

Working title: Using a simulation test to assess warning signal detection in the real-world noise environment of Dutch train cabins

## **Discussion**

*What are the most important findings? Are the two hypotheses confirmed or not?*

*Regarding the first experiment (reproducibility of the test), how can the findings be compared with what others have found?*

To our knowledge, this is the first study that describes the reproducibility of a simulation test that assesses the detectability of warning signals in the acoustic environment of a specific workplace.

- How is the reproducibility of our test compared to the reproducibility of tone audiometry?
- How is the reproducibility of our test compared to the reproducibility of standardized signal-in-noise tests?
- In what way did the reproducibility differ between the different train types? Are there explanations for it?
- Has the test sufficient reliability to continue further development
- What is the clinical implication of the described smallest detectable change?

*Regarding the first hypothesis of the second experiment (highest association for the frequencies at which the signal shows it most prominent signal peaks), how can the findings be compared with what others have found?*

- Who have indicated that frequency specificity might be relevant?
- It has been argued that a warning signal should not contain signal peaks above 3100 Hz, why not? Does our study underline this recommendation?
- What is a difference between our study and the statement of others about frequency specificity?

*Regarding the second hypothesis of the second experiment (adding speech in noise would increase the explained variance of the test), how can the findings be compared with what others have found?*

- Who have indicated that tone audiometry and speech in noise would be both relevant? Were they relevant?
- Why is the explained variance (much) higher for the ATP signal detection test compared to the ATP signal detection test?
- What can explain the fact that the hearing-in-noise tests have a higher association with the aided masked thresholds, than for the unaided measurements? What does this imply?
- What factors could improve the regression model?

*What are the limitations of the study? What other factors could have influenced my findings. Have I reported everything that could make my findings invalid?*

- Is there a possible source for measurement bias?
- Is there a possible source for selection bias?
- What are the implications of using a simulation test as the functional ability to detect warning signals?
- How was the power

*Do my interpretations contribute some new understanding of the problem that I have investigated? In which case do they suggest a shortcoming in or an advance on the work of others?*

- What are the advantages of using a simulation test rather than the outcomes of conventional hearing tests?
- So, can the simulation test already be used to assess auditory fitness in locomotive engineers?
- Could a simulation test be useful for other professions too?
- What future research would be needed?