it use in this article" is intended to reveal possible ambiguity problems in the MTI indexing. This statement was selected 17 times (2.1% of terms evaluated). The statement "Selection of this term not consistent with its annotation." was selected only 9 times(1.1%). Table 13 and Table 14 show the terms selected and the Related score of each.

TABLE 13. Suggested Terms Not Consistent with Annotation

Term	Related Score
Molecular Sequence Data	0
Aortic Diseases	0
Heart	1
Protein Isoforms	4
Fever	5
Base Sequence	5
Amino Acid Sequence	5
Central Nervous System	5
Aorta, Abdominal	5

TABLE 14. Article Concept Different from Scope Note of These Suggested Terms

Term	Related Score
Power (Psychology)	1
Vacuum	1
Paper	1
Regeneration	1
Shock	1
Clone Cells	1
Drinking	1
Linkage (Genetics)	1
Angina Pectoris, Variant	1
Reference Standards	2
Phototropism	3
Family	3
Residence Characteristics	3
Spasm	3
Heart Atrium	4
Cell Fusion	5
Blood Vessel Prosthesis Implantation	5

Specificity. Consistent with their own general observations, the indexers rated 210 (26%) of the top five rejected terms as "too general to use for this article." Only 33 terms (4.1%) had the statement "is to specific to use for the article" assigned.

This distribution of these scores for the selected statements reveals some informal correlation. This information is shown in Table 15. The Annotation statement was assigned to two terms that did not receive Related scores This was also true for two for "too general" and one for "too specific." Note that 61% of the closely related terms (5) were "too general.".

TABLE 15. Related Scores by Statement

Statement score	1	2	3	4	5
MeSH Different	9	1	4	1	2
Annotation	1			1	4
Too General	8	10	10	27	153
Too Specific	1	3	8	5	15

3.6.3 Explanations

The data presented in this section are the indexer explanations of their selections for the descriptive statements. We summarize the comments in the context of the descriptive terms that were selected.

TABLE 16. Explanation of Descriptive Statement Selection

Selected Statement	N	Comments	
Distraction	41	Not found	1
		Not discussed in results	6
		Not really the point	5
		Background or Introduction only	4
		Used more specific term	2
Wrong Category	4	MTI unable to separate an organ, from an organ disease or a finding from a disease term	2
Another Might	10	{Got added to MEDLINE later}	2
		MTI term a fragment of the needed term	3
Not Important	16	Appear in Introduction	2
		Non-depth indexing	2
		Indexing Subtlety	
MeSH Different	3	Linkage_(Genetics): linkage involves two genes. this article attempted to link a gene to the disease.	
		Family: in mesh family is a social group. The subheading /ge refers to familial and genetic aspect of the disease. Genetic predisposition to disease further defines it in terms of familial inheritance.	
		Cell_Fusion: scope isn't right	
Annotation	3	Protein_Isoforms: ANNOTAT SAyS THAT FOR ENZYMES USE ISOENZYMES	
		Base_Sequence: don't index unless base seq data shown in article	
		Linkage_(Genetics): linkage involves two genes. this article attempted to link a gene to the disease.	
Too General	31	Many listed the more specific term.	18
		Suggested term is fragment of specific term	3
Too Specific	4	Non-depth indexing required group term	2
No statement selected	55	(These are characterized in next table.)	

TABLE 17. Explanations without any descriptive statement selections

Stereotyped Comment			
Related Score	**1,2**	**3,4**	**5**
Not there (not found)	10	4	1
Not discussed	1	4	
Not really the point		1	
From Background or Introduction	2	1	
3rd Tier	2	1	
Fragment of needed term	3		
Used more specific term		2	
Heading Mapped to			3
Publication Type			1
Gene vs. Protein			2
Indexing subtleties	2	6	6
Total Explanations	**23**	**18**	**14**

The full set of 167 explanations are available in Appendix D.

It is worth noting here that fortunately one of the selected journals has many short communications that were indexed “non-depth.” As a result the indexer may only use a few terms in total and so they often gave this a reason for not selecting an MTI suggested term.

3.6.4 Other Reasons

The data presented in this section are the indexer supplied reasons for not using the suggested term. We summarize the comments in Table 18. The full 66 comments are in Appendix D.

TABLE 18. Other Reasons for Rejecting Terms

Stereotyped Comment	n
Not discussed in results:	
Related score 1	5
Related score 2	2
Related score 3	8
Not found	
Related score 1	5
Related score 2	3
Indexed more specific term	5
HM for chemtool SCR maps to it	3
Fragment or Part of name of protein	4
Specialty term	2
Publication type	2
Annotation: “not used for indexing” [Age Group]	1
Covered by a subheading	1

3.6.5 Term Used Instead

The indexers provided for 216 preferable terms or explanations for the rejected terms.(27%). Here are some observations from the inspection of these comments.

- Usually a more specific term was used; one exception was also noted. The relationship between the chosen term and the rejected term in the MeSH trees needs future investigation.
- PT was suggested by MTI [MTI can easily filter these out.]
- MTI does less well on article with titles only. [Suggested list could be shortened.]
- Indexers index the chemical not the HM. [MTI should not apply to the Restrict to MeSH algorithm to SCR terms.]
- History of Medicine is too general. [MTI should look for dates and refine this.]

The full set of Term Used Instead comments are in Appendix D.

3.6.6 The Rejected Terms

There were 680 distinct terms evaluated in this experiment. The Table 19 presents a list of terms rejected more than once.

TABLE 19. Multiply Rejected Terms

Term	N	Term	N
Academies_and_Institutes	2	Neurons	2
Adenosine_Triphosphate	2	Nuclear_Matrix	2
Adolescent_Nutrition	2	Occupations	2
Angina_Pectoris	2	Poverty	2
Antigens,_CD8	2	Pregnancy_Outcome	2
Association	2	Prions	2
Atrial_Fibrillation	2	Prion_Diseases	2
Atrial_Flutter	2	Protein-Serine-Threonine_Kinases	2
Birth_Weight	2	PrPC_Proteins	2
Breast	2	Punishment	2
Ca(2+)-Transporting_ATPase	2	Pyridines	2
Cells	2	Repressor_Proteins	2
Cesium	2	Sensitivity_and_Specificity	2
Coitus	2	Social_Class	2
Connexins	2	Substance-Related_Disorders	2
Coronary_Disease	2	Time	2
Coronavirus	2	Tomography,_X-Ray_Computed	2
Costs_and_Cost_Analysis	2	Trans-Activators	2
Counseling	2	Translocation_(Genetics)	2
Creutzfeldt-Jakob_Syndrome	2	Trial_of_Labor	2
Cysteine_Endopeptidases	2	Vaginal_Birth_after_Cesarean	2
Death	2	Viruses	2
Diabetes_Mellitus	2	Arrhythmia	3
Embryo_Transfer	2	Base_Sequence	3
Emergency_Service,_Hospital	2	Biological_Transport	3
Endopeptidases	2	Delivery	3
Evaluation_Studies	2	Epitopes	3
Fever	2	Family	3
Gestational_Age	2	Fetal_Development	3
Guidelines	2	Gene_Expression	3
Heart_Atrium	2	Mutation	3
Hemoglobin_A,_Glycosylated	2	Nuclear_Proteins	3

TABLE 19. Multiply Rejected Terms

Term	N	Term	N
Hippocampus	2	Parents	3
Infant,_Low_Birth_Weight	2	Physicians	3
Interferon_Type_II	2	Protein_Kinase_C	3
Ischemic_Preconditioning,_Myocardia	2	Tachycardia	3
Language	2	Viral_Proteins	3
Membranes	2	Carrier_Proteins	4
Menorrhagia	2	Transcription,_Genetic	4
Mice,_Transgenic	2	Amino_Acid_Sequence	5
Moths	2	Molecular_Sequence_Data	5
Naphthalenes	2	Proteins	5
Nerve_Growth_Factors	2	DNA-Binding_Proteins	6
		Transcription_Factors	6
		Genes	7

Follow up on the most frequently rejected terms should include the search for shared characteristics and the pulling together indexer responses from the questionnaires for those with a frequency greater than two.

4.0 Analysis

We have employed both objective and subjective measures to evaluate the MTI results.

4.1 Indexing Statistics

The indexing statistics are encouraging but the large standard deviation values suggest a fundamental weakness. 64% of the recall scores are likely between 0.37 and 0.73, but the lower end performance is not likely to be acceptable to the MEDLINE indexers. The current data may help identify the sources of this inconsistency.

Preliminary examination of the differences in MTI performance for the various journals suggests that the broad range of indexer opinion may reflect inconsistent performance of MTI. Although generally MTI provided more useful sets of suggestions for the clinical

journals and fewer useful sets for pre-clinical journals. On this metric, MTI performed best on **International Journal of Gynaecology and Obstetrics** and **Pediatrics.** MTI did least well on **Nature** and **Clinical Genetics**. We found that from 5-33% of the sets of suggested terms were above average in recall and precision. The vocabularies of the UMLS Metathesaurus are primarily clinically oriented and this may partially explain the better performance in those journals. A significant number of the articles in Nature are on genetics topics. The up coming work to extend the MetaMap data to include more genetic information may improve MTI's performance on those weakest journals. Additional examination of MTI's performance on the different journals may reveal other patterns pointing to ways to improve MTI's consistency.

We planned to check whether an indexer revised their own work after completing the evaluation. The indexing reported at completion of the article might not have been final since indexers were able to return to an article before they released a journal and could modify the indexing. As noted above, this was reported by indexers in two cases. Forty-nine (18%) of the articles were revised at least once. The nature of the modifications are left for future investigation.

4.2 Other Objective Measures

Excluding the time used for the surveys, we have collected performance times for the indexers on each article. Excluding times that are disproportionally long, this data establishes a baseline for comparison of future MTI support of indexing.

4.3 Negative Case Analysis

We briefly explore some negative cases, both those poor suggestions that were rejected by the indexers and the missed IM terms that were not suggested.

4.3.1 Poor Suggestions

Several kinds of information may be extracted from the study of the reject questionnaire responses.

The assignment of the "too general" statement to 26% of the top five rejected terms affirms the presence of a generality problem. (61% of closely relaed terms)The low, 4%, assignment of the "too specific" statement suggests that we do not have to worry about being too specific. These results indicate a need for improvement of the tree related tweaks to support semi-automatic indexing. Examination of these particular cases may reveal reasons why MTI is mapping to the more general term while the indexers seem to be able to find a more specific one. One hint from the comments is that MTI often identifies a fragment of the phrase that identifies the correct, more specific concept. Some are simple: MTI gives "Death" when "Cell death" is closer. Some are more complex: MTI suggests "Iodobenzenes" when MIBG (3-Iodobenzylguanidine) is needed.

Comments accompanying these selections revealed that the term "Genes" is usually expressed with the subheading "genetics." This suggests that MTI should place this term on a list of terms too general to index. Its tree code is G05.275; maybe all second level terms are too general to index. It was the most frequently rejected term, 7 times, in the experiment.

The assignment of "is from the wrong category (tree)" statement to only 4.5% of the terms evaluated suggests that ambiguity is not a significant problem in the MTI indexing.

The particular terms to which the "meaning in MeSH" and "not consistent with annotation" statements will be investigated to find indexing guideline that could be extracted automatically for use by MTI. The volunteer indexers have given us the first class of terms that can be avoided: Terms whose annotation says, "Do not use in indexing." The data give us another correction:Do not index with these terms because consistency with annotation is not possible with the abstract alone: Molecular Sequence Data, Base Sequence, Amino Acid Sequence

Patterns relating to the use of terms and their ranking score need to be investigated. Such patterns may suggest a ranking threshold for selecting terms to present from the MTI output.

A previously unrecognized problem became evident from the indexer supplied reasons for the rejection and other comments. The abstract usually includes a brief introduction/background "section." Terms from there make it into the MTI results, but are not suitable for indexing. An investigation of discourse analysis on these very short documents may lead to ways of identifying and suppressing these terms.

4.3.2 Missed Important Terms

The example given above to illustrate the objective measures,(see Table 5 on page 12) gives an example of how hard the mapping problem really is. MTI did not provide *Enterostomy* which was one of the indexers IM terms. The term *Ostomy* stands out in the table as an offered term that was not taken; the only one in the top seven MTI terms. The indexers review of this term gave it a 4 for relevance but no comments. The first word in the title: *Appendicostomy* is not in MeSH, but does appear in the UMLS Metathesaurus. Restrict to MeSH maps it to *Ostomy*, the parent concept of *Enterostomy. Ostomy* is a broader term for both *Enterostomy* and the concept to which *Appendicostomy* belongs. So the Restrict-To_MeSH algorithm could probably not do better. I think the indexer decided that along with *Colostomy* and *Ileostomy* (both in the MTI list, the former in the indexers indexing) that *Appendicostomy* belonged as a child of *Enterostomy* and pick it as the closest substitute.

4.4 Questionnaire Assessment

The final questionnaire includes five pairs of statements that are intended to measure the same aspect of the MTI performance. The topics cover are noted here:

- Judgemental Performance: Overall Quality (13 - 5),
  Completeness (16 - 8);
- Opinion on Usability: Confidence (9 - 15),
  List size (11- 2),
- Organizational Impact: Efficiency (6 - 12),

The statement number listed first is a positive statement and the second is a negative statement about that topic. This design is intended to allow a partial assessment of the reliability of the questionnaire using a split-half reliability measure.

The correlation of the positive to negative scores was 0.4093.

The Cronback's Alpha (coefficient of reliability) was 0.5809

A score of 1.0 would indicate perfect reliability, so these scores indicate some problems with the questionnaire as a reliable instrument.

For several of the paired statements, some indexers agreed strongly with both the positive and negative statements. This is illustrated in Table 20 on page 30 where individual responses to paired statements were not similar as expected but were on opposite sides dff 5. or the comparison the negative statement scores were complemented (10-n). If a

TABLE 20. Individual Responses with Divergent Scores for Paired Statements

Statement Pair	MRI	Score for Positive	10 - Score for Negative	
6 - 12	NLM019059943	6	4	*
9 - 15	ONL000922823	4	7	
9 - 15	NLM009414999	2	9	
11 - 2	NLM019059943	1	8	*
13 - 5	NLM019407895	2	7	
13 - 5	NLM021334634	1	6	+
16 - 8	NLM019122495	6	1	
16 - 8	NLM019111680	6	2	
16 - 8	NLM021334634	7	4	+
16 - 8	ONL000921878	7	2	

indexer agreed strongly with the positive statement (score of say 7), then the expectation was that they would disagree with the negative statement and give a score of 3. (7 = 10-3). In the table these responses are negatively correlated, the actual responses were similar to statements intended to be opposites.

The marks in the right most column indicate responses from the same indexer.

For the pair dealing with the length of the suggested list(#11- 2), one indexer being happy with length disagreed with both, "I would rather see a longer list of suggested terms" and "...the list of suggested subject headings was too long." The indexer explained that longer was "Not necessary" and that, "I like the length, it gives terms not needed but thats ok."

The 16 - 8 pair seem not to be true opposite expressions of completeness.:

Positive Statement: 16. Coverage of significant topics was good.

Negative Statement: 8. Important subject areas were sometimes missing.

Four of the ten indexers agreed with both, agreeing more strongly with the negative statement.

The 13 - 5 pair was more divergent for a few. This is unfortunate because these statements were to assess the indexers' view of the overall quality of the MTI indexing. One indexer disagreed with both because for her, "I dont think it is the quality that is a problem, but the system."

4.5 User Satisfaction

User satisfaction assessment is simplified in this situation because the concept of useful and the measures of utility are clearer that is often the case. If the term is used in the final indexing then it was useful. Overall satisfaction is a desirable measure, but harder to assess.

The first of the scored statements was the only one that attempted to measure helpfulness. The indexers were divided in their opinions:

- Average Score: 5.4 ($\sigma = 2.76$)
- Scores: 0 2 5 5 5 6 6 6 9 10

The reactions of the indexers in this area were mixed. Some would use this tool regularly. Others would find it easier to type in the obvious terms and faster too because they would not risk being distracted the unrelated terms.

Also, user satisfaction with the suggested subject heading is confounded with their presentation and other aspects of the human-computer interface. The indexer complaints and suggestions for improvement of the interface should be addressed.

4.6 Future MTI Indexing Presentation

There are three issues describing the key elements of the presentation of MTI results: Selection, Order, and Content.

Selection. In the 1997 experiment, a ranking score threshold of 10 or more was used to pick the terms to be presented. The current experiment applied no threshold. To address this question of whether to select the appropriate terms to display from a selected ranking score threshold or a term count limit, we will determine the lowest ranked score for terms selected from the MTI list of suggestions. Analysis of these values may suggest a possible threshold.

Indexer complaints in the earlier experiment, and now, have emphasized inappropriate terms. Increased precision may be favored. MTI will always miss some significant terms, including especially check tags, because they appear only in the body of the article and not in the title and abstract. This also suggests that any future tuning of MTI should be in favor of precision over recall. Some of the tweaks may also contribute to enhanced accuracy. But analysis alone will not yield the best values so additional experiments may be run to determine the best selection.

The current experiment provides a baseline for an *F* measure of 0.2. We will determine the list size that maximizes F-measure for $\alpha = 0.2$. (This balance of recall and precision should be experimentally evaluated by repeating this experiment with each indexer seeing suggestions with different selections schemes and being asked to compare the sets.)

Order. In this experiment the terms were presented in rank order. Indexers who act as revisers are use to seeing the indexing presented in the order that the term were entered. Indexers pretty much start with the title and go through the article adding significant terms as they find them. Therefore, the subject headings tell the story of the article best if they appear in the same order that the concepts were introduced in the article. So the suggested subject headings could be present to the indexers in this experiment in appearance order. An indexer has suggested that this list be in alphabetical order. The fourth options is to make the first three selectable by the indexer as an option in DCMS. These options could be discussed with the volunteers before any future experiments.

Content. In the experiment of 1997 indexers were often confused when they could not tell where the suggested term came from and they have repeated these concerns this time. So to make the suggestions more understandable, we will consider presenting, in future trials, the noun phrase to which the term was matched. If the term has no MetaMap evidence supporting its selection, then the note <Rel. Cit.> could be displayed instead.

The MTI results are returned with the preferred form of the MeSH term. This experiment makes clear that indexers would prefer the entry term closest to the text of the article since that is often the one they use. This fits well with the previous suggestion; the match from the term to the text is obvious if they are the same words.

Another piece of information that has been considered is the ranking score. However, the comments of the indexers suggest that they have no interest whatsoever in knowing how strongly the MTI felt the term was related to an article.

An indexer has also suggested that the terms in the list be hot links to enable easy access to the MeSH records for that term.

Similarly, our investigation of the tree position of terms as in Tweak 2, could be used to mark those terms that are not leaves. This could be as in the MeSH browser with a ‘+’ at the end of the term. Indexers could then use the MeSH link to consider the more specific terms.