**Independent Review Report, Reviewer 1**

I would like to congratulate the authors for this very important and much needed study. I have great respect that the authors have completed this complex study in the light of the pandemic. The study is very important and contributes to our understanding of SLIs' aptitude. It will serve as the basis for other researchers to build their studies on the current study.

This study has a very strong theoretical background and sound methodology for this complex topic. The chosen methods are appropriate, even though I have a remark regarding the correlations below. The results are presented in a comprehensive way (see also one comment below re the figures) as is the data interpretations.

Thank you for these positive remarks, which we are glad to read!

Here a few limitations and suggestions for revisions:

1) The authors argue quite late in the review of the literature why they did not focus equally on personality traits as much as they did for cognitive and linguistic variables (even though they included one measure). I would recommend mentioning somewhere earlier in the review of the literature why the main focus lies on cognitive/linguistic variables and not on personality factors.

This is an important point and should have been clarified earlier in the text. We have moved the following sentence, explaining the focus on cognitive and linguistic factors over personality factors, forward a few sections to the second paragraph of the Introduction: "*Here we follow López Gómez et al. (2007), who found that perceptual-motor skills and cognitive verbal abilities played a greater role than personality in predicting SLI students’ sign language proficiency, and suggest that greater focus should be placed on cognitive predictors of signing and interpreting outcomes, as does Stone (2017).*"

2) Figures: It would be helpful to include the information about the sample size(s) to the different figures.

This point was also raised by Reviewer 2, and we have added in more details to the figure legends, including sample sizes, statistical significance and more explanation of what the figures indicate to aid in understanding the data visualizations.

3) Correlations: We all know that correlations are not the strongest statistical measure to investigate the complexity of the different variables associated with SLIs’ aptitude that were collected in the present study. The explanation of the authors that other statistical models cannot be applied because of the number of missings (further complicated by the pandemic) is comprehensible. I suggest that the authors include the effect size of the correlations (Cohen) in the paper. Another issue is that the authors sometimes write that there is an association between variable A and B, but that it is not statistically significant. Isn’t it more common to report only if a correlation is statistically significant or not?

We appreciate Reviewer 1’s note of caution around the reporting of the correlations and the suggestion to include effect sizes. It should be noted that the coefficient R2, which we have used to report the correlations in the final section of the Results (‘Predictors of BSL and SLI performance’), is already a standardized effect size. Cohen's *d* is generally used for measuring the strength of a difference between the means of two groups, not for calculating the size of the effect of one variable on another variable. However, we have added Cohen's *d* effect sizes for the analyses in the second section of the results ‘Change in cognitive and linguistic skills during SLI program’, where performance at the initial pre-degree session is compared to performance at each subsequent session in turn.

Furthermore, we agree that it is important to avoid inadvertently implying that relationships between variables are meaningful due to moderately large R2 values which are nonetheless statistically insignificant (likely in part due to the small sample sizes). We have therefore removed mentions of ‘associations’ between variables with moderate R2 values but where p>.05, and for extra clarity have stated that these are statistically insignificant.

**Check List**

**a. Is the quality of the figures and tables satisfactory?**Yes

**b. Does the reference list cover the relevant literature adequately and in an unbiased manner?**Yes

**c. Are the statistical methods valid and correctly applied? (e.g. sample size, choice of test)**Yes

**d. Is a statistician required to evaluate this study?**No

**e. Are the methods sufficiently documented to allow replication studies?**Yes

**QUALITY ASSESSMENT:**

**Rigor**5

**Quality of the writing**5

**Overall quality of the content**5

**Interest to a general audience**4

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Independent Review Report, Reviewer 2**

*Requests for revisions:*

*1. Aptitude testing will be very relevant to interpreter trainers and it would be useful if this article were written with them in mind as a potential audience. It doesn't need to be less rigorous, but it should include more post processing of the specific information and guidance in reading the figures. The text is dense, even for someone familiar with such articles, so it would be of general benefit. Especially because the unfortunate impact of COVID and its added effect on already low retention has made this more of a pilot/preliminary study, tapping into the audience of interpreter trainers will yield future subjects and hopefully potential expanded data from additional programs. Given the already existing problems with retention of subjects over longitudinal studies, this would be recommended. This would also highly motivate publishing this as a preliminary longitudinal study with some missing data points, which I hope would be an option because of its relevance.*

**We agree that it is vital that this work reaches the audience of interpreter trainers and thank Reviewer 2 for these suggestions to make the manuscript more accessible. In order to make the data visualizations clearer, we have expanded the legends with more details on how to read the figures and interpretation of the results, as suggested by both reviewers. We have also rewritten much of the results section to make clearer which results are significant and of particular note, as well as redrafting the first paragraph of the Discussion, which is now a much more specific summary of the key findings of our work. In terms of organization, we have added additional sub-headings in the Discussion ‘Summary of key findings’ and ‘Issues with data collection and attrition’ in order to guide the reader through these sections more clearly. We hope that having made these changes, the take-home messages are both clearer and easier to find, and that the results have been post-processed for the reader better throughout the Results and Discussion sections. Thank you for the positive comments about the relevance of our study, even with missing data points.**

*2. In the legends to Figures, it would be useful to not only indicate year and task but what they indicate and significance. I am looking for ways to give the text more of a narrative and flow in terms of what the findings mean and to make the visuals more understandable (more stand alone) when one looks back at them. Currently it is very dense reading.*

**Thank you for this suggestion, which relates to a similar comment by Reviewer 1. It is indeed important that our readers understand the figures and that they can stand alone. We have added details regarding sample size and statistical significance to the figure legends, and attempted to explain what the figures indicate more clearly in the legends.**

*3. In a similar vein, in the Discussion section, the summary of findings needs to be less vague. There is an assumption that all of this hangs together in the Results section, but the reader gets lost in the details. So it needs to be re-organized with deliberate efforts to synthesize and lucidly present the data. As an example of where to punch up the specifics of the findings for readers, look at the vagueness of findings in the first paragraph of the Discussion section.*

**Thank you for pointing out that the summary of findings is too vague at present. The lack of specificity in the Discussion was mostly due to an overabundance of caution given the small sample sizes and lower statistical power in our preliminary study. We have completely redrafted the opening paragraph of the Discussion in order to make the specific take-away findings from our three analyses clear. As previously stated we have also tried to reorganize the Discussion section better via additional sub-headings and have added in more detail throughout to make sure that the findings are less vague.**

Strengths:

1. Prior aptitude studies have started the testing at entry into interpreter training, but the authors here have recognized the benefit of looking earlier to entry or early progress in learning the signed language. In terms of response to aptitude rather than using this information as an exclusion factor, which the authors do not advocate at this early stage of aptitude assessment, it can be used for early identification of SL student who should be encouraged to consider pursuing interpreting as well as motivation for directing other students toward SL careers that don't involve interpreting. The latter are often not emphasized early enough in students' careers and they gravitate toward interpreting without considering the alternatives.

We agree and we have emphasized this point again in the ‘Implications and future directions’ section: "*Additionally, SLI aptitude testing could also help highlight other related careers that do not involve SLI, which may be more appropriate for some students.*"

2. The inclusion of mental rotation tasks in aptitude testing is also a strength. It is interesting that 3-D rotation proved relevant to third-year as opposed to second-year BSL grades. One factor that wasn't addressed was the correlation between 3D rotation skills and grammatical constructions covered in various years of the BSL curriculum. Also, 2D rotation skills seem more relevant to admission to a program than ongoing. Many second-language learners rely (incorrectly) on 2D rotation in roleshifting, when Deaf signers use a shifting of perspective (less in line with the Shepard, et al. studies and more in line with the PIazza del Duomo studies of Bisiach). In future, consider pairing rotation with perspective tasks.

Thank you for highlighting this point. We have added a sentence about the correlation between 3D-MR and the BSL-SRT, which tests progressively more difficult grammatical constructions: "*Both mental rotation and visuospatial WM are likely implicated in this task where increasingly complex grammatical constructions in BSL must be reproduced.*" The point about perspective shifting vs. rotation is very interesting. Is there a study we could cite that suggests L2 signers rely on 2D rotation during role-shifting whereas Deaf signers use perspective-shift? In the future it would certainly be interesting to include perspective-taking tasks.

3. The independent measure of consecutive interpreting done in session 4 were important to the student assessment. It would be good to confirm that the SLI instructors were not instructors from the programs studied, give the length of the texts, as well as characteristics of the texts would be worth noting in a footnote. The focus on chunking was very appropriate. Kudos for not getting waylaid by looking at simultaneous interpreting, which would detract from seeing processing.

Thank you for the positive comments about the consecutive interpreting tasks and the focus on chunking. We have added more details on the length of the texts and the formality of the register in the Methods section. In terms of the marking, we agreed it would be best to have a consistent single marker for the tasks, who was a lecturer at the University of Wolverhampton. Only 3 of the 9 participants who completed these final-session interpreting tasks were Wolverhampton students. Students on this degree have multiple instructors and contact with this one specific lecturer was minimal, so we do not think this has had any meaningful impact on the grading for this task.

4. COVID and the shift to on-line classes as well as the expansion of virtual interpreting makes this article very timely. Interpreting programs have been hit harder than most academic programs and the increasing scarcity of professional interpreters is a major public concern. The shift in these studies from face-to-face testing and the necessary shift to on-line testing for the research is also worthy of note for future studies as a pro-active consideration. We were all forced to re-work testing to a virtual setting, but the reworking can inform systematic changes in future studies. This also allows for access to a wider population of subjects.

We agree with Reviewer 2 about the benefits of online testing and plan to make use of this in future larger-scale studies of interpreter students, even if lab-based testing has now largely resumed. Furthermore, we have added some extra detail in the Implications section about what worked particularly well for us online; the impact of the pandemic on SLI education and the hope that our preliminary study shows online aptitude testing is feasible: “*The tasks hosted on Pavlovia and GoReact were particularly successful. Future online aptitude testing could facilitate access to a larger population of SLI students, however, factors such as equity of access to technology and space to participate online should be considered carefully. Given the extra adaptations to SLI and sign language teaching programs which were required due to the pandemic (see e.g. Katz, 2021; Hornstra 2021) and the increasing use of online L2 sign language teaching methods even pre-pandemic (Ackerman et al., 2018), we hope the present preliminary study demonstrates that online SLI aptitude testing is also feasible*”.

5. Interest of the research to a general audience is very high, but accessibility is limited by the exegesis.

See our previous answers regarding accessibility for an interpreter educator audience and better elaboration of what the findings mean throughout the Discussion and in the figure legends.

6. Having the link to the individual dataset at OSFHOME is a real plus. The link worked and the ability to actually download and sort data was great. A README with the tests, where to find them, and norms would be a nice addition even though much is recoverable from the article. Also, a summary of the meaning of BSL scores would be useful to those not familiar with the BSL system. Are classroom grades correlated with SQCF levels?

We are glad to read that Reviewer 2 has benefitted from the data being openly shared on OSF. The suggestion of a README is an excellent idea, and we have now added this to the OSF repository. Regarding the meaning of the BSL scores, we have added in the Methodology section that the module grades for interpreting and BSL modules are simply marked according to the UK university grading system, i.e. this is not a system specific to BSL: "*Grades follow the standard UK university grading system (0-100, whereby most marks fall between the range 40-80, see e.g Yorke et al., 2000).*" It is more difficult to (briefly) add details of the criteria for the BSL and interpreting modules themselves, since the scores used in our study are averaged across multiple module grades within each academic year, and the modules/content differs at each of the two institutions. However, we have already stated in the text which modules these marks come from, and further information about the modules is now provided in a footnote where the relevant university web pages are linked. We are not sure what is meant by “SQCF levels”, could you clarify please?

Weaknesses:

1. This is a very promising and comprehensive study. While not their fault, the impacts of COVID in conjunction with an already strong tendency for attrition in interpreter training compromised the significance and reliability of the findings. Without the impact of COVID the large initial cohorts tested pre-selection might have yielded the numbers needed. Hopefully, the authors have continued to collect data from new cohorts for further analysis.

Thank you for the positive words about our work and for acknowledging the difficulties caused by the pandemic’s effects on participant attrition. We certainly plan to test new cohorts of interpreting students as part of a larger new research study, grant funding permitting, but this is obviously beyond the scope of the present preliminary study. Ideally we will also have funds to re-test our continuing students in this study again, hopefully at the end of their degrees and at the beginning of their working careers.

2. As found in Macnamara et al, and Stone, personality tests like the BIS can be strong predictors of success in interpreting programs, but this effect is most likely to be found pre-entry rather than along the way.

The lack of a baseline measure of impulsiveness/risk-taking was an unfortunate result of not having enough time to test all measures at the first testing session without causing testing fatigue. We have now acknowledged this more explicitly in the Discussion: “*Due to testing session time constraints, we only introduced the BIS from session 3 onwards, where all testing was already online due to the pandemic, but ideally we would have also taken a baseline measure of risk-taking.*” However, we note that Stone (2017) only found a group difference in Barratt Impulsiveness Scale scores between those who continued with interpreting and those who switched to a deaf studies course in semester 3, and not at the first testing session early in the course. In the Introduction, we have also clarified that the test used by Macnamara et al. (2011) is actually a different measure of risk-taking (yet confusingly with the same BIS acronym; "Behavioral Inhibition System").

3. Grades in BSL and interpreting classes, especially during COVID may not be the best dependent measure of success. To argue for this, the criteria for grading and the interpreting tasks used should be presented. One of the best dependent measures of success would be 1) entry into the interpreting field and staying there for 2+ years and 2) certification. These obviously cannot be included in the study at hand, but the authors should be encouraged to track these subjects beyond the study for a potential future report.

We have added details of the marking criteria for the interpreting tasks: "*The total score for each consecutive interpreting task was calculated based on poise, style, consecutive management, comprehension, conceptual rendition, vocabulary, accuracy, repairs and an overall mark.*" Furthermore, we have added a footnote with links to the descriptions of the modules for which grades were included, on the websites of both universities in the present study. Thank you for the suggestion to track students further. As previously stated, although the remaining students are few in number due to the high levels of drop-out, we hope to re-test them again at the end of their studies / at the beginning of their working careers as part of a larger research grant.

Re: **Q2**: It is difficult to respond with a yes or no to these questions

**a. Is the quality of the figures and tables satisfactory?** The actual visuals of the figures and tables are satisfactory. However, the Legends to the figures need to be a bit more detailed.

As noted above, we have now made efforts to add more detail to the figure legends, both in terms of statistical details such as sample size and significance, as well as interpretation, so that the meanings of the visualizations are more immediately apparent to the reader.

**b. Does the reference list cover the relevant literature adequately and in an unbiased manner?**

The literature is covered and discussed extremely well and in an unbiased manner. The authors were very thorough. I cannot bring to mind any reference that was not included. There is one reference in the text that does not appear in the references: Macnamara and Conway 2016.

Thank you for highlighting this omission. We have added Macnamara and Conway (2016) to the References.

**c. Are the statistical methods valid and correctly applied?** The choice of methods and planned application are well justified. The only problem follows from the high attrition rate and the reduction from an n=33 to n=9 over the course of the study. This is not a reason to not publish the results provided that they are cast as pilot or preliminary. It is well worth getting them out there. The BIS test was important for replication purposes, but its true value would have been with application from the outset of the study.

We have additionally re-emphasised in a number of places throughout the manuscript that the present study is a preliminary study. See our earlier response above regarding the timing of BIS testing.

**d. Is a statistician required to evaluate this study?** Yes. There clearly was effective use of a statistician in development of the study and discussion of findings, but I would want to rely upon a statistician in review of the findings as presented. A statistician may also be able to make suggestions regarding organization and presentation of the data.

While we agree that it would be beneficial for a statistician to look over the findings as presented, we have not been able to find someone qualified and available within the past few weeks to do this. Hopefully this is something that the editors can assist us with.

**e. Are the methods sufficiently documented to allow replication studies?** Yes, definitely. In fact, one is inspired to replicate or at least experiment with application of these tests in the context of interpreter training. One suggestion is to look for a reliable way to direct readers to the tests. Links have a tendency to expire. For example, when clicking on the link mentioned here " The task is freely available at: [https://pavlovia.org/freyawatkins/block\_rotation.](http://links.email.frontiersin.org/ls/click?upn=AAaFa03elZRFPXQ6ShiKwAHG33AjCKNhL-2B7Z8ftXOpnQAj9Zlt5RFB0hLfa1gGW1Yu5mKTjkve6ssOyZ8cuN7A-3D-3DDrOh_K7xONFJIs6Z2YdKYxHigZN55yRHi5tbvXa2puvP5LuWFFi5BRujTDHAN75b6Jf4MsUzJc8H7XQLdamkPva1WeemtqTc9LnfDdEmB7wc-2BBE5FZrHhDwUQEkIKx5A7pGJDODedmqoBoeI8qiDzvAQvad5GDUjNdyCE0jnrC5BIMJc6Ks-2BvcRfWoX-2FrL-2FfeTWyAVoKsYmDaWcRqNM-2B0FE1J1LN5sHpd7fPSpwPAZdmBojV6iTFh4heAfjYOIYzER7zdOpSXNVIDEp0Yx27A17PG9UwYN7iydQeWo70PrgTcEfAM3bLw9Ql6Yl9X4tYT9yHri5e7ph307IHdOWTtFf7q-2BA-2FBenerKR-2BDbI-2FKlYm-2FYWGAPHnY0L-2FOG6NBV3HtClRxpMMv-2FJb9v4ZU2z3GhD2AEHnVgsuR4NAvq066-2F6m70gk-3D) The link does not work and yields the following: "No information is available for this experiment. Either it does not exist or you do not have access to it." The article does also reference Ganis and Kievit (2015) where the materials can be found.

Thank you for making us aware that the link was not working (the Pavlovia account was accidentally still logged in). The task is now public and we have double-checked the URL is working. As Reviewer 2 states, access to all the other tasks is already described in the methodology and the links should be stable (many of the tasks are from PEBL, or from cited publications/assessment manuals, e.g. BIS, KBIT-2). Nevertheless, we have followed Reviewer 2’s earlier good suggestion of adding a README file to the OSF repository describing the location of these tasks, as well as explaining the data file and analysis scripts). We are pleased to read that Reviewer 2 is feeling inspired to attempt to replicate some of the assessments with interpreting students.

**Q4: Quality of the writing** and **Q5: Interest to a General Audience** have an interaction. While the writing is sound, there is a need to take into consideration accessibility to a wider audience.

See our previous answers regarding making the text more accessible to interpreter educators.

**Check List**

**a. Is the quality of the figures and tables satisfactory?**No

**b. Does the reference list cover the relevant literature adequately and in an unbiased manner?**Yes

**c. Are the statistical methods valid and correctly applied? (e.g. sample size, choice of test)**Yes

**d. Is a statistician required to evaluate this study?**Yes

**e. Are the methods sufficiently documented to allow replication studies?**Yes

**QUALITY ASSESSMENT:**

**Rigor**4

**Quality of the writing**3

**Overall quality of the content**5

**Interest to a general audience**5