
Respirator Speech Intelligibility Testing with an Experienced Speaker

Karen M. Coyne* and Daniel J. Barker

U.S. Army Edgewood Chemical Biological Center
Aberdeen Proving Ground, MD 21010

Corresponding author E-mail: karen.m.coyne4.civ@mail.mil

ABSTRACT

The United States' National Institute for Occupational Safety and Health (NIOSH) assesses speech intelligibility of commercial chemical, biological, radiological and nuclear air-purifying respirators using the Modified Rhyme Test (MRT). The speaker's loudness, enunciation, accent, and pronunciation may adversely affect respirator sound conveyance and speech intelligibility. It was postulated that an experienced speaker might induce higher intelligibility scores than a less trained speaker. A modified version of the NIOSH MRT was conducted to evaluate this hypothesis. Twelve respirators that were evaluated previously using the NIOSH method were tested with a unique listener panel on four separate test days. On each test date, the speaker read one word list while wearing each of the twelve masks and one word list without the respirator to the same panel of three listeners. A one way Analysis of Variance (ANOVA) showed that performance ratings (PR) with the experienced speaker ($86.5\% \pm 0.6\%$) were significantly higher than those for the standard method ($84.0\% \pm 0.6\%$). However, the average PRs across all respirators by listener for trials with the experienced speaker were 78.2% to 93.5%. So, while an experienced speaker will yield a small increase in speech intelligibility, there is still a large amount of variability in listener performance.

Keywords: Speech intelligibility, Air-Purifying respirator, Modified Rhyme Test.

INTRODUCTION

Personal protective equipment (PPE) can interfere with the ability of the wearer to perform occupational tasks. First responders such as law enforcement officers and fire fighters sometimes need to wear chemical, biological, radiological, and nuclear (CBRN) PPE when responding to incidents or performing specific operations (e.g., terrorist events, clandestine laboratory responses, and personnel recovery missions). Respirators worn during these operations can interfere with speech transmission and reception, hindering the first responders' ability to communicate with each other and the public.

The National Institute for Occupational Safety and Health (NIOSH) air-purifying respirator (APR) CBRN standard (NIOSH, 2005) assesses speech intelligibility during mask wear using the Modified Rhyme Test (MRT) (House et al., 1965). The MRT consists of 50 six-word lists of monosyllabic English words, most having three sounds in a consonant-vowel-consonant sequence. Listeners must correctly identify single-syllable words spoken by a respirator wearer.

Both the talkers and listeners are variables in speech intelligibility testing (House et al., 1965). The NIOSH test standard prescribes training for both speakers and listeners to try to reduce their variability as much as possible. Test speaker sound output level, enunciation, accent, and pronunciation impact the sound conveyance and speech intelligibility scores of a given respirator. Experience with MRT testing has shown that some speakers tend to over-enunciate and speak more loudly when they are wearing the respirator. To control the speaking portion of the test, House et al. (1965) states that "superior" or "experienced" talkers could be used. Listeners must have normal hearing, be familiar with all the test words, and be attentive throughout the test. However, the MRT was designed so that novice listeners can be used (House et al., 1965; Nye and Gaitenby, 1973).

A prior study that followed the NIOSH format indicated that MRT scores were highly variable (Coyne and Barker, 2010; Coyne and Barker, 2014). It was postulated that an experienced speaker who had participated in many speech intelligibility trials might induce higher intelligibility scores than a less trained speaker. Therefore, a modified version of the NIOSH MRT was conducted to determine whether listener speech intelligibility scores were higher and less variable with an experienced speaker.

METHODS

Twelve NIOSH-certified CBRN APRs used in a previous study (Coyne and Barker, 2010; Coyne and Barker, 2014) were used in this study. The respirators were: 3M FR-M40 (St. Paul, MN), Avon C50 (Avon Protection Systems, Inc., Cadillac, MI), AOSafety Peltor M-TAC (Aearo Technologies, Indianapolis, IN), Draeger CDR 4500 (Draeger Safety Inc., Pittsburgh, PA), MSA Millenium (Mine Safety Appliances Company, Cranberry Township, PA), Survivair Optifit (Sperian Protection, Honeywell Safety Products Australia Pty Ltd, Victoria, Australia), MSA Ultra Elite, Scott M120 (Scott Health and Safety, Monroe, NC), North 5400 (North Safety, Honeywell, Cranston, RI), 3M FR-7800B, Scott M110, and Avon FM12.

Volunteers

Volunteers were recruited from the civilian and military workforce of the U.S. Army Edgewood Chemical Biological Center (ECBC), Aberdeen Proving Ground, MD. The test protocol was approved by the institutional review board at ECBC. All participants completed the Occupational Safety and Health Administration respirator medical evaluation questionnaire. A medical professional reviewed the questionnaires and determined that all volunteers were healthy and free of coronary risk factors. Each of the 13 male volunteers (23.2 ± 5.7 years) gave written informed consent prior to participation in the study. All volunteers were native speakers of American English and had normal hearing. The speaker spoke

without a regional accent and had participated in over five hundred trials over the course of two decades of testing. None of the listener volunteers had participated in any prior intelligibility testing.

Test Procedures

All testing was conducted in the Respiratory Protection Branch's Individual Protective Equipment Test Facility. The trials followed a modified NIOSH MRT communications certification procedure. Deviations from the standard test procedures were: one male volunteer served as the speaker for all trials rather than having one female and four male speakers; the listener panels consisted of three male volunteers rather than one female and two males; and, the speaker read one of twelve word lists while wearing each of the twelve masks and one additional word list without the respirator to the same panel of listeners. There are six MRT word lists in House et al. (1965). Per the ANSI standard (1999), additional lists may be made from these six. A total of 12 word lists were used. With 13 trials in one day, a different list was repeated on each of the four test days. Mask condition was the same for both the speaker and the listeners. All twelve respirators were tested on four separate test days with a different group of listener volunteers for each test day. Therefore, each listener participated in 13 trials. The order of respirators tested on each day was randomized. However, the no respirator control condition was always the first condition tested.

Equipment

All equipment used corresponded to that listed in the NIOSH standard and was the same as that used in the previous study (Coyne and Barker, 2010; Coyne and Barker, 2014). In brief, a stereo amplifier, two loudspeakers, and a pink noise generator (Bruel and Kjaer Type 1405, Naerum, Denmark) were used to transmit pink noise at 60 ± 2 dBA. The noise levels were measured at the position of the listener's ear. The A-weighted fast response setting was used on the sound level meters (Model 322, Center Technology Corporation, Taipei, Taiwan). The meters were calibrated at the start of each test day. Once the meters were calibrated, the room background noise was set.

Data Analyses

Performance Ratings

Performance ratings were computed using the method outlined in the NIOSH (2005) test procedure. Scores were corrected for words misspoken by the speaker and for chance or guessing made possible by the multiple-choice format of the answer sheet. Misspoken words were identified by the test administrator. The score for an individual listener for each trial was calculated using the following equation:

$$\text{Adjusted Score (\%)} = (\# \text{ Correct} - \# \text{ Correct} / 5) / \# \text{ Words Spoken Correctly} \times 100$$

The performance rating for each listener was then calculated as:

$$\text{Performance Rating (\%)} = \% \text{ Adjusted Score with Respirator} / \% \text{ Adjusted Score without Respirator} \times 100$$

The scores from the three listeners were then averaged to obtain the performance rating of the mask.

Statistical Analyses

A one way Analysis of Variance (ANOVA) was performed (SigmaPlot v 12.0, Systat Software, San Jose, CA) to determine if there was a difference between the NIOSH and single speaker methods. A Friedman Repeated Measures (RM) ANOVA on Ranks was performed to determine if there were differences in performance ratings among respirators. A one way ANOVA on Ranks was also performed to determine if there were differences among the listeners. A one way ANOVA was performed to determine if there were differences among test dates for performance ratings and for the MRT scores for the respirator conditions and for the control condition. Two separate two-way RM ANOVA were performed to determine if there were differences in speaker sound level or background noise level among the test dates and from the beginning to the end of each trial. ANOVA on Ranks was used whenever the data were not normal. The Tukey Multiple Comparison Procedure was used to identify groups that differed. All tests were performed at the $p < 0.05$ level.

RESULTS

Performance ratings are shown for each of the twelve respirators in Table I. Data from the prior study (Coyne and Barker, 2010; Coyne and Barker, 2014) using the NIOSH method are also provided. There was a statistically significant difference between the two methods. Performance ratings with the single experienced speaker ($86.5\% \pm 0.6\%$, mean $\pm$ standard error of the mean) were statistically significantly higher than those for the NIOSH method ($84.0\% \pm 0.6\%$). Differences between the average performance ratings for the two methods ranged from 2.7% to 9.0%.

The performance ratings and MRT scores for individual listeners are shown in Table II. For individual listeners, the performance ratings ranged from $78.2\% \pm 4.5\%$ to $93.5\% \pm 9.6\%$. A one-way ANOVA on Ranks with Tukey post-hoc test showed that subjects 287 and 288 both had significantly higher performance ratings than subjects 337 and 338.

There was a significant difference among the test dates as determined by the ANOVA on Ranks. Average performance ratings on the fourth date ($80.8\% \pm 8.2\%$) were significantly lower than the performance rating on the other three dates ($91.2\% \pm 8.3\%$, $87.1\% \pm 5.9\%$, $86.9\% \pm 8.3\%$). MRT scores for the respirator conditions were also significantly lower on the fourth test date ($79.5\% \pm 7.9\%$) than the scores on the other three dates ($86.1\% \pm 7.6\%$, $87.1\% \pm 5.9\%$, $85.5\% \pm 8.2\%$). Average control scores for days 1 through 4 respectively, were $94.4\% \pm 2.8\%$ (mean $\pm$ standard deviation), $100\% \pm 0\%$, $98.4\% \pm 1.4\%$, and $98.4\% \pm 1.4\%$. An ANOVA showed that there was a statistically significant difference in the control scores among test days. The Tukey post hoc test indicated that average control scores on the second date were different from those on the first test date. No other differences were indicated.

There were no statistically significant differences in either background noise or speaker sound level among test dates or during the duration of each test (beginning, middle, and end of each trial). Background noise levels for the four tests dates were 60.2 ± 0.5 , 60.3 ± 0.5 , 60.5 ± 0.4 , and 60.2 ± 0.3 , respectively. Speech sound levels for the four test dates were 80.1 ± 3.5 , 80.6 ± 3.2 , 80.4 ± 2.6 , and 80.3 ± 2.7 , respectively.

Table I. Performance Ratings for Single Speaker MRT and NIOSH Method

Mask	Single Speaker Performance Rating (%)	NIOSH Format Performance Rating¹ (%)	Difference² (%)
Millenium	92.5 ± 6.4	90.5 ± 1.9	2.0
FR-M40	91.4 ± 8.0	87.8 ± 3.2	3.6
Ultra Elite	84.4 ± 9.4	86.7 ± 3.5	-2.5
M110	90.9 ± 7.6	85.8 ± 5.9	5.1
5400	84.5 ± 5.1	84.9 ± 5.3	-0.5
M120	90.9 ± 6.0	84.8 ± 3.3	6.0
C50	86.9 ± 5.2	84.1 ± 6.5	2.8
FM12	86.5 ± 4.5	83.8 ± 2.7	2.7
Optifit	83.4 ± 9.5	80.8 ± 5.0	4.5
4500	81.2 ± 5.6	80.8 ± 1.3	0.3
M-TAC	88.1 ± 9.3	79.1 ± 4.0	9.0
FR-7800B	75.2 ± 9.2	78.1 ± 4.2	-2.7

¹ Data are from Coyne and Barker (2010)² Difference is the single speaker performance rating minus the NIOSH performance rating

Table II. Average Performance Ratings and MRT Scores for Individual Subjects

Subject #	% PR	Score %	Control Score
287	93.3 ± 7.3	86.8 ± 6.7	92.8
288	91.6 ± 9.4	85.0 ± 8.7	92.8
289	88.5 ± 7.9	86.4 ± 7.7	97.6
308	88.8 ± 4.5	88.8 ± 4.5	100
309	87.8 ± 6.5	87.8 ± 6.5	100
310	84.6 ± 6.3	84.6 ± 6.3	100
322	88.7 ± 9.3	84.1 ± 6.5	97.6
323	85.9 ± 8.5	86.6 ± 9.1	97.6
324	86.0 ± 7.6	86.0 ± 7.6	100
336	83.8 ± 6.8	81.8 ± 6.6	97.6
337	80.5 ± 7.7	78.6 ± 7.5	97.6
338	78.2 ± 9.6	78.2 ± 9.6	100

DISCUSSION

The NIOSH MRT test standard specifies training of the speakers. However, speakers may not have experience with MRT trials prior to being asked to participate in a certification test. For this study, one experienced speaker who had participated in many MRT trials was used rather than five speakers (who may be experienced or novice) as specified by the NIOSH test standard. The trials with the experienced speaker resulted in performance ratings that were 2.5% higher on average for the twelve tested respirators. However, differences between the trials with the experienced speaker and the trials with the standard speaking panel ranged from -2.7% to 9%. While scores for nine of the respirators were higher with the experienced speaker, two were less, and one was essentially the same. So, while the difference between the two test methods was significant, the experienced speaker trials did not always result in higher performance ratings than the NIOSH method.

While the difference was statistically significant, it is important to consider whether or not the difference is practically significant. For ten of the respirators, the experienced subject did worse than the highest performing speaker during the NIOSH-formatted trials. And, for 7 respirators, the second highest performing speaker from the standard trials scored higher than the experienced speaker. The average score of the highest performing speaker from each NIOSH-formatted trial was $90 \pm 4\%$ while it was $87 \pm 5\%$ for the second highest performing speaker. Additionally, performance rating rankings between the two methods were similar. The Millenium and the FR-M40 received the highest two performance ratings for both the single speaker trials and for the NIOSH-formatted trials. The FR-7800B ranked #12 by both methods. The M-TAC did much better with the single speaker (rank #5) than with the panel of speakers (rank #11). The Ultra-Elite dropped from rank #3 with the NIOSH speakers to #9 with the single speaker. The other respirators had similar performance ranks between the two methods and were within 1-3 position changes between the two methods. So, it does not seem that the statistically significant difference between the two methods is of practical significance.

Listeners were trained on the test procedures and familiarized with the word lists. Even so, there was variability in the listener scores. Subjects 287 and 288 had performance ratings significantly higher than subjects 387 and 388. It was noted that subjects 287 and 288 had lower control scores (92.8%) than the other subjects. In fact, their control scores were almost two standard deviations from the average control score across all subjects. It was possible that their performance ratings were higher than the other subjects because their control scores were the lowest. The performance rating is determined by dividing the MRT score from the respirator trial by the MRT score without the respirator (control). If one subject has a lower control score than another, the same MRT score with a respirator will result in a higher performance rating. For example, if subject 1 has a control score of 90 and subject 2 has a control score of 100, an MRT score with a respirator of 85% for both subjects will result in a 94% performance rating for subject 1 ($85\%/90\%$) and a 85% performance rating for subject 2 ($85\%/100\%$). At least part of the reason subjects 287 and 288 had significantly higher performance ratings may have been due to their lower control scores. The ANSI Standard (1999) recommends that listener training trials be completed until there is a plateau in the listener scores. Perhaps future tests should include multiple control trials to ensure that a true baseline has been assessed.

The two subjects with the highest scores participated on the first day while the two with the lowest scores participated on the fourth day. The performance ratings and respirator MRT scores were significantly lower on the fourth test date compared to the other three test dates. While the ANOVA indicated that there were significant differences in control scores on the test dates, the Tukey post-hoc test failed to find a difference at the $p < 0.05$ level. All testing took place in the same room with the same speaker and equipment. Speech sound levels and background noise levels were not significantly different among the test days. The test administrator did not note anything that could explain the differences other than individual subject variability. While the speaker was highly trained, it is possible that he enunciated more clearly on some days, although this was not noticed at the time. The differences

are likely just due to the highly subjective nature of the MRT test and the fact that many factors impact the performance rating.

As noted in House et al. (1965), Fairbanks and Miron (1957) found that when talkers change their vocal effort, the consonant-vowel ratio may change. Therefore, it is possible that when the volunteer is wearing the respirator, the natural inclination to raise one's voice may also be changing the time duration of vowels and impacting speech intelligibility. In fact, House et al. (1965) found that listeners had consistently higher scores with one experienced talker compared to another. Further analysis showed that one talker had a lower consonant-vowel ratio for speech sound level and duration, tending to draw out the vowels. So, even "experienced" talkers may yield different results. The NIOSH test requires five speakers. Averaging scores across five speakers would tend to minimize this effect.

The ANSI (1999) standard states that at least 5 talkers and 5 listeners will be used, with the number of talkers being equal to or greater than the number of listeners due to the fact that there is greater variability in intelligibility scores among talkers compared to listeners. The NIOSH standard specifies five speakers and three listeners and requires all listeners and speakers to go through a minimum of three lists without the respirator and three lists with the respirator. Based on the results of this limited study, the standard level of training required by NIOSH appears to be sufficient. However, while House et al. (1965) and Nye and Gaitenby (1973) state that novice listeners can be used, the ANSI standard (1999) specifies that listeners be trained until their performance reaches a plateau. It may help to reduce listener variability if test administrators recorded listener's scores during training and continued that training until the scores had leveled off.

CONCLUSIONS AND RECOMMENDATIONS

Performance ratings with the single experienced speaker were statistically significantly higher than those for the NIOSH method. However, the average performance rating across all respirators by listener during the trials with the experienced speaker ranged from $78.2\% \pm 9.6\%$ to $93.5\% \pm 7.3\%$. So, while an experienced speaker will yield a small increase (2.5%) in performance rating, there is still a large amount of variability in listener performance even though speaker variability was controlled as much as possible. This small difference would not impact the results much and it is debatable whether it is a meaningful difference. It is likely that the training specified in the NIOSH standard is sufficient. Future work will examine listener performance with an automated speaking headform, so that speaker enunciation, timing, and sound level are the same for all listeners regardless of the respirator condition or test day.

Acknowledgements

The authors would like to thank the Office of Law Enforcement Standards, National Institute of Standards and Technology (Gaithersburg, MD), and the Department of Homeland Security, Science and Technology Directorate (Washington, DC) for funding this project.

REFERENCES

- American National Standards Institute. (1999) *American National Standard Method for Measuring The Intelligibility of Speech over Communication Systems*. American National Standards Institute (ANSI). New York, New York, ANSI S3.2-1989 (R1999).
- Coyne KM, Barker DJ. (2010) Speech Intelligibility While Wearing Civilian Full-Facepiece Air-Purifying

-
- Respirators; ECBC-TR-779; U.S. Army Edgewood Chemical Biological Center: Aberdeen Proving Ground, MD, UNCLASSIFIED Report.
- Coyne KM, Barker DJ. (2014) Speech Intelligibility While Wearing Full-Facepiece Air-Purifying Respirators. *J. Occup. Environ. Hyg.* 11:751-6.
- Fairbanks G, Miron MS. (1957) Effects of Vocal Effort upon the Consonant-Vowel Ratio within the Syllable. *J. Acoust. Soc. Am.* 29:621-626.
- House AS, Williams CE, Hecker MHL, and Kryter KD. (1965) Articulation Testing Methods: Consonantal Differentiation with a Closed-Response Set. *J. Acoust. Soc. Am.* 37:158-166.
- National Institute for Occupational Safety and Health. (2005) *Determination of Communication Performance Test for Speech Conveyance and Intelligibility of Chemical Biological Radiological and Nuclear (CBRN) Full-Facepiece Air-Purifying Respirator Standard Test Procedure (STP)*. National Institute for Occupational Safety and Health (NIOSH): Washington, DC. Procedure No. CET-APRS-STP-CBRN-0313.
- Nye PW, Gaitenby JH. (1973). Consonant Intelligibility in Synthetic Speech and in a Natural Speech Control (Modified Rhyme Test Results. In: *Haskins Laboratories Status Report on Speech Research SR-33*. New Haven, CT: Haskins Laboratories.