Appendix A. Statistical Aspects of Formulating the TIL Criteria
Version 1, Tuesday, February 6, 2007

I. Overall Objectives

This document evaluates the main statistical issues in formulating the TIL criteria, i.e. evaluation of different sample sizes and requirements for the percentage of subjects needing to achieve an acceptable penetration level. In this document, the acceptable penetration level is specified as 0.05; the feasibility a 0.05 cut-off, and other descriptive analyses of penetration levels from the available TIL data, are detailed in Appendix B. Sampling strategies for subject selection are detailed in Appendix C.

For purposes of this document, the percentage of subjects required to meet the 0.05 penetration level is subsequently referred to as the 'cut-off percentage', and denoted by p^*. The observed percentage of subjects meeting that cut-off for a given respirator is subsequently referred to as the 'observed percentage' and denoted by $\hat{p}$. The analysis also depends on defining the distribution of $\hat{p}$ under different assumptions about the percentage of subjects that truly achieve the 0.05 penetration level for the given respirator; those percentages are subsequently referred to as the 'assumed percentage' and denoted by p. For instance, we might wish to evaluate the probability of observing $\hat{p} > p^*$ (i.e. the respirator passes the test) if the respirator truly worked for only 40% of the population of workers (i.e. was 40% effective); alternatively, we could evaluate the probability of observing $\hat{p} > p^*$ if the respirator truly worked for 90% of the population of workers (i.e. was 90% effective). The first example should lead to a relatively low probability, whereas the second example should lead to a relatively high probability.

II. Specific Objectives:

1) Determine the minimum number of subjects (denoted by n) needed for achieving sufficient statistical power under different assumptions.
2) Determine the optimal cut-off percentage for the fraction of subjects achieving an acceptable penetration level (i.e. below 0.05).

These two objectives have to be addressed simultaneously since different sample sizes may change the resulting statistical properties and may thus lead to different choices for cut-off percentage across different sample sizes.

III. Statistical Methods:

To evaluate different sample sizes and cut-off percentages, the analysis utilizes the following strategy:

1) View the selection of a minimum sample size and an optimal cut-off percentage in a hypothesis testing framework, i.e. choose the cut-off percentage that maximizes the statistical power (true positive results) and minimizes the type I error rate (false positive results) for a given sample size
2) To accomplish this overall goal, first specify a range of assumed percentages (e.g. 40%, 50%, …, 90%), and define degrees of acceptability; for instance, a respirator that achieves a penetration ≤ 0.05 for 90% of the subjects (i.e. was 90% effective) would be defined as highly acceptable, and should therefore pass the TIL criteria for a high percentage (if not all) of the tests, whereas an assumed

percentage of 40% would be defined as unacceptable, and should therefore fail the TIL criteria for a high percentage (if not all) of the tests; the test may allow for a higher degree of chance results with more intermediate values for the assumed percentage (e.g. 70%)

3) For each assumed percentage p (as specified in #2), and a given sample size (starting with $n = 25$), use the binomial distribution (simulated in S-Plus software with 10,000 random numbers generated for each choice of n and p) to simulate observed percentages for a hypothetical population of respirators (with the given assumed percentage); the simulation results will give the known sampling distribution of $\hat{p}$ under the given choice for n and p
4) Based on the results of #3 choose a cut-off percentage which gives a high probability of passing an "acceptable" respirator (i.e. observing $\hat{p} > p^*$ for a respirator with a high value of p) and low probability of passing an "unacceptable" respirator (i.e. observing $\hat{p} > p^*$ for a respirator with a low value of p)
5) Select a sample size that balances feasibility with minimizing the degree of chance results; larger sample sizes will lead to a more narrow range of values for p where the probabilities of passing the test are relatively near 0.5 (i.e. involve a high degree of chance), whereas smaller sample sizes will require p to be closer to 0% or 100% to achieve definitive results

For any hypothesis test, we typically choose a criteria (to fail or pass the statistical test) based on two statistical properties.

1) Minimize the chances of falsely rejecting (or failing the model) if it is really above the minimum required percentage; typically we limit the probability of this 'type I error' to 5%.
2) Maximize the chances of failing a respirator if it is in fact below the required percentage; typically we want this probability, called the 'power of the test' to be at least 80 or 90%. Another way of stating this is to say that we want to minimize the type II error, i.e. limit the chances of not rejecting a false hypothesis to 10 or 20%.

Although rejecting a true hypothesis, i.e. the type I error, is typically consider to be the "worst" error, formulation of the TIL criteria depends more on achieving high statistical power, since failing to reject an ineffective respirator equates to life-time certification of a suboptimal model, and subsequently increased risk to the individual worker. Therefore, we should consider choosing a cut-off percentage that limits the type II error to < 5%, perhaps at the cost of a larger type I error, or equivalently, maximize the power to fail an unacceptable respirator to at least 95%.

IV. Results:

In terms of specifying the range of assumed percentages (e.g. 40-90%) and defining degrees of acceptability, 80-90% effective was deemed to be an acceptable range, where we would like to minimize the percentage of times such models will fail the test. In contrast, 60% or less was deemed to be unacceptable, where we would like to maximize the percentage of times such models will fail the test. As previously discussed, failing models in the unacceptable range was deemed to be more important, since an error in that

direction would lead to life-time certification of a model with an unacceptable level of effectiveness. Assumed percentages near 70% were deemed to represent an intermediate range, where a higher degree of chance results could be tolerated.

The following tables present the simulation results for the percentage of times that a given respirator (with some assumed percentage of effectiveness between 40% and 90%) will meet a given cut-off percentage using sample sizes of 25, 35 and 50.

Table 1. Percentage of tests that a given respirator will fail using a sample size of 25

Minimum # of Subjects Required to Pass[1] (%)	Assumed Percentage of Subjects Achieving the Required Penetration					
	90%	80%	70%	60%	50%	40%
13 (52%)	*<0.1%*	*<0.1%*	*1.7%*	*15.5%*	*50.0%*	*84.4%*
14 (56%)	*<0.1%*	*0.2%*	*4.4%*	*26.8%*	*65.4%*	*92.3%*
15 (60%)	*<0.1%*	*0.6%*	*10.0%*	*41.4%*	*78.8%*	*96.6%*
16 (64%)	*<0.1%*	*1.7%*	*19.0%*	*57.4%*	*88.7%*	*98.7%*
17 (68%)	*<0.1%*	*4.7%*	*32.1%*	*72.7%*	*94.6%*	*99.6%*
18 (72%)	*0.3%*	*10.9%*	*48.9%*	*84.6%*	*97.8%*	*99.9%*
19 (76%)	*0.9%*	*22.2%*	*66.0%*	*92.7%*	*99.2%*	*>99.9%*
20 (80%)	*3.4%*	*38.5%*	*80.6%*	*97.2%*	*99.8%*	*>99.9%*
21 (84%)	*9.7%*	*58.1%*	*91.0%*	*99.0%*	*99.9%*	*>99.9%*
22 (88%)	*23.5%*	*76.7%*	*96.6%*	*99.8%*	*>99.9%*	*>99.9%*
23 (92%)	*46.4%*	*90.2%*	*99.1%*	*>99.9%*	*>99.9%*	*>99.9%*
24 (96%)	*72.9%*	*97.2%*	*99.8%*	*>99.9%*	*>99.9%*	*>99.9%*
25 (100%)	*92.8%*	*99.6%*	*>99.9%*	*>99.9%*	*>99.9%*	*>99.9%*

[1]For an individual subject, passing the test is defined as achieving a penetration ≤ 0.05.

The above results seem to indicate that, for a sample size of 25, we should choose a cut-off of at least 18/25 (72%) to achieve optimal results. Using that cut-off, respirators that are effective for 80% of subjects only fail 10.9% of the time, whereas respirators that are effective for only 60% of subjects will fail 84.6% of the time. Using a lower criteria (e.g. 17/25 or less) has insufficient power to reject a poorly performing respirator. Increasing the cut-off may be needed, since the 84.6% failure rate (for model effective on only 60% of subjects) may be too low. Increasing the cut-off to 19/25 (76%) increases our power to 92.7% (for rejecting models effective on only 60% of subjects). However, it also sharply increases the type I error rate; for instance, models effective on 80% of subjects will fail the test 22% of the time. This suggests a potential need for increasing the sample size.

Table 2. Percentage of tests that a given respirator will fail using a sample size of 35

Minimum # of Subjects Required to Pass[1] (%)	Assumed Percentage of Subjects Achieving the Required Penetration					
	90%	80%	70%	60%	50%	40%
18 (51%)	*<0.1%*	*<0.1%*	*0.6%*	*11.5%*	*50.1%*	*88.6%*
19 (54%)	*<0.1%*	*<0.1%*	*1.6%*	*19.3%*	*63.2%*	*93.8%*
20 (57%)	*<0.1%*	*0.1%*	*3.5%*	*30.1%*	*75.0%*	*96.9%*
21 (60%)	*<0.1%*	*0.2%*	*7.2%*	*42.6%*	*84.3%*	*98.7%*
22 (63%)	*<0.1%*	*0.5%*	*13.4%*	*56.3%*	*91.3%*	*99.5%*
23 (66%)	*<0.1%*	*1.4%*	*22.9%*	*69.4%*	*95.5%*	*99.8%*
24 (69%)	*<0.1%*	*3.5%*	*34.6%*	*80.4%*	*98.0%*	*>99.9%*
25 (71%)	*<0.1%*	*7.4%*	*49.2%*	*89.0%*	*99.2%*	*>99.9%*
26 (74%)	*0.2%*	*14.5%*	*63.6%*	*94.3%*	*99.7%*	*>99.9%*
27 (77%)	*0.6%*	*25.4%*	*76.4%*	*97.4%*	*99.9%*	*>99.9%*
28 (80%)	*2.0%*	*40.0%*	*86.7%*	*98.9%*	*>99.9%*	*>99.9%*
29 (83%)	*5.6%*	*56.6%*	*93.4%*	*99.6%*	*>99.9%*	*>99.9%*
30 (86%)	*13.3%*	*73.0%*	*97.2%*	*99.9%*	*>99.9%*	*>99.9%*
31 (89%)	*26.9%*	*85.6%*	*99.1%*	*>99.9%*	*>99.9%*	*>99.9%*
32 (91%)	*46.9%*	*93.9%*	*99.7%*	*>99.9%*	*>99.9%*	*>99.9%*
33 (94%)	*69.5%*	*98.1%*	*>99.9%*	*>99.9%*	*>99.9%*	*>99.9%*
34 (97%)	*87.7%*	*99.6%*	*>99.9%*	*>99.9%*	*>99.9%*	*>99.9%*
35 (100%)	*97.4%*	*>99.9%*	*>99.9%*	*>99.9%*	*>99.9%*	*>99.9%*

A sample size of 35 improves noticeably on the above results. For instance, from Table 1, a cut-off of 18/25 (72%) meant that respirators which are effective for 80% of subjects only fail 10.9% of the time, whereas respirators that are effective for only 60% of subjects will fail 84.6% of the time. With a sample size of 35, a similar cut-off of 25/35 (71%) yields a lower type I error (7.4% versus 10.9%) in terms of failing an 80% effective respirator and a higher power (89% versus 84.6%) to fail a 60% effective respirator. Increasing the cut-off to 26/35 (74%) increases the power to fail a 60% effective respirator to 94.3% at the cost of failing an 80% effective respirator 14.5% of the time.

Table 3. Percentage of tests that a given respirator will fail using a sample size of 50

Minimum # of Subjects Required to Pass[1] (%)	Assumed Percentage of Subjects Achieving the Required Penetration					
	90%	80%	70%	60%	50%	40%
26 (52%)	*<0.1%*	*<0.1%*	*0.3%*	*9.8%*	*55.7%*	*94.3%*
28 (56%)	*<0.1%*	*<0.1%*	*1.2%*	*23.5%*	*75.8%*	*98.4%*
30 (60%)	*<0.1%*	*<0.1%*	*4.7*	*43.8%*	*89.9%*	*99.7%*
32 (64%)	*<0.1%*	*0.2%*	*14.1%*	*66.5%*	*96.7%*	*99.9%*
34 (68%)	*<0.1%*	*1.4%*	*31.6%*	*84.4%*	*99.2%*	*>99.9%*
36 (72%)	*<0.1%*	*6.0%*	*55.5%*	*94.6%*	*99.9%*	*>99.9%*
37 (74%)	*<0.1%*	*11.0%*	*67.1%*	*97.2%*	*>99.9%*	*>99.9%*
38 (76%)	*0.1%*	*18.5%*	*77.5%*	*98.7%*	*>99.9%*	*>99.9%*
40 (80%)	*1.0%*	*41.7%*	*92.2%*	*99.8%*	*>99.9%*	*>99.9%*
42 (84%)	*5.7%*	*69.3%*	*98.3%*	*>99.9%*	*>99.9%*	*>99.9%*
44 (88%)	*22.8%*	*89.7%*	*99.7%*	*>99.9%*	*>99.9%*	*>99.9%*
46 (92%)	*56.8%*	*98.1%*	*>99.9%*	*>99.9%*	*>99.9%*	*>99.9%*
48 (96%)	*88.8%*	*99.9%*	*>99.9%*	*>99.9%*	*>99.9%*	*>99.9%*
50 (100%)	*99.5%*	*>99.9%*	*>99.9%*	*>99.9%*	*>99.9%*	*>99.9%*

A sample size makes further improvements on reducing the type I error (say for an 80% effective respirator) and increasing the statistical power (to fail a 60% effective model). For instance, using a cut-off of 36/50 (72%), an 80% effective respirator will fail the test 6% of the time, while a 60% effective respirator will fail the test 94.6% of the time. Increasing the cut-off to 37/50 (74%) increases the percentage of time a 60% model will fail to 97.2%, at the cost of failing an 80% effective model 11% of the time.

V. Conclusions

Based on the previously described framework for selecting the optimal sample size and cut-off percentage, results indicate that a sample size of 25 may be insufficient for discriminating between an acceptable and unacceptable model. To achieve over 90% power (92.7%) for failing a model which is only 60% effective, the cut-off percentage needs to be at least 19/25; however, that cut-off percentage leads to failing a model which is 80% effective over 22% of the time. Increasing the sample size to 35, and using a cut-off percentage of 26/35 would slightly increase the power (to fail a respirator which is only 60% effective) to 94.3%, while noticeably reducing the percentage of failures for a respirator which is 80% effective to under 15%. Using a cut-off of 36/50 yields very similar power (94.6%), but only fails a respirator which is 80% effective 6% of the time. Considering these statistical properties, in conjunction with feasibility constraints, the cut-off of 26/35 was deemed to be the optimal rule for the TIL criteria.