**Supplementary material B. Simulation of power of our study**

The test statistics’ size depends on the effect size and the number of observations. Generally speaking, with more observations, a small effect is more detectable through the statistical test. Because our sample size is relatively small, it is possible that the statistically significant difference in scores between children with CIs and children with normal hearing was not detected by our statistical test. To assess the power of the test (i.e., the probability of correctly rejecting the null hypothesis), we run two simulations by replicating our method described in the Data Analysis section. **Supplementary material B1** shows the relationship between the power and the average score difference between the children with CIs and the reference score, conditioned if the standard deviation of the population score of children with CIs is one. By refereeing the power of 0.8, which is the conventional reference point, it is evident that a more than 0.8 average score difference is statistically detectable for any of our age groups.

**Supplementary material B1. The simulation results regarding the power by the effect size in our study**

グラフ, 折れ線グラフ

自動的に生成された説明

Note: the score in the population is assumed to be distributed with standard deviation of 1.

Nevertheless, the power also depends on the population’s size’s standard deviation. If the population’s standard deviation is large, a large average difference is necessary to achieve the same statistical power (i.e., 0.8) (see **Supplementary material B2**). Notably, the descriptive statistics of our samples’ domain-level scores is reported in **Supplementary material B3**, which shows that the standard deviation G2 and G3 are below 1 for all the item except for the total score of G3. G1 shows higher standard deviation for many items than G2 and G3 but the sample size of G1 is larger than G2 and G3. Therefore, even the moderately higher standard deviation of 1.25, the minimally required difference is 0.75.

**Supplementary material B2. The simulation about the relationship between the size of standard deviation and the required size of average difference given the power of 0.8**

グラフ, 折れ線グラフ

自動的に生成された説明

**Supplementary material B3. Descriptive statistics of domain level scores by each age group**

|  |  |  |  |
| --- | --- | --- | --- |
|  | G1 | G2 | G3 |
| A | -0.96  (1.75) | 0.07  (0.92) | -0.37  (0.86) |
| B | -0.70  (0.90) | 0.28  (0.95) | -0.61  (1.06) |
| C | -0.78  (1.22) | -0.10  (0.89) | -0.75  (1.12) |
| D | -0.72  (1.26) | -0.08  (1.07) | -0.48  (1.43) |
| Total | -0.92  (1.38) | 0.01  (0.96) | -0.64  (1.21) |

Note: The number without and with parenthesis indicate the average score and standard deviation, respectively.

The practical insights from these simulations and comparison with the descriptive statistics are twofold. First, suppose the average score of children with CIs score is less than -0.8 from the reference score (i.e., children with normal hearing). In that case, we may have detected its statistically significant difference through our data analysis. Conversely, if the difference were smaller than 0.8, it may have been the case that we could have missed observing the statistical difference. Second, and relatedly, we have reported in the main text that G3 marked a lower score in (D) *communication skills* and *total skill* at the domain-level score, but these were statistically insignificant. This may be attributed to the smaller sample size of G3 and its larger standard deviation (i.e., 1.43 for communication skills and 1.21 for Total score). On the other hand, (A) *language skill* may not be statistically significant even if the sample size would be larger because the difference is relatively small (i.e., -0.37) and its standard deviation is not large (i.e., 0.86).

***Technical detail of the simulation***

Let the number of children in a group be . To conduct our simulation study, we first generated test scores with the same length of that follow the normal distribution with a particular average and a standard deviation. We then randomly resampled the test scores from1,000 times and calculated the bootstrap average. If the 95% quantile of this average is outside of zero, we conclude this as a true judgment; otherwise, it is a false judgment. We repeated this procedure 1,000 times and calculated how many times the true judgment was achieved, which can be regarded as power at the given level of difference. To make the figure in Supplementary material B1, we repeated this simulation by changing the difference from 0 to 1 and increasing the difference by 0.01. Similarly, for Supplementary material B2, we repeated the simulation by changing the range of the standard deviation (from 0.5 to 2.0) and the range of the difference (from 0 to 2.5). We then grid-searched the combination that achieved the power of 0.8.