ASHA American Speech-Language-Hearing Association

LSHSS

Research Note

Does Assessor Masking Affect Kindergartners' Performance on Oral Language Measures? A COVID-19 Era Experiment With Children From Diverse Home Language Backgrounds

Sarah Surrain,[a] Michael P. Mesa,[a] Mike A. Assel,[a] and Tricia A. Zucker[a]

[a] The Children's Learning Institute, The University of Texas Health Science Center at Houston

ARTICLE INFO

Article History:
Received December 10, 2022
Revision received March 12, 2023
Accepted April 3, 2023

Editor-in-Chief: Amanda J. Owen Van Horne
Editor: Kelly Farquharson

https://doi.org/10.1044/2023_LSHSS-22-00197

ABSTRACT

Purpose: The ongoing COVID-19 pandemic has prompted changes to child assessment procedures in schools such as the use of face masks by assessors. Research with adults suggests that face masks diminish performance on speech processing and comprehension tasks, yet little is known about how assessor masking affects child performance. Therefore, we asked whether assessor masking impacts children's performance on a widely used, individually administered oral language assessment and if impacts vary by child home language background.
Method: A total of 96 kindergartners (5–7 years old, n = 45 with a home language other than English) were administered items from the Clinical Evaluation of Language Fundamentals Preschool–Second Edition Recalling Sentences subtest under two conditions: with and without the assessor wearing a face mask. Regression analysis was used to determine if children scored significantly lower in the masked condition and if the effect of masking depended on home language background.
Results: Contrary to expectations, we found no evidence that students scored systematically differently in the masked condition. Children with a home language other than English scored lower overall, but masking did not increase the gap in scores by language background.
Conclusions: Our results suggest that children's performance on oral language measures is not adversely affected by assessor masking and imply that valid measurements of students' language skills may be obtained in masked conditions. While masking might decrease some of the social determinants of communication (e.g., recognition of emotions), masking in this experiment did not appear to detract from children's ability to hear and immediately recall verbal information.
Supplemental Material: https://doi.org/10.23641/asha.23567463

Masking in schools is a complex and politicized issue among educators, policymakers, and parents and of particular importance to clinicians and educators making decisions based on assessments administered when wearing a face mask (American Speech-Language-Hearing Association, 2022; Vergara et al., 2022). In the interest of public health during the COVID-19 pandemic, many educators and clinicians conducted in-person assessments while wearing face masks; others conducted assessments remotely over videoconferencing applications. Although there is emergent evidence that language scores obtained from remote testing are comparable to those obtained face-to-face (Castilla-Earls et al., 2022), no studies to date have examined the effect of assessor masking on young children's performance on individually administered language measures. Given that masking in schools may continue because of new COVID variants, seasonal surges in respiratory viruses, and the

Correspondence to Sarah Surrain: sarah.surrain@uth.tmc.edu. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

need for immunocompromised individuals to protect themselves, it is critical that we understand whether assessor masking affects child performance and if some groups of children are affected more than others. If children score worse when their assessor is masked, this would increase overidentification of children needing services. However, if children's scores are largely unaffected by masking, then scores obtained during the pandemic or future masked administrations can be taken as valid measures of child performance.

The Importance of Visual Information for Language Comprehension

It makes intuitive sense that face masking might interfere with children's comprehension, as we know that children and adults use visual information when processing speech. One well-studied phenomenon demonstrating audiovisual integration in speech perception is the McGurk effect (McGurk & MacDonald, 1976). In this paradigm, participants who are presented with mismatched audio and visual stimuli—for example, they hear the syllable *ba* while seeing a face articulate the syllable *ga*—tend to perceive a third syllable (e.g., *da*). In a related line of research, studies have found that infants and toddlers look at speakers' mouths more than at their eyes, as observed with 18- to 20-week-olds (Kuhl & Meltzoff, 1982), 4- to 12-month-olds (Lewkowicz & Hansen-Tift, 2012), and 14- to 18-month-olds (Hillairet de Boisferon et al., 2018). Adults, in contrast, tend to focus slightly more on speakers' eyes than their mouths (Morin-Lessard et al., 2019), which could suggest that children rely more on mouth movements than more mature language users. Evidence from children learning two languages also supports this idea. Although the evidence is mixed, some studies have found that simultaneous bilingual children fixate on mouths more than monolingual children, especially when their developing languages are linguistically similar (e.g., Catalan–Spanish; Birulés et al., 2019, with children aged 15 months and 4–6 years; Pons et al., 2015, with children aged 4–12 months). Given children's propensity to look at speakers' mouths, it follows that removing this source of information from view by masking could attenuate their ability to process speech.

On the other hand, research on the developmental course of audiovisual integration has found that the ability to use mouth movements to discern what is being said (i.e., lipreading) develops slowly (Massaro et al., 1986, with children aged 4–6 and 6–10 years; Ross et al., 2011, with children aged 5–14 years). If young children are less proficient than adults at integrating visual cues from mouth movements with the sounds they hear, it is plausible that masking would have a lesser impact on children's comprehension compared to adults. However, these studies were conducted with small samples of primarily monolingual English-speaking children. The current sample included both monolingual kindergartners and children using English at school and other languages at home.

The Effect of Masking on Speech Perception and Comprehension

Most research on the effect of masking on speech perception and comprehension has been conducted with adults. These experiments have found that face masks make language comprehension more effortful and error-prone (Giovanelli et al., 2021; Haider et al., 2022; Toscano & Toscano, 2021; Truong et al., 2021). The effect of masking on adult comprehension is attributed to two main factors: degrading the audio signal itself (e.g., Corey et al., 2020) and obscuring visual information that adults rely on to supplement the audio signal, especially in noisy environments (Sönnichsen et al., 2022). Given the consistent finding that masking interferes with comprehension in adults, it follows that children, whose communication skills are still developing, would potentially be impacted even more by face masking.

Few studies have investigated the effect of face masking on child language comprehension, with two recent exceptions. Lalonde et al. (2022) examined the effect of masking on consonant recognition in three groups: children with hearing loss (aged 7–18 years), their adult family members with normal hearing, and their siblings with normal hearing (aged 7–19 years). They found that masking affected the adults and children with normal hearing equally. However, when transparent masks were used that allowed the mouth to be seen, the adults with normal hearing benefited the most and the children with normal hearing benefited the least. Another study by Schwarz et al. (2022) tested adults' and 8- to 12-year-old children's ability to repeat the last word of a sentence under different masking conditions. They found that both adults and children were slower and made more mistakes with an acoustic mask (sound was degraded), but children were less affected by a visual mask (mouth was covered). Importantly, the children in both extant studies were in middle childhood or older. We currently lack research on the effect of masking on speech comprehension in younger children and children who are learning multiple languages, two groups that are routinely assessed in schools using individually administered oral measures.

In summary, the research suggests that children's comprehension is affected by masking, but perhaps due more to degradation of the audio signal than the absence of visual information. Although research on young children's fixations suggests that children rely on mouth movements to supplement auditory information, evidence

for the protracted developmental timeline of lipreading and the emerging evidence on masking imply that children do not benefit from seeing speakers' mouths to the same degree as adults. The current study aims to provide ecologically valid evidence of the effect of masking on child language performance in school settings.

The Current Study

This study is the first we know of to examine the effect of assessor masking on child performance on oral language measures with younger children (aged 5–7 years). In a sample of 96 kindergartners from both monolingual and bilingual backgrounds, we administered items from the Clinical Evaluation of Language Fundamentals Preschool–Second Edition (CELF-P2) Recalling Sentences subtest in a quiet location, with and without the assessor wearing a mask. We address the following questions: (a) Do children perform systematically worse when the assessor is wearing a face mask? (b) Does masking increase discrepancies between children from different language backgrounds? We hypothesized that children would perform lower on items in the masked condition compared to the unmasked condition and that this effect would be larger for children learning an additional language at home.

If children perform systematically worse on items in the masked condition, or if the effect of masking is greater for children learning an additional language at home, this would imply that scores collected by masked assessors may not be valid representations of child language ability. Conversely, if there are no differences by masking condition and no interaction effects with language background, this would imply that the effect of masking may not be large enough to systematically deflate scores and, thus, valid scores on language measures may be obtained by adults wearing face masks for children from different language backgrounds.

Method

The participants in this study were kindergartners (aged 5–7 years) who were part of two larger research studies in three urban public school districts in the southern United States in the spring of 2022 (referred to hereafter as Study 1 and Study 2, although the data were collected concurrently). Study 1 (n = 43) was a shared book-reading intervention supporting English vocabulary development in typically developing children who met a minimum threshold of English proficiency. Eligible children had raw scores of 16 or higher (equivalent to a standard score of 68; age equivalent of 3;6 [years;months]) on the Woodcock Johnson Picture Vocabulary subtest (Schrank et al., 2014). Study 2 (n = 53) was a reading intervention for struggling readers. Eligible children met two criteria based on kindergarten benchmarks in early literacy skills: (a) They scored below 17 on the Dynamic Indicators of Basic Early Literacy Skills, Eighth Edition Letter Naming Fluency subtest (University of Oregon, 2019), on which a score of 16 corresponds to a percentile ranking of 41 at the beginning of kindergarten; (b) they scored below 6 on the Comprehensive Test of Phonological Processing Sound Matching subtest (Wagner et al., 1999), which corresponds to a scaled score of 8 (i.e., 25th percentile). Children classified as English learners and those with mild-to-moderate disabilities (e.g., attention-deficit/hyperactivity disorder, speech delays) were included in Studies 1 and 2, but children with severe learning disabilities were excluded.

Our total sample included 96 children (43% girls, 57% boys; n = 41 girls, 55 boys) who were, on average, 74.51 months old (SD = 4.59 months, exact age missing for five students). Based on parent report, 47% of the children (n = 45) lived in homes where only English was spoken, whereas 45% (n = 43) lived in homes where a non-English language was spoken (the remaining 8% did not provide home language information). The non-English languages spoken at home were Spanish (n = 40), Amharic (n = 1), or not specified (n = 2). Children's most used language was English (78%, n = 75), Spanish (11%, n = 11), or both Spanish and English (2%, n = 2), with the remaining 8% missing. Precise data on the extent of children's exposure to a non-English home language were not collected. Based on parent report, 55% (n = 53) of the children were Hispanic/Latino, and 35% (n = 34) were not Hispanic/Latino, with 9% missing. On a separate item about their child's race, parents reported that their child was Black or African American (33%, n = 32), White or Caucasian (23%, n = 22), American Indian or Alaska Native (2%, n = 2), or Asian (1%, n = 1) or selected more than one race or Other (14%, n = 13), with 27% (n = 26) missing.

Study Tasks and Procedure

To test the effect of assessor masking on child performance, we administered items from the Recalling Sentences subtest from the English CELF-P2 (Wiig et al., 2004) in two conditions: with and without the assessor wearing a face mask. This procedure was added to the posttest battery of the two larger studies. Parents provided written informed consent and children provided verbal assent following procedures approved by our local institutional review board (MS-19-0527, MS-18-0392). As noted, all testing occurred one-on-one in quiet locations within children's schools, such as a conference room or empty classroom.

The Recalling Sentences subtest is an individually administered oral language measure in which children are

asked to repeat sentences of increasing length and complexity, tapping into children's language knowledge and phonological working memory. The assessor counts the child's errors, with each word omitted, added, substituted, or changed counted as one error. Items are scored based on the number of errors (four or more errors = score of 0, two to three errors = 1, one error = 2, and no errors = 3, except for Items 1 and 2 that each have a maximum score of 2), with raw scores ranging from 0 to 37. The CELF-P2 manual reports a test–retest reliability of .88 and an average internal consistency (coefficient alpha) of .88. This measure is widely used for both research and clinical purposes and has been found to have better diagnostic accuracy than CELF-P2 composite scores for identifying bilingual children with atypical language development (Rose et al., 2022). We selected this measure because it was not already part of the posttest battery for either of the larger studies and because judgments based on it (and other similar measures) hold consequences for individual children, school systems, and research findings. The CELF-P2 version of Recalling Sentences was selected instead of the Clinical Evaluation of Language Fundamentals–Fifth Edition (CELF-5) version of the same subtest because it has fewer items (13 vs. 26), and thus, we were able to administer all items to all children without overly taxing the children.

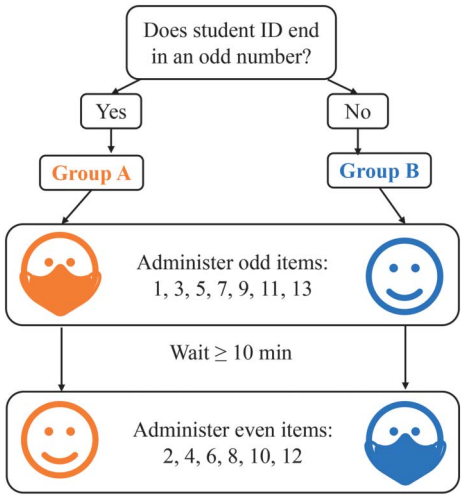
Standardized Language Measure

The Sentence Comprehension subtest from the CELF-5, which was included in the posttest battery for both of the larger studies, was used to control for children's receptive language comprehension in English without variation in assessor masking. The Sentence Comprehension subtest measures children's ability to understand spoken sentences of increasing difficulty. The assessor orally presents a sentence, and the child selects a picture that illustrates the meaning of the sentence. The Sentence Comprehension subtest has acceptable reliability for younger students (Wiig et al., 2013). This measure was administered while masked for all children, as it was not part of the masked/unmasked experiment.

Masking Experiment Procedure

Children's ID numbers were used to randomly assign them to two groups (A and B), and the 13 Recalling Sentences items were administered in two blocks (odd-numbered items and even-numbered items; see Figure 1). Children with ID numbers ending in an odd number were assigned to Group A, and their assessor administered the first block of items wearing a face mask and the second block without a face mask. Children with ID numbers ending in an even number were assigned to Group B, and their assessor administered the first block with no face mask and the second block with a face mask. The blocks were administered about 10 min apart, with an unrelated measure administered between blocks and no additional training or practice between the blocks. Thus, all children experienced both the masked and unmasked conditions, with their group determining the order. The assessors were research staff trained to (a) maintain similar volume and tone in both masking conditions, (b) avoid any verbal reference to their change in masking status between blocks, and (c) wear a surgical face mask or an N95 mask.

Figure 1. Masking experiment procedure flowchart.



Because we were only interested in the effect of assessor masking, child masking was not manipulated as part of this experiment. At the time of data collection in Spring 2022, masking was optional and the proportion of students who continued to mask varied by school district. Assessors reported that approximately 20% of Study 1 children and 75% of Study 2 children were masked during the assessment. For this experiment, instead of following the discontinue rule of three consecutive zeros, we administered all items to obtain complete item-level data. We then calculated raw sum scores of all items for each block, which can range from 0 to 20 for the seven odd items and 0 to 17 for the six even items. To ensure that this strategy resulted in two groups that did not significantly differ in observed characteristics, we tested for significant differences across groups in study sample (1 or 2), gender, age, home language, race/ethnicity, and sentence comprehension score and found none (see Table 1 for descriptive statistics by group).

Analytic Approach

We first conducted exploratory analyses to determine the interitem reliability of each block of items, describe children's performance overall, visually inspect

Table 1. Demographic characteristics and descriptive statistics by group.

Variable	Group A (n = 47) M (SD) or %	Group B (n = 49) M (SD) or %	Significant differences
Age (in months)	73.80 (4.57)	75.19 (4.59)	*ns*
Female	40.43	44.90	*ns*
Ethnicity			*ns*
Hispanic/Latino	59.57	51.02	
Not Hispanic/Latino	31.91	38.78	
Missing ethnicity	8.51	10.20	
Race[a]			*ns*
Black or African American	36.17	30.61	
White or Caucasian	27.66	18.37	
American Indian or Alaska Native	0.00	4.08	
Asian	2.13	0.00	
More than one race or other	12.77	14.29	
Missing race	21.28	32.65	
Home language other than English	44.68	44.90	*ns*
Child's most used language not English	14.89	12.24	*ns*
Missing language data	8.51	8.16	
CELF-P2 RS odd-item sum score	10.64 (4.77)	10.80 (5.06)	*ns*
CELF-P2 RS even-item sum score	l8.91 (4.41)	9.18 (3.88)	*ns*
CELF-5 SC raw sum score	16.39 (7.63)	17.90 (5.41)	*ns*

Note. *ns* = not significant, based on chi-square tests for categorical variables and two-sample t tests for continuous variables; CELF-P2 RS = Clinical Evaluation of Language Fundamentals Preschool–Second Edition, Recalling Sentences subtest; CELF-5 SC = Clinical Evaluation of Language Fundamentals–Fifth Edition, Sentence Comprehension subtest.

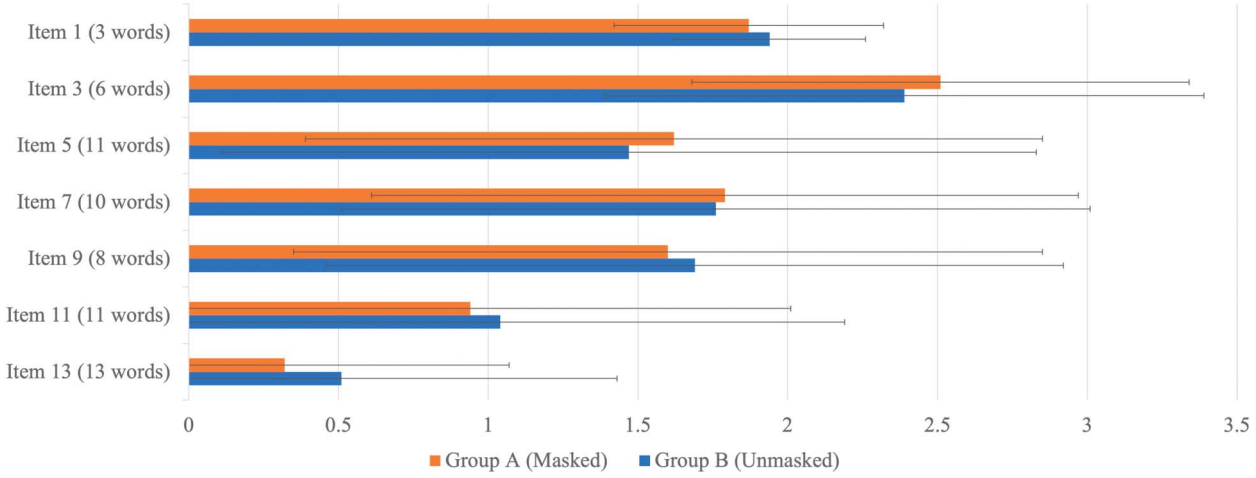
[a]Missingness in this variable was particularly high among parents who identified their child as Hispanic/Latino (n = 19).

each group's performance on odd versus even numbered items, and investigate correlations among variables. We then fit a series of linear ordinary least squares models, in Stata Version 17.0, predicting the raw sum scores for odd and even items separately. The key independent variable was group, as we expected that Group A would score lower on the odd items and Group B would score lower on the even items, if assessor masking negatively affected child performance. We controlled for child gender, home language, and children's performance on the CELF-5 Sentence Comprehension subtest to increase the precision of our estimates, as these variables were correlated with the dependent variables. In addition to testing for statistical significance, we interpreted the effect size using the unstandardized coefficients of the key predictor variables (i.e., Group A or B) in our final models. These coefficients represent the predicted point difference in raw scores between the group that was administered the items by a masked assessor and the group that was administered the items by an unmasked assessor. Predicted raw score differences were interpreted by determining their potential impact on scaled scores for children at the sample mean in age and raw score, and the magnitude of the estimated impact on scaled scores was interpreted based on clinical judgment. We also calculated the change in R^2 when the key predictor (group) was removed from the final model to report the percentage of the variance explained by masking. Next, to address whether masking effects varied by home language, we tested interactions between home language and group. If masking interferes with the performance of children with a non-English home language more than their peers who only speak English at home, we would expect to see a significant, negative Group × Non-English Home Language interaction. Finally, we conducted two sensitivity checks: (a) We tested for interactions between sentence comprehension score and group to determine whether students with lower language skills were disproportionately affected by assessor masking, and (b) we refit all models controlling for intervention group: Study 1 treatment group, Study 2 treatment group, or the business-as-usual control group for Studies 1 and 2 (see Supplemental Materials S1–S5).

Results

First, we calculated the internal consistency for all 13 items and for each block of items separately, using Cronbach's alpha. Values were $\alpha = .89$ for all items, $\alpha = .80$ for the seven odd-numbered items, and $\alpha = .81$ for the six even-numbered items, suggesting strong internal consistency in our sample. Second, we calculated the overall

Figure 2. Performance on odd-numbered items by group. The highest score was 2 for Item 1 and was 3 for all other items.



raw scores for the full sample. The mean raw score was 19.77 (SD = 8.54), which is equivalent to a 7 scaled score for a child in the 6;0–6;5 age range, or 1 SD below the norming sample mean of 10. We also looked at the mean overall raw scores of those who spoke only English at home (M = 23.16, SD = 7.84) and those who spoke a language other than or in addition to English at home (M = 16.86, SD = 8.37) and noted that children who heard a non-English language at home scored lower overall. For all children, the raw sum score for the odd items was 10.72 (SD = 4.89) and the raw sum score for the even items was 9.04 (SD = 4.13). Third, we examined item-level scores to ensure that they fell within the expected ranges and reflected the expected item-level difficulty. We visually inspected the mean score of each item by group as a first look at whether children appeared to score systemically lower on items when their assessor was masked. As shown in Figures 2 and 3, children scored close to ceiling on the first item in each block (Items 1 and 2 had a maximum score of 2). On subsequent items, which each had a maximum score of 3, children scored higher on the earlier items and lower on the later items, as expected. Our visual inspection of differences in item-level scores by group did not reveal a pattern of lower scores on items that were administered in the masked condition compared to the unmasked condition.

Next, we fit a series of multiple regression models predicting the sum score for each block as a function of group, controlling for children's sentence comprehension, gender, and home language. Table 2 displays the estimates for the block of odd-numbered items. In Model 1, the coefficient on Group A (who saw the odd items masked) was negative, nonsignificant, and small, b = −0.16, $t(94)$ =

Figure 3. Performance on even-numbered items by group. The highest score was 2 for Item 2 and was 3 for all other items.

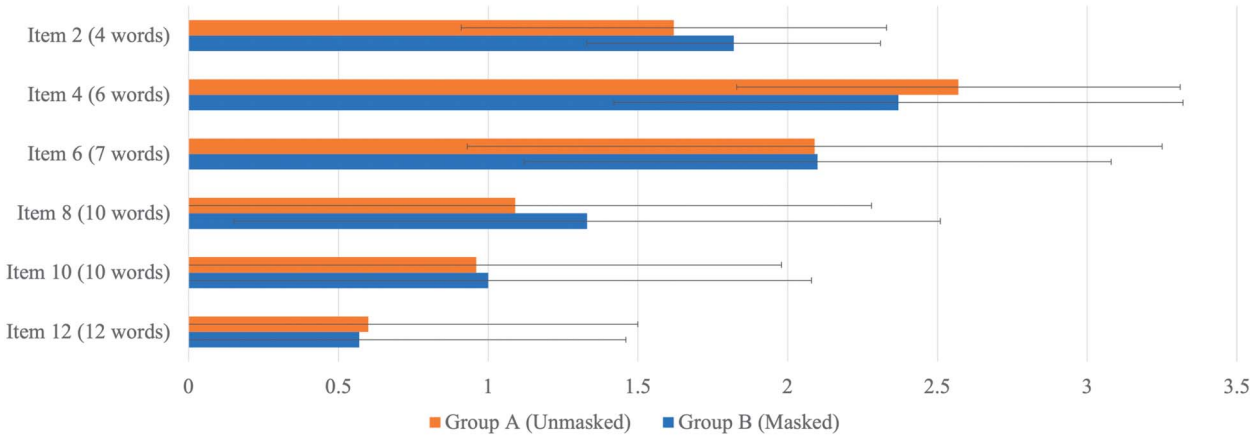


Table 2. Multiple regression predicting odd-item sum score as a function of Group A (masked).

Variable	Model 1		Model 2		Model 3		Model 4	
	Est.	(*SE*)	Est.	(*SE*)	Est.	(*SE*)	Est.	(*SE*)
Group A (masked)	−0.16	(1.00)	0.59	(0.78)	0.52	(0.77)	0.58	(0.80)
CELF-5 Sentence Comp.			0.49***	(0.06)	0.48***	(0.06)	0.43***	(0.07)
Female					−1.54*	(0.77)	−1.63*	(0.80)
Home language not English							−2.14*	(0.85)
Intercept	10.80***	(0.70)	10.44***	(0.54)	12.67***	(1.24)	13.95***	(1.34)
R^2	< .01		.42		.45		.49	
F	0.02		33.72		24.52		19.54	

Note. CELF-5 Sentence Comp. is centered at the sample mean. In this way, the intercept is interpretable as the predicted raw score of a male child in Group B with English only at home and an average score on the CELF-5 Sentence Comp. Est. = estimate; *SE* = standard error; CELF-5 Sentence Comp. = Clinical Evaluation of Language Fundamentals–Fifth Edition, Sentence Comprehension.

*$p < .05$. ***$p < .001$.

−0.16, $p = .88$. When sentence comprehension was added in Model 2, the coefficient on Group A became positive but remained nonsignificant and small. Adding subsequent controls in Models 3 and 4 did not change the sign, significance, or strength of the coefficient for group. The estimates in our final model (Model 4) suggest that children who scored higher on sentence comprehension also tended to score higher on Block 1 of recalling sentences, and while children who were female or who spoke a non-English language at home tended to score slightly lower, there was no significant effect of assessor masking on child performance. Masking explained 0.3% of the total variance (based on the change in R^2 when group is removed from the final model), and the predicted difference in raw scores was about half a point higher for children in the masked condition, controlling for all else, $b = 0.58$, $t(82) = 0.73$, $p = .47$.

Table 3 displays the estimates for the even-numbered items. In Model 1, the coefficient on Group B (who saw the even items masked) was positive, nonsignificant, and small, $b = 0.27$, $t(94) = 0.32$, $p = .75$. When sentence comprehension was added in Model 2, the coefficient on Group B became negative but remained nonsignificant and small in Models 2–4. In our final model (Model 4), the only significant predictor of even-item sum score is sentence comprehension, with children who scored higher on this measure predicted to score higher on the even items. As with the odd items, there was no significant effect of assessor masking on child performance on the even items. In our final model, masking again explained 0.3% of the total variance. The predicted difference in raw scores was less than half a point lower for children in the masked condition, controlling for all else, $b = -0.44$, $t(82) = -0.63$, $p = .53$.

To answer our second research question about whether the effect of masking varied by home language background, we added a Group × Home Language interaction term to the final models for the odd and even items (see Table 4). There were no significant interactions, $b = 2.14$, $t(81) = 1.33$, $p = .19$, for odd items, and $b = -0.86$, $t(81) = -0.60$, $p = .55$, for even items, suggesting that assessor masking did not have a greater impact on the children whose families spoke a language other than

Table 3. Multiple regression predicting even-item sum score as a function of Group B (masked).

Variable	Model 1		Model 2		Model 3		Model 4	
	Est.	(*SE*)	Est.	(*SE*)	Est.	(*SE*)	Est.	(*SE*)
Group B (masked)	0.29	(0.85)	−0.33	(0.66)	−0.33	(0.66)	−0.44	(0.70)
CELF-5 Sentence Comp.			0.41***	(0.05)	0.41***	(0.05)	0.42***	(0.06)
Female					−0.10	(0.66)	−0.30	(0.70)
Home language not English							−0.02	(0.74)
Intercept	8.89***	(0.61)	9.21***	(0.47)	9.36***	(1.05)	9.83***	(1.14)
R^2	< .01		.43		.43		.44	
F	0.12		34.29		22.63		15.92	

Note. CELF-5 Sentence Comp. is centered at the sample mean. In this way, the intercept is interpretable as the predicted raw score of a male child in Group A with English only at home and an average score on the CELF-5 Sentence Comp. Est. = estimate; *SE* = standard error; CELF-5 Sentence Comp. = Clinical Evaluation of Language Fundamentals–Fifth Edition, Sentence Comprehension.

***$p < .001$.

Table 4. Testing interactions between group and home language (HL).

Variable	Odd items		Odd items		Even items		Even items	
	Est.	(*SE*)	Est.	(*SE*)	Est.	(*SE*)	Est	(*SE*)
Group (masked)	0.58	(0.80)	–0.42	(1.10)	–0.44	(0.70)	–0.04	(0.97)
CELF-5 Sentence Comp.	0.43***	(0.07)	0.44***	(0.07)	0.42***	(0.06)	0.43***	(0.06)
Female	–1.63*	(0.80)	–1.45	(0.81)	–0.30	(0.70)	–0.23	(0.71)
Home language not English	–2.14*	(0.85)	–3.12**	(1.12)	–0.02	(0.74)	0.44	(1.07)
Group × HL			2.14	(1.61)			–0.86	(1.42)
Intercept	13.95***	(1.34)	14.17***	(1.35)	9.827***	(1.14)	9.512***	(1.26)
R^2	.49		.50		.44		.44	
F	19.54		16.13		15.92		12.71	

Note. Group is coded as A = 1 for models predicting odd items and B = 1 for the models predicting even items, so that for both sets of models, the estimates on group are for the masked condition and the reference category is the unmasked condition. Est. = estimate; *SE* = standard error; CELF-5 Sentence Comp. = Clinical Evaluation of Language Fundamentals–Fifth Edition, Sentence Comprehension.

*$p < .05$. **$p < .01$. ***$p < .001$.

English at home. As a sensitivity check, we also tested for significant interactions between group and sentence comprehension score (as both a continuous variable and a dichotomous variable) to determine if children with lower language skills were disproportionately affected by assessor masking. No significant interactions were found (see Supplemental Materials S1 and S2). We also reran all analyses controlling for treatment condition in the two larger studies, but the results remained substantively unchanged (see Supplemental Materials S3–S5).

Discussion

The purpose of the current experiment was to test whether assessor masking impacted child performance on a widely used oral language measure. In our sample of 96 kindergartners from diverse ability and linguistic backgrounds, we did not find evidence for a clinically meaningful, negative impact of masking on sentence recall. Moreover, while children with a home language other than or in addition to English scored slightly lower on average on this English-language measure, this gap was not exacerbated by assessor masking. These findings were contrary to our expectations based on past studies on the effect of masking with adults. Below, we discuss potential explanations for these findings, limitations to consider, and implications for clinical practice and research.

One interpretation of our results is that masking did not adversely affect performance on this task in meaningful ways because kindergarten-age children are not yet proficient at integrating visual information from speakers' mouths and thus do not benefit from the unmasked condition compared to the masked condition. This is consistent with research on the slow development of audiovisual integration in children (e.g., Ross et al., 2011), as well as emerging evidence that masking affects child comprehension by degrading the audio signal, but not by obscuring visual information (Lalonde et al., 2022; Schwarz et al., 2022). It is also possible that the changes to the audio signal due to masking were not large enough to compromise child performance. Our assessors used surgical or N95-style masks, which have been found to degrade the audio signal less than thicker cloth or transparent plastic masks (Corey et al., 2020). Moreover, studies with adults consistently found larger masking effects in the presence of background noise (e.g., Toscano & Toscano, 2021). Because we conducted the experiment in a quiet room at the child's school without competing background noise, the audio signal may have been sufficiently clear for the children in both the masked and unmasked conditions. Although the assessors were instructed to deliver the items in the same conversational tone in both conditions, it is also possible that they subconsciously compensated for the mask by enunciating more (Pycha et al., 2022).

Given that children in this sample entered formal schooling during the pandemic, another possibility is that the children in our sample had developed their capacity to understand speech from masked speakers to a greater degree than participants in the extant literature. The children in our sample were 6 years old in the spring of 2022, meaning that the majority entered school as preschoolers in Fall 2020 or kindergartners in Fall 2021, at which time mask mandates were in place in the school districts where this research was conducted. Therefore, they likely had extensive experience listening to masked adults and peers in school settings, for at least 6 months. While we do not yet have empirical evidence on children's ability to adapt to adult masking over time, one study investigated whether the effect of masking on adult speech comprehension waned during the pandemic (Crinnion et al., 2022). This study did not find that the adults improved at

understanding masked speech over time, but it is possible that young children are more resilient and able to adapt to changes in their audio and visual input.

Another possibility is that we were underpowered to detect significant effects. However, even if our sample had been large enough for the differences between the groups to reach significance, the direction and magnitude of the effects do not suggest that masking had a clinically meaningful impact on young children's scores. Across all models predicting odd-item sum scores, the estimated coefficients on group ranged from −.16 to .58. For models predicting even-item sum scores, the estimated coefficients on ranged from −.44 to .29. If we take the lowest estimates obtained from each to represent the lower-bound plausible values for the effect of masking, the combined lower bound estimate is still not large enough to be clinically meaningful (−.16 + −.44 = −.60). This amounts to a difference of less than 1 point on children's raw scores and, in most cases, would not be equivalent to a full 1-point difference in the scaled scores. For example, a child aged 6;2 with a raw score of 19 would receive the same scaled score of 7 as a child aged 6;2 with a raw score of 20. Importantly, these data provide initial evidence that assessments administered for clinical and research purposes during times of widespread masking may provide valid indicators of young children's language recall skills.

Because these assessments were completed in a quiet setting at the child's school, our findings cannot be generalized to settings in which background noise and other distractions may compromise students' attention and speech comprehension. Likewise, these findings should not be used to justify decisions by policymakers or school administrators in making decisions about masking in light of future public health emergencies. It is possible that our decision to use a measure developed for slightly younger children (CELF-P2) resulted in different findings than would have occurred with other versions of the CELF. However, this is unlikely, as the raw scores were normally distributed, and we did not see evidence of ceiling effects. Another limitation is our lack of detailed information about the degree of children's exposure to and use of the languages they heard at home, which would have allowed us to describe the heterogeneity of language backgrounds in our sample more precisely. These findings do not generalize to other child language comprehension measures or literacy and social–emotional tasks that may be differentially affected by assessor masking. Future studies should replicate this experiment with other measures to see if the results are generalizable. Finally, we assessed the impact of masking at a single time point; in the future, researchers should conduct longitudinal analyses to understand if children's ability to adapt to masking increases over time.

Although masking in schools was a complex and politicized issue among educators, policymakers, and parents during the pandemic (Frenkel, 2022), these findings suggest that one subtest of a widely used language assessment can likely be interpreted as a valid, meaningful indicator of child skills if it was individually administered during COVID-related masking. These data also suggest that, when assessing young and linguistically diverse children's language recall skills, future assessors can choose to mask without significantly impacting dual language learners' assessment results. To inform broader clinical and research assessment procedures, we encourage other researchers and clinicians to consider how variations in masking affect children's performance in other language domains and other important areas of early learning.

Data Availability Statement

The de-identified data set and syntax to replicate our analysis are available on OSF at https://osf.io/g3rjk/.

Acknowledgments

This research was supported by research and training grants from the Institute of Education Sciences (R305A190065, R305A180094, and R324B200018) to Tricia Zucker, principal investigator. Opinions expressed do not represent the views of the U.S. Department of Education. The authors are grateful to the students and staff who took part in this study with particular thanks to the study coordinators, Nancy VanderLinden, Cindy Elias, and Ivet Hirlas. They also want to thank Chris Schatschneider for consulting with us on the design of the experiment.

References

American Speech-Language-Hearing Association. (2022). *Using masks for in-person service delivery during the COVID-19 pandemic: What to consider*. https://www.asha.org/practice/using-masks-for-in-person-service-delivery-during-covid-19-what-to-consider/

Birulés, J., Bosch, L., Brieke, R., Pons, F., & Lewkowicz, D. J. (2019). Inside bilingualism: Language background modulates selective attention to a talker's mouth. *Developmental Science, 22*(3), Article e12755. https://doi.org/10.1111/desc.12755.

Castilla-Earls, A., Ronderos, J., McIlraith, A., & Martinez, D. (2022). Is bilingual receptive vocabulary assessment via telepractice comparable to face-to-face? *Language, Speech, and Hearing Services in Schools, 53*(2), 454–465. https://doi.org/10.1044/2021_LSHSS-21-00054

Corey, R. M., Jones, U., & Singer, A. C. (2020). Acoustic effects of medical, cloth, and transparent face masks on speech signals. *The Journal of the Acoustical Society of America, 148*(4), 2371–2375. https://doi.org/10.1121/10.0002279

Crinnion, A. M., Toscano, J. C., & Toscano, C. M. (2022). Effects of experience on recognition of speech produced with a face mask. *Cognitive Research: Principles and Implications, 7*(1), Article 46. https://doi.org/10.1186/s41235-022-00388-4

Frenkel, S. (2022, August 1). *How some parents changed their politics in the pandemic*. The New York Times. https://www.nytimes.com/2022/08/01/technology/anti-vax-parents-political-party.html

Giovanelli, E., Valzolgher, C., Gessa, E., Todeschini, M., & Pavani, F. (2021). Unmasking the difficulty of listening to talkers with masks: Lessons from the COVID-19 pandemic. *I-Perception, 12*(2), 2041669521998393. https://doi.org/10.1177/2041669521998393

Haider, C. L., Suess, N., Hauswald, A., Park, H., & Weisz, N. (2022). Masking of the mouth area impairs reconstruction of acoustic speech features and higher-level segmentational features in the presence of a distractor speaker. *NeuroImage, 252,* 119044. https://doi.org/10.1016/j.neuroimage.2022.119044

Hillairet de Boisferon, A., Tift, A. H., Minar, N. J., & Lewkowicz, D. J. (2018). The redeployment of attention to the mouth of a talking face during the second year of life. *Journal of Experimental Child Psychology, 172,* 189–200. https://doi.org/10.1016/j.jecp.2018.03.009

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science, 218*(4577), 1138–1141. https://doi.org/10.1126/science.7146899

Lalonde, K., Buss, E., Miller, M. K., & Leibold, L. J. (2022). Face masks impact auditory and audiovisual consonant recognition in children with and without hearing loss. *Frontiers in Psychology, 13,* Article 874345. https://doi.org/10.3389/fpsyg.2022.874345

Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences of the United States of America, 109*(5), 1431–1436. https://doi.org/10.1073/pnas.1114783109

Massaro, D. W., Thompson, L. A., Barron, B., & Laren E. (1986). Developmental changes in visual and auditory contributions to speech perception. *Journal of Experimental Child Psychology, 41*(1), 93–113. https://doi.org/10.1016/0022-0965(86)90053-6

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746–748. https://doi.org/10.1038/264746a0

Morin-Lessard, E., Poulin-Dubois, D., Segalowitz, N., & Byers-Heinlein, K. (2019). Selective attention to the mouth of talking faces in monolinguals and bilinguals aged 5 months to 5 years. *Developmental Psychology, 55*(8), 1640–1655. https://doi.org/10.1037/dev0000750

Pons, F., Bosch, L., & Lewkowicz, D. J. (2015). Bilingualism modulates infants' selective attention to the mouth of a talking face. *Psychological Science, 26*(4), 490–498. https://doi.org/10.1177/0956797614568320

Pycha, A., Cohn, M., & Zellou, G. (2022). Face-masked speech intelligibility: The influence of speaking style, visual information, and background noise. *Frontiers in Communication, 7*. https://doi.org/10.3389/fcomm.2022.874215

Rose, K., Armon-Lotem, S., & Altman, C. (2022). Profiling bilingual children: Using monolingual assessment to inform diagnosis. *Language, Speech, and Hearing Services in Schools, 53*(2), 494–510. https://doi.org/10.1044/2021_LSHSS-21-00099

Ross, L. A., Molholm, S., Blanco, D., Gomez-Ramirez, M., Saint-Amour, D., & Foxe, J. J. (2011). The development of multisensory speech perception continues into the late childhood years. *European Journal of Neuroscience, 33*(12), 2329–2337. https://doi.org/10.1111/j.1460-9568.2011.07685.x

Schrank, F. A., McGrew, K. S., & Mather, N. (2014). *Woodcock Johnson IV*. Riverside.

Schwarz, J., Li, K. K., Sim, J. H., Zhang, Y., Buchanan-Worster, E., Post, B., Gibson, J. L., & McDougall, K. (2022). Semantic cues modulate children's and adults' processing of audio-visual face mask speech. *Frontiers in Psychology, 13,* 879156. https://doi.org/10.3389/fpsyg.2022.879156

Sönnichsen, R., Llorach Tó G., Hochmuth, S., Hohmann, V., & Radeloff, A. (2022). How face masks interfere with speech understanding of normal-hearing individuals: Vision makes the difference. *Otology & Neurotology, 43*(3), 282–288. https://doi.org/10.1097/mao.0000000000003458

Toscano, J. C., & Toscano, C. M. (2021). Effects of face masks on speech recognition in multi-talker babble noise. *PLOS ONE, 16*(2), Article e0246842. https://doi.org/10.1371/journal.pone.0246842

Truong, T. L., Beck, S. D., & Weber, A. (2021). The impact of face masks on the recall of spoken sentences. *The Journal of the Acoustical Society of America, 149*(1), 142–144. https://doi.org/10.1121/10.0002951

University of Oregon. (2019). *Dynamic Indicators of Basic Early Literacy Skills (DIBELS)* (8th ed.). Author.

Vergara, D., Antón-Sancho, Á., Maldonado, J.-J., & Nieto-Sobrino, M. (2022). Impact of using facemasks on literacy learning: The perception of early childhood education teachers. *European Journal of Investigation in Health, Psychology and Education, 12*(6), 639–654. https://doi.org/10.3390/ejihpe12060048

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (1999). *Comprehensive Test of Phonological Processing: CTOPP*. Pro-Ed.

Wiig, E. H., Secord, W. A., & Semel, E. (2004). *Clinical Evaluation of Language Fundamentals Preschool* (2nd ed.). Pearson.

Wiig, E. H., Semel, E., & Secord, W. A. (2013). *Clinical Evaluation of Language Fundamentals–Fifth Edition (CELF-5)*. Pearson.