Determination of Sample Size and Passing Criteria for Respirator Fit Test Panels

D Landsittel, Z Zhuang, W Newcomb, and R BerryAnn

Keywords: fit test panels, fit testing, sample size

INTRODUCTION

Face sizes and shapes have long been a significant area of research for those responsible for respirator design, testing and certification. Prior to 1972, the U.S. Bureau of Mines (USBM) was responsible for testing and approval of respirators used in the United States. The USBM testing program included the testing of respirators on three human subjects with "varying facial shapes and sizes," one full, one average and one lean, a provision that was criticized by many as being too vague. [1]

In 1972, the National Institute for Occupational Safety and Health (NIOSH) took over the USBM's responsibilities for respirator performance testing and certification, with those functions now carried out by the NIOSH National Personal Protective Technology Laboratory (NPPTL). The current NIOSH respirator certification regulations for three different mask types (42 CFR 84.124(a), 84.205(a), and 84.1156(b)(1)) indicate that the masks are to be tested on a panel of human subjects with "varying facial shapes and sizes" for a qualitative fit test using isoamyl acetate as the challenge vapor. [2]

Subsequent recommendations for fit test panels led to proposal of a 25-subject panel, developed by Los Alamos National Laboratory (LANL).[3,4] The full-facepiece panel was based on the bivariate distribution of face length and face width. The half-facepiece panel was based on the bivariate distribution of the face length and lip length. Each panel has 10 cells. The recommendations were that 25 subjects be recruited according to the face dimension criteria in the LANL panels for half- or full-facepiece respirators. It is a compromise between the need for a sufficient number of tests to achieve optimal statistical properties and the requirement to test all devices submitted for approval in a reasonable length of time. The pass/fail criteria for the panel suggested that a single respirator model designed to fit 95% of the user population should have only a single failure for the 25 subjects tested.[3] NIOSH regulations for testing various other respirator types use even smaller numbers of subjects.

The above LANL specifications, however, when proposed by NIOSH for certification regulations, drew criticism during the public comment period. [5] In response, NIOSH committed to a program of research to specifically address fit test methods. Some of that research has been completed and published.[6-8] In more recent years, NIOSH issued performance requirements for self-contained breathing apparatus (SCBA) respirators intended for emergency responses (NIOSH, 2001).[9] The LANL panel, which is defined by the face length and face width (for full-facepiece devices), was incorporated into a NIOSH certification facepiece fit test.[9]

Concern was raised about the applicability of military data and corresponding test panels to civilian workers. Military personnel have to meet strict entry and fitness criteria and also tend to be younger than the general civilian workforce. Military population may not represent the great diversity in face size in civilian population. As a result, personal protective equipment designed and sized for a military population may not provide the same level of fit to civilian workers. In addition, the demographics of the U.S. population have changed substantially over the last 30 years.

In 2001, the NIOSH National Personal Protective Technology Laboratory recognized the difficulties inherent in using the old military data, and initiated a study to develop an anthropometric database of the heads and faces of civilian respirator users.[8] The new database was also used to evaluate the ability of the LANL respirator fit test panels to represent the current U.S. civilian workers in a companion study.[9] Comparisons were made on age and race distributions as well as key face dimensions (face length, face width and lip length) between the USAF and NIOSH surveys. Age and race distributions of the USAF data were different from those of the NIOSH data. The bivariate distribution of face length and face width for full-facepiece applications and the face length and lip length for half-facepiece applications were different between the two surveys. Furthermore, the LANL full-facepiece panel excluded 15.3 percent of NIOSH survey subjects. Subjects in the NIOSH survey had larger key face dimensions (face length and face width) than the subjects in the 1967-68 USAF survey. Thus, it was concluded that the LANL respirator fit test panels did not adequately represent the current U.S. civilian workforce. The NIOSH survey on face anthropometry of respirator users is more representative of the age and racial/ethnic distributions of the current civilian population.

Based on the NIOSH anthropometric survey, Zhuang et al.[11] developed criteria for two new fit test panels for both half-mask and full-facepiece respirators. The NIOSH bivariate Respirator Fit Test Panel (NRFTP) was defined by face length and face width (previously used by LANL for full face masks). For a second panel (the PCA, or Principal Component Analysis Panel), facial dimensions were categorized based on two independent linear combinations (principal components) of 10 facial dimensions. The panel size in either was initially planned to be 25 subjects for testing one-size fits all models (based on the previous definition of the LANL panel).

Although these developments in respirator fit test panel design have devoted considerable attention to issues such as incorporation of anthropometric features, relatively little research has focused on the critical issues of sample determination and specification of appropriate passing criteria; these two issues are inherently connected since statistical properties of using a given critical fraction needed to pass depends on the total sample size. The Food and Drug Administration (2007) developed recommendations for assessment of quantitative fit of filtering facepiece respirators via a random effects analysis of variance (ANOVA) model, where fit factors are log-normal with two components of variability: the within- and between-subject variability (Nicas and Neuhaus, 2004). Bootstrap methods were then recommended for interval estimation and subsequent testing of the fit factor parameters. Zhang and Kolz (2008) subsequently presented a closed-form normal approximation method for the same underlying model and objectives.

The random effects ANOVA can then be applied for inference about the parameter of interest (i.e. percentage of subjects meeting the specified fit factor requirement over a certain percentage of repeated fits/donning) based on a specific data set, or for sample size estimation if we can specify a number of variance estimates; Zhang and Kotz (2008) specify equations for each approach using the normal approximation method. These methods may be extremely useful for assessment of fit for a given respirator model, where corresponding preliminary data exists. However, a more common need in this setting is specifying both passing criteria and minimum sample size requirements for a test panel to be used across many different respirator models. Accomplishing this task is not tractable within the random effects ANOVA framework, since doing so would require a number of specific estimates about the variance components, which will vary across different populations being targeted for different respirator models, and across different respirators fit to the same target population.

For instance, a given manufacturer may wish to test a single-size model designed to fit the entire population, whereas another manufacturer may wish to test a model for larger facial dimensions. As another illustration, consider two respirator models fit to the same target population, where one model uniformly fits the population with minimally adequate fit, whereas the second model fits most subjects extremely well but performs especially poorly on a minority of subjects. This type of scenario may result in practice due to different shapes or widths associated with the given respirator design. Each problem would lead to potential very different estimates of the different variance components; in some cases (where perhaps the model is designed to fit less common facial dimensions), it is unlikely that any reasonable estimates would exist.

Some specific problems of public health importance are also difficult to formulate in terms of distributional properties for underlying fit factors and assumptions in the ANOVA model. For instance, a significant public health question might be to assess whether, for a given test, a respirator model which truly fits only 50 percent of the target population will fail the test a high percentage of times. Conversely, from the manufacturer's perspective, will a respirator model which truly fits over 90 percent of the target population pass the test a high percentage of times? Such questions cannot be addressed directly using the random effects model without extensive assumptions about the underlying data. Given the fact that such assumptions may be difficult to specify, and may vary widely across different target populations, a more flexible methodology is needed for specifying both the test panel size and minimum passing criteria.

This study therefore aims to outline an alternative approach, based on simple binomial probabilities, which is feasible for specifying passing criteria across a range of sample sizes with minimal assumptions. The corresponding significance level and power of the test can then be assessed across different sample sizes to choose reasonable passing criteria. Although this approach will likely be less powerful when applicable variance components can be specified, such cases are uncommon in practice, and intractable for specifying certification criteria across a range of respirator models, thus motivating the need for more a flexible approach. Application of this approach is illustrated for the proposed NIOSH Total Inward Leakage (TIL) test, which is currently being proposed and reviewed for certification of half-mask respirators.

METHODS AND MATERIALS

Calculation of Binomial Probabilities

Denote $p_{c,n}$ as the true (or assumed) proportion of n individual subjects who achieve a passing result, i.e. achieve a penetration below some specified value of c, such as 0.01 (i.e. a fit factor above 100), and $\hat{p}_{c,n}$ as the sample fraction of test subjects who achieve a sufficiently high fit. A given respirator then passes the overall test if $\hat{p}_{c,n} > p_c^*$, where $p_{c,n}^*$ is specified to reject a sufficiently high percentage of respirators with unacceptably low p_c (i.e. achieve a sufficiently high power of 1 - β_{p_c}). The corresponding type I error rates (denoted by α_{p_c}) are then evaluated across a range of sample sizes. In contrast to the usual hypothesis testing framework, where we set the significance level at 5% (typically for a single parameter value under the null), and then assess the corresponding power to reject the null (which is then possibly fixed at 80 or 90% to calculate a minimum sample size), the problem of respirator testing places a higher priority on rejecting an ineffective model as opposed to not rejecting an effective model. Further, rather than specifying a single value of $p_{c,n}^*$, for fixed values of α, β, and specific null and alternative assumptions about p_c, we examine a range of scenarios (across different n) with different assumptions about p_c (as will be specified in the following section).

The probability of a given model, with true effectiveness p_c, failing the overall test, over n subjects, is given by Equation 1.

Equation 1. Probability of a Test Failure

$$P\left(\hat{p}_{c,n} < p_{c,n}*\right) = P(X < np_{c,n}* \mid p_c, n) = \sum_{k=1}^{np_{c,n}*-1} \binom{n}{k} p_c^{\,k} (1-p_c)^{n-k}$$

Probabilities using Equation 1 were then calculated for each possible value of $\hat{p}_{c,n}$ across different sample sizes and assumed values for p_c. In practice, reasonable values of n and $p_{c,n}^*$ can then be empirically determined to achieve sufficiently optimal power and significance levels across a range of applicable proportions for the different null and alternative hypotheses as described above and illustrated with the following example. For all subsequent analyses, c is specified to be 0.01; the value of n is varied but dropped from the notation for convenience.

In applying binomial probabilities to specify a test panel size and cut-off for the passing criteria, the use of multiple alternative and null hypotheses yields a more flexible approach but prevents us from providing specific formula for the minimum sample size and passing criteria. Instead, the approach can be summarized via the following algorithm:

1) Specify a range for n
2) Specify k type I error rates, i.e. $\alpha_{p_1} < \alpha_{p_2} < \ldots < \alpha_{p_k}$, for levels of fit, $p_1 > p_2 > \ldots > p_k$, deemed to be ineffective

3) Specify *l* type II error rates, i.e. $\beta_{p_{k+1}} < \beta_{p_{k+2}} < \ldots < \beta_{p_{k+k}}$, for levels of fit, $p_{k+1} < p_{k+2} < \ldots < p_{k+l}$, deemed to be effective
4) For each *n* and each value of p^* above 0.5, calculate the probability of a test failure using Equation 1 for each value of *p* under consideration
5) For each *n*, find the maximum p^* which satisfies the specified type I error rates in Step 2
6) Among all values of p^* yielding sufficiently low type I error rates, identify which criteria (i.e. p^*) also yield sufficient power as specified in Step 3

Application of Probability Calculations to the Total Inward Leakage Test

In order to apply the results of the above-described probability calculations to the proposed TIL certification test, we considered only a limited range of sample sizes which were deemed to be feasible (when considering the substantial time and financial constraints associated with TIL testing). Further, only a limited combination was considered since the selected sample size would have to be distributed over the 10 cells of the NRFTP in a representative fashion (i.e. in manner that closely matched the population distributions from the NIOSH/Anthrotech survey). Specifically, we considered sample sizes of 25 to 45 in increments of 5. Sample sizes below 25 were deemed inadequate to appropriately cover the NRFTP in a fashion that was sufficiently representative of the target population, and sample sizes of 50 or more were considered impractical in terms of eventually implementing the certification test procedure for the hundreds of applicable respirator models.

Respirators achieving sufficient fit for at least 80% of the population were considered "effective" models, and thus should pass the subsequently specified criteria a high percentage of the time (specified to be at least 80% of the time). Respirators achieving sufficient fit for at least 90% of the population were considered highly effective models, and should pass the subsequently specified criteria a very high percentage of the time (specified to be at least 90% of the time). In contrast, respirators achieving sufficient fit for no more than 60% of the population were considered "ineffective" models, and should pass the subsequently specified criteria a low percentage of the time (specified to be no more than 5% of the time). Respirators achieving sufficient fit for no more than 50% of the population were considered highly ineffective models, and should pass the subsequently specified criteria a very low percentage of the time (specified to be no more than 1% of the time). Respirators achieving sufficient fit for between 60 and 80% of the population cannot be clearly differentiated as either effective or ineffective, and we were therefore willing to accept some uncertainty as to their probability of passing the subsequently defined test criteria. In terms of the type I and type II error rates, the above assumptions correspond to $\alpha_{0.8} = 0.20$, $\alpha_{0.9} = 0.10$, $\beta_{0.6} = 0.05$ (or $1 - \beta_{0.6} = 0.95$), and $\beta_{0.5} = 0.01$ (or $1 - \beta_{0.5} = 0.99$).

RESULTS

The following tables presents the probability calculations for the percentage of times that a given unit, or respirator, which is effective for $100 \times p$% of the applicable subjects, will fail to achieve passing results for a given fraction, p^*, of the $n = 25$ to 45 test subjects.

Table 1. Percentage of times that a given unit will fail using a sample size of 25

Minimum # of Subjects Required to Pass[1] (%)	Assumed Percentage of the Population Achieving Passing Results					
	90%	80%	70%	60%	50%	40%
13 (52%)	*<0.1%*	*<0.1%*	*1.7%*	*15.4%*	*50.0%*	*84.6%*
14 (56%)	*<0.1%*	*0.2%*	*4.4%*	*26.8%*	*65.5%*	*92.2%*
15 (60%)	*<0.1%*	*0.6%*	*9.8%*	*41.4%*	*78.8%*	*96.6%*
16 (64%)	*<0.1%*	*1.7%*	*18.9%*	*57.5%*	*88.5%*	*98.7%*
17 (68%)	*<0.1%*	*4.7%*	*32.3%*	*72.6%*	*94.6%*	*99.6%*
18 (72%)	*0.2%*	*10.9%*	*48.8%*	*84.6%*	*97.8%*	*99.9%*
19 (76%)	*0.9%*	*22.0%*	*65.9%*	*92.6%*	*99.3%*	*>99.9%*
20 (80%)	*3.3%*	*38.3%*	*80.7%*	*97.1%*	*99.8%*	*>99.9%*
21 (84%)	*9.8%*	*57.9%*	*91.0%*	*99.1%*	*>99.9%*	*>99.9%*
22 (88%)	*23.6%*	*76.6%*	*96.7%*	*99.8%*	*>99.9%*	*>99.9%*
23 (92%)	*46.3%*	*90.2%*	*99.1%*	*>99.9%*	*>99.9%*	*>99.9%*
24 (96%)	*72.9%*	*97.3%*	*99.8%*	*>99.9%*	*>99.9%*	*>99.9%*
25 (100%)	*92.8%*	*99.6%*	*>99.9%*	*>99.9%*	*>99.9%*	*>99.9%*

The above results indicate that, for a sample size of 25, we should choose a cut-off of no more than 18/25 ($p^* = 0.72$) to achieve sufficiently low type I error rates ($\alpha_{0.8} = 0.109$ and $\alpha_{0.9} = 0.002$, which are both well below the specified values for $\alpha_{0.8}$ and $\alpha_{0.9}$). However, for the cut-off of $p^* = 0.72$, the power is too low for more effective respirators; specifically, $1 - \beta_{0.6} = 0.846$ (which is less than the specified level of 0.95) and $1 - \beta_{0.5} = 0.978$ (which is less than the specified level of 0.99). Increasing the cut-off to 19/25 ($p^* = 0.76$) leads to only slightly exceeding the specified type I error rate for 80% effectiveness ($\alpha_{0.8} = 0.22$) and still fails to achieve sufficient power for rejecting a respirator which is 60% effective ($1 - \beta_{0.6} = 0.926$, which is less than the specified level of 0.95). Further increasing p^* leads to unacceptably high error rates for $\alpha_{0.8}$ and $\alpha_{0.9}$. A sample size of 25 is therefore insufficient to meet the specified properties, although a cut-off of 19/25 ($p^* = 0.76$) leads to over 90% power for rejecting ineffective respirators and low error rates for highly effective respirators.

Table 2. Percentage of times that a given unit will fail using a sample size of 30

Minimum # of Subjects Required to Pass[1] (%)	Assumed Percentage of the Population Achieving Passing Results					
	90%	80%	70%	60%	50%	40%
16 (53%)	<0.1%	<0.1%	1.7%	17.5%	57.2%	90.3%
17 (57%)	<0.1%	<0.1%	4.0%	28.5%	70.8%	95.2%
18 (60%)	<0.1%	0.3%	8.4%	42.2%	81.9%	97.9%
19 (63%)	<0.1%	0.9%	15.9%	56.9%	90.0%	99.2%
20 (67%)	<0.1%	2.6%	27.0%	70.9%	95.1%	99.7%
21 (70%)	<0.1%	6.1%	41.1%	82.4%	97.9%	99.9%
22 (73%)	0.2%	12.9%	56.8%	90.6%	99.2%	>99.9%
23 (77%)	0.8%	23.9%	71.9%	95.6%	99.7%	>99.9%
24 (80%)	2.6%	39.3%	84.0%	98.3%	99.9%	>99.9%
25 (83%)	7.3%	57.2%	92.3%	99.4%	>99.9%	>99.9%
26 (87%)	17.5%	74.5%	97.0%	99.8%	>99.9%	>99.9%
27 (90%)	35.3%	87.7%	99.1%	>99.9%	>99.9%	>99.9%
28 (93%)	58.9%	95.6%	99.8%	>99.9%	>99.9%	>99.9%
29 (97%)	81.6%	98.9%	>99.9%	>99.9%	>99.9%	>99.9%
30 (100%)	95.8%	99.9%	>99.9%	>99.9%	>99.9%	>99.9%

The above results indicate that, for a sample size of 30, we should choose a cut-off of no more than 22/30 ($p^* = 0.73$) to achieve sufficiently low type I error rates ($\alpha_{0.8} = 0.129$ and $\alpha_{0.9} = 0.002$, which are both well below the specified values for $\alpha_{0.8}$ and $\alpha_{0.9}$). However, for $p^* = 0.73$, the power is again too low for more effective respirators; specifically, $1 - \beta_{0.6} = 0.906$ (which is less than the specified level of 0.95, although $1 - \beta_{0.5} = 0.992$, which is above than the specified level of 0.99). Increasing the cut-off to 23/30 ($p^* = 0.77$) leads to exceeding the specified type I error rate for 80% effectiveness ($\alpha_{0.8} = 0.239$) but does give sufficient power for rejecting ineffective respirators, with $1 - \beta_{0.6} = 0.956$ and $1 - \beta_{0.5} = 0.997$. A sample size of 30 is therefore still insufficient to meet all specified properties, although a cut-off of 23/30 ($p^* = 0.77$) leads to sufficient power for rejecting ineffective respirators and low error rates for highly effective respirators.

Table 3. 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%	11.4%	50.0%	88.6%
19 (54%)	<0.1%	<0.1%	1.6%	19.3%	63.2%	93.8%
20 (57%)	<0.1%	0.1%	3.6%	30.0%	75.0%	97.0%
21 (60%)	<0.1%	0.2%	7.3%	42.7%	84.5%	98.7%
22 (63%)	<0.1%	0.5%	13.5%	56.4%	91.2%	99.5%
23 (66%)	<0.1%	1.4%	22.7%	69.4%	95.5%	99.8%
24 (69%)	<0.1%	3.4%	34.8%	80.5%	98.0%	>99.9%
25 (71%)	<0.1%	7.5%	49.0%	88.8%	99.2%	>99.9%
26 (74%)	0.2%	14.6%	63.5%	94.2%	99.7%	>99.9%
27 (77%)	0.6%	25.5%	76.6%	97.4%	99.9%	>99.9%
28 (80%)	2.0%	40.1%	86.7%	99.0%	>99.9%	>99.9%
29 (83%)	5.5%	56.7%	93.5%	99.7%	>99.9%	>99.9%
30 (86%)	13.2%	72.8%	97.3%	99.9%	>99.9%	>99.9%
31 (89%)	26.9%	85.7%	99.1%	>99.9%	>99.9%	>99.9%
32 (91%)	46.9%	93.9%	99.8%	>99.9%	>99.9%	>99.9%
33 (94%)	69.4%	98.1%	>99.9%	>99.9%	>99.9%	>99.9%
34 (97%)	87.8%	99.6%	>99.9%	>99.9%	>99.9%	>99.9%
35 (100%)	97.5%	>99.9%	>99.9%	>99.9%	>99.9%	>99.9%

The above results indicate that, for a sample size of 35, we should choose a cut-off of no more than 26/35 ($p^* = 0.74$) to achieve sufficiently low type I error rates ($\alpha_{0.8} = 0.146$ and $\alpha_{0.9} = 0.002$). However, for $p^* = 0.74$, the power is slightly less than the specified level of 95% for more effective respirators; specifically, $1 - \beta_{0.6} = 0.942$ (although $1 - \beta_{0.5} = 0.997$, which is above than the specified level of 0.99). Increasing the cut-off to 27/35 ($p^* = 0.77$) leads to exceeding the specified type I error rate for 80% effectiveness ($\alpha_{0.8} = 0.255$) but does give sufficient power for rejecting ineffective respirators, with $1 - \beta_{0.6} = 0.974$ and $1 - \beta_{0.5} = 0.999$. A sample size of 35 therefore gives a test which is extremely close to meeting all specified properties with only a very slight deficit in the power for rejecting an ineffective model (with a result that is only 0.8% lower than the specified 95% level).

Table 4. Percentage of times that a given unit will fail using a sample size of 40

Minimum # of Subjects Required to Pass[1] (%)	Assumed Percentage of the Population Achieving Passing Results					
	90%	80%	70%	60%	50%	40%
21 (53%)	*<0.1%*	*<0.1%*	*0.6%*	*13.0%*	*56.3%*	*92.6%*
22 (55%)	*<0.1%*	*<0.1%*	*1.5%*	*20.9%*	*68.2%*	*96.1%*
23 (58%)	*<0.1%*	*<0.1%*	*3.2%*	*31.1%*	*78.5%*	*98.1%*
24 (60%)	*<0.1%*	*0.1%*	*6.3%*	*43.2%*	*86.6%*	*99.2%*
25 (63%)	*<0.1%*	*0.3%*	*11.5%*	*56.0%*	*92.3%*	*99.7%*
26 (65%)	*<0.1%*	*0.8%*	*19.3%*	*68.3%*	*96.0%*	*99.9%*
27 (68%)	*<0.1%*	*1.9%*	*29.7%*	*78.9%*	*98.1%*	*>99.9%*
28 (70%)	*<0.1%*	*4.3%*	*42.3%*	*87.1%*	*99.2%*	*>99.9%*
29 (73%)	*<0.1%*	*8.8%*	*55.9%*	*92.9%*	*99.7%*	*>99.9%*
30 (75%)	*0.1%*	*16.1%*	*69.1%*	*96.5%*	*99.9%*	*>99.9%*
31 (78%)	*0.5%*	*26.8%*	*80.4%*	*98.4%*	*>99.9%*	*>99.9%*
32 (80%)	*1.5%*	*40.7%*	*88.9%*	*99.4%*	*>99.9%*	*>99.9%*
33 (83%)	*4.2%*	*56.3%*	*94.5%*	*99.8%*	*>99.9%*	*>99.9%*
34 (85%)	*10.0%*	*71.4%*	*97.6%*	*99.9%*	*>99.9%*	*>99.9%*
35 (88%)	*20.6%*	*83.9%*	*99.1%*	*>99.9%*	*>99.9%*	*>99.9%*
36 (90%)	*37.1%*	*92.4%*	*99.7%*	*>99.9%*	*>99.9%*	*>99.9%*
37 (93%)	*57.7%*	*97.2%*	*99.9%*	*>99.9%*	*>99.9%*	*>99.9%*
38 (95%)	*77.7%*	*99.2%*	*>99.9%*	*>99.9%*	*>99.9%*	*>99.9%*
39 (98%)	*92.0%*	*99.9%*	*>99.9%*	*>99.9%*	*>99.9%*	*>99.9%*
40 (100%)	*98.5%*	*>99.9%*	*>99.9%*	*>99.9%*	*>99.9%*	*>99.9%*

The above results indicate that, for a sample size of 40, we should choose a cut-off of no more than 30/40 ($p^* = 0.75$) to achieve sufficiently low type I error rates with $\alpha_{0.8} = 0.161$ and $\alpha_{0.9} = 0.001$. The power, for $p^* = 0.75$, is also sufficient to meet the pre-specified conditions, with $1 - \beta_{0.6} = 0.965$ and $1 - \beta_{0.5} = 0.999$. A sample size of 40 therefore gives a test which meets all specified properties, both in terms of rejecting ineffective respirators and passing effective models. Results for larger sample sizes (not shown here) obviously provide further gains in power with similar type I error rates. For instance, with $n = 45$, several cut-offs, namely 33/45 ($p^* = 0.73$) and 34/45 ($p^* = 0.76$) meet all specified error rates. For $p^* = 0.73$, $\alpha_{0.8} = 0.099$, $\alpha_{0.9} < 0.001$, $1 - \beta_{0.6} = 0.955$ and $1 - \beta_{0.5} = 0.999$. For $p^* = 0.76$, $\alpha_{0.8} = 0.174$, $\alpha_{0.9} = 0.001$, $1 - \beta_{0.6} = 0.978$ and $1 - \beta_{0.5} > 0.999$.

DISCUSSION

The previously-noted random effects models recommended by the FDA (2007) and Zhang and Kolz (2008) may seem to represent a more informative and potentially more powerful approach to modeling fit factors using repeated measures (or donnings) on a subject to estimate the proportion of users exceeding some specified level of fit. Based on a sample of *n* subjects, the hypothesis of sufficient fit is rejected or accepted based on the endpoints of the resulting confidence interval. Estimating the required sample size for such models (prior to conducting the experiment), however, requires knowledge of both the within and between subject variability, which tends to be specific to a given population of users (say with a certain range of facial dimensions or ethnic traits) and the specific respirator being tested. Different respirators, or respirator types, might then equate to different sample sizes for testing effectiveness with sufficient power. For instance, filtering facepiece models might tend to have more consistent fit across a test panel of different facial dimensions as compared to elastomeric models where fit might tend to be vary more with large versus small faces. Hence, for organizations, such as ISO or NIOSH, seeking to develop or implement general recommendations for sample sizes and test panels to be used in respirator testing, the random effects models may not be practically feasible.

Although conservative estimates for the variance components could be used for applying the random effects model (to circumvent lack of preliminary data across various models and subject populations), other aspects of fit testing complicate the use continuous fit factors and confidence limits for setting a passing criteria. One such limitation is measurement error in devices used to assess the total inward leakage. Specifically, fit factors above 200 are commonly measured in practice, but known to be unreliable around or above that threshold. Although use of a log-transformation aids somewhat in mitigating this error, fit factors above any such detection limit can still substantially influence the confidence limits. Although imputation methods (as often used for work exposures below a detection limit) might apply to this problem, we are not aware of any such analyses being conducted with respect to fit testing methods. In contrast, categorization of a subject's fit as either passing or failing, and subsequent application of binomial probabilities, avoids any such measurement issues since a passing result is well within limits of accurately measured fit factors.

Utilizing binomial probability calculations for this application only assumes that a given model has some constant probability, of achieving sufficient fit, across the target population. The validity of this assumption rests to some degree (as is always the case) in testing a representative sample of the target population. To address this issue, the proposed NIOSH test uses the NRFTP, and thus aims to assemble a panel of *n* subjects which approximates the facial dimensions of the working U.S. population, and thus hopefully avoids any systematic differences between the sample data and target population. Although doing so cannot guarantee any specific degree of correspondence between the underlying *p* for the sample versus the underlying *p* for the population, any other methods (such as the random effects model) will require analogous assumptions about the distribution of fit factors across the target population versus the sample or sampling frame. Although one might explore model adjustment for a range of facial dimensions or other predictors (which could for instance be accomplished via a logistic model for predicting the proportion with sufficient fit), doing so would no longer yield a specific passing criteria, and would therefore be difficult to implement in practice.

CONCLUSIONS

The proposed approach uses binomial probabilities to calculate error rates across different assumptions about the respirator's underlying effectiveness, different sample sizes, and different minimum requirements for the percentage of subjects required to pass fit testing. The total inward leakage test is used as an illustration for applying this approach, and a specific algorithm is outlined for determining the required sample size and percentage of subjects required to pass. This approach, as compared to the previously proposed random effects model, substantially simplifies the problem of sample size estimation for respirator fit testing, and simultaneously identifies a specific cut-off for the percentage of test subjects required to pass. The analysis makes minimal assumptions and does not require any preliminary data or knowledge about the underlying distribution of fit factors or partitioning of the variance components; development of the test criteria is based solely on achieving adequately high probabilities for rejecting ineffective respirators, and passing effective respirators, and is thus well suited for estimating sample sizes and passing criteria which are easily interpreted and implemented in practical settings.

DISCLAIMER

The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the National Institute for Occupational Safety and Health.

REFERENCES

1. Hyatt, E.C., J.A. Pritchard, B.J. Held, D.A. Bevis, T.O. Davis, L.A. Geoffrion, A.L. Hack, P.L. Lowry, T.O. Moore, C.P. Richards, and L.D. Wheat: 1974. Respiratory studies for the National Institute for Occupational Safety and Health, July 1, 1972-June 3, 1973, LA-5620-PR. Los Alamos, New Mexico: Los Alamos Scientific Laboratory.

2. National Institute for Occupational Safety and Health (NIOSH): "Approval of Respiratory Protection Devices," *Code of Federal Regulations* Title 42, Part 84.

2003.

3. Hack, A.L., E.C. Hyatt, B.J. Held, T.O. Moore, C.P. Richards, and J.T. McConville: *Selection of respirator test panels representative of U.S. adult facial size.* New Mexico: Los Alamos Scientific Laboratory of the University of California (LA5488); 1974.

4. Hack, A.L., and J.T. McConville: Respirator protection factors: part I - development of an anthropometric test panel. *Am. Ind. Hyg. Assoc. J. 39*:970-975 (1978).

5. **National Institute for Occupational Safety and Health (NIOSH):** Preamble to revised 42 CFR Part 84. *Federal Register* 60, 30355. June 8, 1995.

6. **Coffey C.C., D.L. Campbell, W.R. Myers, Z. Zhuang, and S. Das:** Comparison of six respirator fit-test methods with an actual measurement of exposure in a simulated health care environment: part I – protocol development. *Am. Ind. Hyg. Assoc. J. 59:*852-861 (1998a).

7. **Coffey C.C., D.L. Campbell, W.R. Myers, and Z. Zhuang:** Comparison of six respirator fit-test methods with an actual measurement of exposure in a simulated health care environment: part II - method comparison testing. *Am. Ind. Hyg. Assoc. J. 59:*862-870 (1998b).

8. **Zhuang, Z., C.C. Coffey, P.A. Jensen, D.L. Campbell, R.B. Lawrence, and W.R. Myers:** Correlation between quantitative fit factors and protection factors measured under actual workplace environments at a steel foundry. *Am. Ind. Hyg. Assoc. J.* 64*:730*-738 (2003).

9. **National Institute for Occupational Safety and Health (NIOSH):** Approval of

Self-Contained Breathing Respirators for Emergency Workers in Terrorist Attacks [Online]. NIOSH 42 CFR Part 84, Subpart H, Part 84.63(c). http://www.cdc.gov/niosh/npptl/standardsdev/cbrn/scba/ (2001). Accessed March 2008.

10. **Zhuang, Z. and B. Bradtmiller:** Head and face anthropometric survey of U.S. respirator users. *J Occup. Environ. Hyg.* 2:567-576 (2005).

11. **Zhuang Z., J. Guan, H. Hsiao, and B. Bradtmiller:** Evaluating the representativeness of the LANL respirator fit test panels for the current U.S. civilian workers. Journal of the International Society for Respiratory Protection, 21:83-93 (2004).

12. **Zhuang, Z. and B. Bradtmiller, and R.E. Shaffer:** New respirator fit test panels representing the current U.S. civilian work force. *J Occup. Environ. Hyg.* 4:647-659 (2007).