|  |  |  |
| --- | --- | --- |
| **Table S2**  *Phase1- Identifying functional features.* | | |
| Learning | Example in data | Impact on app development |
| Children had difficulty attending to the  ‘disembodied’ audio in the test. | Testers referred to the audio as “the lady” as it  was a female voice. This became confusing for  participants if there was a ‘lady’ in the video also.  Participant 1 made a judgement of “that was wrong” about a video with a ‘lady’ as one of the agents in it even though there was an error, and the audio did not play. | An animated bunny character *Bella* was  designed to give instructions and speak the test items. Bella was designed to have no resemblance to characters that appeared in the animations of the test items. |
| Children made errors in responding as they  did not remember which button  represented ‘right’ and which was for  ‘wrong’. | Participant 5 was observed to say “right” verbally  as her answer to the item but then subsequently pressed the button for wrong. This occurred a number of times and required the tester to assist her to go back and change her answer. | A training game was designed for the start  of the test, with a focus on teaching children the function of the buttons. The training game builds from asking children to simply “touch here” to asking them to make their own judgements on the answer and press the corresponding button without prompting. |
| Children tended to judge the morality of  what happened in the video rather than  make a truth-value judgement about the  spoken sentence. | For some of the assessments, testers used the  phrasing “Did the lady get it *right* or *wrong*?” to prompt children. This resulted in more morality judgements than phrasing like “Did the lady get it *right* or *mixed-up*?” Participant 2 made morality judgements on several of the items, justifying his answers by saying things like “Because he shouldn’t be climbing trees. He’s little only”. | For instructions and prompts in the final app  the phrasing “Did I get it *right* or did I *make a mistake*?” was used. |
| Children required frequent praise and  redirection to stay on task, particularly  as the assessment progressed. | For most participants, a pattern of praise and  redirection provided by testers was observed as follows: praise or redirection was more often needed on transitions periods in assessment (before and after stars) and more frequently towards end of assessment. Off-task behaviour increased towards end of assessment suggesting more praise or redirection is needed at this stage. For example, participant 7’s off-task behaviours initially occurred roughly 3 to 5 minutes apart in the first two thirds of the test. In the final part of the test off-task behaviours occurred on average 1 minute apart. | Audio praise was incorporated after every  practice test item. Reward/star animations were incorporated more frequently. The star animations included templates/shadows to indicate all the stars that were to be collected to allow children to see how close they were to the end of the test. |
| Swiping was a difficult gesture for children  to master and resulted in frequent errors. | Participant 7 found swiping particularly difficult  and often missed buttons by swiping imprecisely which increased his reliance on the tester for support. | Final version of app was designed so  swiping was not required. The only touch gesture required was single touch of buttons which was intuitive to children. |
| Symbols for right and wrong (a smiley face  and a sad face) caused children to select  the response that reflected any emotions  in the animation rather than truth-value  judgement. | Children were observed to match emotions  presented in the animations to the corresponding response button. | Symbols were changed to a tick and an X (to  indicate when the character got it right or made a mistake respectively). |
| Children generally needed a lot of support  during practice items and practice items  were not always successful in teaching  the truth-value judgement concept of the  test. | The ‘catch’ items, that is, the items with simple  sentences designed to control for a yes bias in responding, in some cases did not ‘catch’ children and they responded incorrectly. | A training game was designed to teach the  concept of truth-value judgement to children. First children would be asked to make judgements on common nouns e.g. “This is a ball” and then simple sentences e.g. “The boy is eating” before moving onto complex practice sentences. |
| Requiring children to press a play button to  progress to the next item delayed testing  and required frequent tester prompting. | The most frequent reason for testers touching the  screen on the child’s behalf was to play the next item and keep the pace of the test going. | The app was designed so that once the child  gave a response it would auto-progress to the next item. |