Article:
Pan, Junlan and Marsden, Emma orcid.org/0000-0003-4086-5765 (2024) Developing internet-based Tests of Aptitude for Language Learning (TALL):An open research endeavour. Language Testing. ISSN: 0265-5322

https://doi.org/10.1177/02655322241241849

Preprint of: Pan, J., & Marsden, E. (2024). Developing internet-based *Tests of Aptitude for Language Learning* (TALL): An open research endeavour. Submitted to *Language Testing*

Accepted and published version citable as:

Pan, J., & Marsden, E. (2024). Developing internet-based *Tests of Aptitude for Language Learning* (TALL): An open research endeavour. *Language Testing*, *0*(0). https://doi.org/10.1177/02655322241241849

Developing internet-based *Tests of Aptitude for Language Learning* (TALL): An open research endeavour

Junlan Pan, Chongqing University, China

University of York, UK

Emma Marsden, University of York, UK

Developing internet-based *Tests of Aptitude for Language Learning* (TALL): An open research endeavour

Abstract

Tests of Aptitude for Language Learning (TALL) is an openly accessible internet-based battery to measure the multifaceted construct of foreign language aptitude, using language domain-specific instruments and L1-sensitive instructions and stimuli. This brief report introduces the components of this theory-informed battery and methodological considerations for developing it into an open research instrument. It also presents the preliminary results from the initial validation of TALL carried out on data collected from Chinese L1 participants (n = 165) from a university setting who took two rounds of tests (with counterbalanced test items) with a minimum 30-day interval. The results of data analyses at subtest, item, and battery levels suggest that, in general, TALL has satisfactory reliability and can be used to measure aptitude conceptualized in the theoretical frameworks on which it has been developed. This report also highlights the value of TALL as a convenient data collection tool openly accessible to any researcher for free, its potential for facilitating an open data pool for high-quality syntheses of aptitude-related research findings, as well as its implications for Open Research practices in testing language-related constructs.

开发基于互联网的语言学能测试（TALL）：一项开放研究工作

摘要

语言学能测试(TALL)是基于互联网并且可供公开访问的学能测试。它是语言特定领域的测试工具，使用基于被试母语的实验指令和题目，可测量由多个部分组成的外语学能构念。本报告简要介绍了 TALL 学能测试的组成部分，以及将其开发成为一个开放研究工具的方法考虑。TALL 的初步验证由母语为中文的大学生（n =165）完成，他们参加了间隔至少为 30 天的两轮测试（使用平衡设计的测试题目）。分项测试、测试题目和整体测试三个层面的数据分析结果表明：作为基于外语学能理论框架的学能测试，TALL 总体上信度良好，可用于测量理论框架中包含的学能构念组成部分。本报告特别强调了 TALL 的工具价值：它方便数据收集，并且可供研究人员免费使用；它有望发展成为一个开放数据资源，用于高质量地整合学能相关研究成果；此外，它也为在语言相关构念测试中实践开放研究提供了启示。

Keywords:

aptitude battery, internet-based test, open research resource, open instrument for data collection, open data pool

1 Introduction

Foreign language aptitude (also referred to as language aptitude or simply aptitude throughout this article) is defined as the special ability to learn an additional language beyond one's first language (L1) efficiently (Carroll, 1981). It has been conceptualised as a componential construct of cognitive individual differences that predict and explain language learning outcomes. Given the significance of aptitude and the need to measure it in order to investigate its role in learning, it is crucial to ascertain the reliability and validity of any aptitude measurements prior to conducting research. However, this step has been surprisingly neglected to date (cf. Bokander & Bylund, 2020). This research gap may primarily be due to limited access to most of the aptitude batteries, which prevents other researchers from scrutinizing their reliability and validity. Ideally, rigorous language test validation should be conducted by researchers who can maintain a higher level of objectivity and scepticism that the test developers themselves may not always provide (Isbell & Kim, 2023)—and this requires that batteries be openly available for (future) independent scrutiny. A notable exception is the widely used LLAMA tests (Meara & Rogers, 2019), which are openly available online. However, item-level data elicited by this battery are not readily available to the researcher (without contacting the test developers) and the data are not openly available to other researchers. The lack of open and easy access to item-level data from aptitude batteries constrains the scrutiny of reliability and validity, eventually reducing confidence in aptitude findings. This logical impasse has been described by Marsden and Morgan-Short (2023) as a "chicken-and-egg conundrum" (p.17): Appropriate validation procedures are highly desirable to ensure the rigour of measures before making them openly available to other researchers. However, in order to reach a consensus about the validity of the measures, it is necessary to make them openly accessible to accumulate validation evidence from a range of contexts and participants.

This brief report introduces an open research endeavour that attempts to begin to address the conundrum. It provides a summary of *Tests of Aptitude for Language Learning* (TALL, n.d.), a theory-informed aptitude battery, and discusses methodological considerations relevant to instrument design and technical development. In addition, it presents a subset of results from the initial validation of the battery (Pan, 2023). Our report highlights the value of TALL as an open research instrument in enhancing the methodological rigour of assessing language-related abilities and its potential as an open data tool for accumulating validation evidence in the long term.

2 TALL as a measurement for aptitude

TALL measures multiple facets of aptitude, drawing together different theoretical models and also informed by previous batteries. Four theoretical components are postulated to represent the aptitude

constructs, i.e., associative memory, phonetic coding ability, language analytic ability, and working memory (WM). Uniquely, TALL's design includes two components of Skehan's (2016) Stages Approach—language analytic ability and working memory (WM)— that are not both captured by other batteries. TALL was also designed to capture two specific components of WM that are not comprehensively accounted for by the Stages Approach but are foregrounded by Wen's (2016) Phonological / Executive Model (i.e., phonological short-term memory and executive control). In summary, TALL was designed to operationalize some of the components of first stage ('Input-oriented') and all of the components of the second stage ('Interlanguage development') of Skehan's Stages Approach, with additionally specified components of WM. It also includes the component of associative memory that is operationalized in most of the existing aptitude batteries. To measure these components, TALL consists of five subtests, each informed by existing batteries: Vocabulary Learning (TALL_VL), informed by LLAMA_B (Meara & Rogers, 2019), measures associative memory; Sound Discrimination (TALL_SD), informed by Part 5 of PLAB (Pimsleur, 1966), measures phonetic coding ability; Language Analysis (TALL_LA), informed by LLAMA_F (Meara & Rogers, 2019), measures language analytic ability; Serial Nonword Recall (TALL_SNWR), informed by the nonword repetition task in Suzuki (2021), measures phonological short-term memory; and Complex Span Tasks (TALL_CST), adapted from the reading span tasks in Gass et al. (2019), measures executive control capacity.

TALL_VL, TALL_SD, and TALL_LA are scored by the software based on the number of accurate responses to test items. For TALL_SNWR, productive data (the recall of nonwords) are manually scored after the completion of the test. TALL_CST's raw data include both processing data (i.e., semantic judgements of sentence plausibility) and storage data (i.e., letters recalled in sequential order), with only the storage data being scored by the software. This scoring approach aligns with the scoring practices in previous research involving complex span tasks (e.g., Gass et al., 2019). Further details on subtest design, example items, and scoring report are available in the test manuals (see the Supplemental material).

The development of TALL considered two important potential confounding factors that have not been systematically investigated or controlled in measuring aptitude. First, to address the confounding effect of the modality (aural or written) of presentation, TALL has two test suites. The aural suite includes all subtests in the aural format; the written suite consists of three subtests (TALL_VL, TALL_LA, and TALL_CST) in the written format, and two subtests (TALL_SD and TALL_SNWR) that are—necessarily—in the aural format. Second, to mitigate potential confounds related to learners' L2 knowledge (as existing aptitude and working memory tests are typically written in English, which is often the L2 of learners), TALL has been specifically designed for Chinese L1 speakers, using Chinese as the

instructional language. Also, in TALL_ SNWR, nonword stimuli for recall conform to the phonology of participants' L1 Mandarin Chinese and avoid real meaning associations (which can be challenging in Mandarin, as individual syllables may correspond to one or more meanings). In TALL_CST, sentence stimuli in the processing task are presented in Chinese and are controlled for sentence length, as well as for the sentential location and usage rate of the lexical cues that learners are likely to use to make semantic judgements. Additionally, language stimuli were developed in three subtests (TALL_VL, TALL_SD, and TALL_LA) using a miniature language (based on Lithuanian) that ensured novelty for these learners (as it is highly unlikely that they knew any Lithuanian).

3 TALL as an open research instrument

Developing TALL into an internet-based instrument for research purposes enables accessibility for other researchers. This endeavour involves technical considerations to minimise threats to test validity, following steps recommended by Newman et al. (2021) to ensure data quality in internet-based research. These considerations include: (i) Archival techniques for recording response times allow the identification of anomalous responses; (ii) Instructions and warnings displayed on screen before each subtest aim to reduce dishonesty in testing behaviours, e.g., note-taking or seeking others' help; and (iii) Assigned single-use test codes prohibit participants from reattempting the test.

With an open research infrastructure in mind, we incorporated functionality to separate access for researchers and invited test takers. Figure 1 shows that, through the Researcher Entry, researchers can generate test codes for their participants, download item-level data that are associated with the test codes they have generated, and upload the scores of TALL_SNWR that they have manually assessed. Crucially for the purposes of sustainability, all of this is possible without relying on administrative support from the owner of TALL, although all functionalities and collected data can be monitored from the backend by the researchers hosting TALL. Researchers can also navigate test manuals and try out demo tests for their own research interests or for pedagogical (e.g., research training) purposes. Meanwhile, access for participants is limited to the Test-taker Entry and requires a test code from a researcher. This controlled access ensures that participants do not have prior knowledge of the test, which is crucial for the test's validity, as access to the test is impossible without a code. The functionalities for access control and researcher-administration not only serve the quality of the data collected, but also the sustainability of TALL as a resource that can be used by the community without relying on an individual.

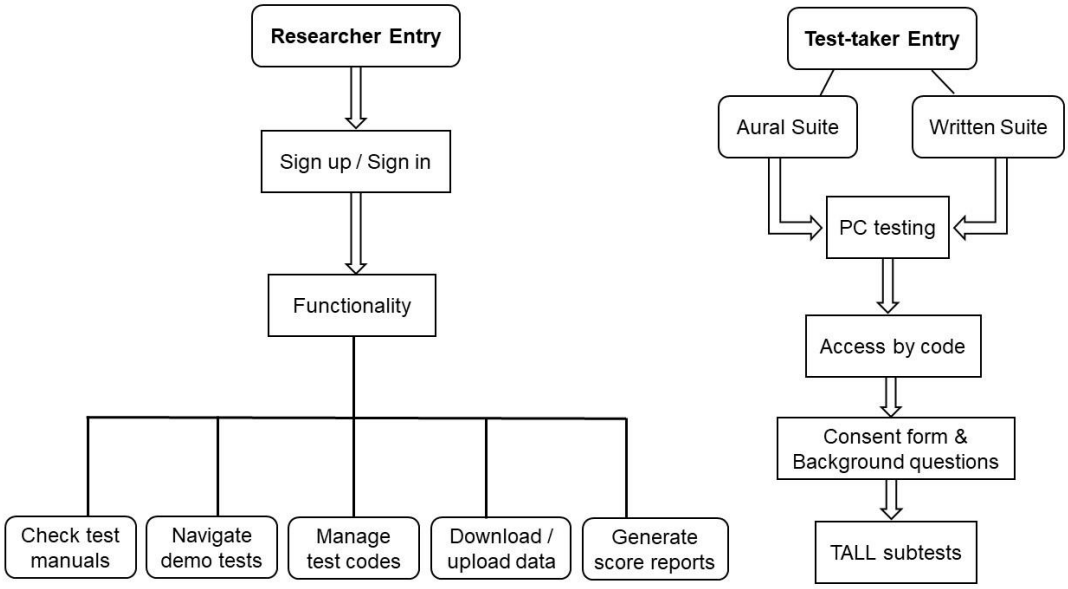

Figure 1 Separate accesses and functionality for researchers and test-takers on TALL

4. Initial validation

To assess the reliability and validity of TALL as an aptitude battery, we conducted an initial validation at subtest, item, and battery levels. This validation process was adapted from a schema used by Bokander and Bylund (2020), which draws upon the argument-based validity framework by Kane (2006) and an earlier adaptation of Kane's framework for L2 learning research by Purpura et al. (2015). In this framework, validation is carried out step by step through logical inferences. The three-level schema allowed us to make (i) generalisation inferences about each subtest's reliability in targeting the intended component, (ii) scoring inferences about the effectiveness of items in discriminating latent abilities with appropriate levels of difficulty, and (iii) explanation inferences about the alignment of the battery's structure with the theoretical frameworks underlying TALL's development.

Methods

The final data for analysis was obtained from 165 participants who were year-one undergraduates from various disciplines, across 11 universities, in China. As we were examining the effects of two modalities on measuring aptitude, we used a within-subject design (whereby each participant experienced both modalities) to attain higher statistical power compared to a between-subject design (whereby only half our participants would have experienced each modality). All 165 participants took two rounds of TALL on the test website at a time and place of their choosing, with a minimum 30-day interval between rounds. In the first session, one suite (either aural or written) of subtests were presented in the fixed order, i.e., TALL_VL, TALL_SD, TALL_LA, TALL_SNWR, and TALL_CST sequentially, while in the

second session, another suite of subtests was presented in the same fixed order. Two test suites and two versions of materials were counterbalanced over two sessions, such that an individual experienced one modality at time one and another modality at time two, with different versions in each to reduce test familiarity. The order of the test items within each subtest was randomised for every individual iteration.

Data analysis and results

Data analysis started at the *subtest* level to examine the reliability of each subtest as a measure for a specific component of aptitude. Analysis at the *item* level determined the fitness of the items in each subtest to Item Response Theory (IRT) models, addressing whether each subtest was composed of well-functioning items that had appropriate levels of difficulty for the participants and were able to discriminate the ability to be measured. Datasets split by version (A or B) and modality (aural or written) were used for analyses at subtest and item levels. Analysis at the *battery* level investigated the extent to which TALL reflected the structure consistent with the theoretical frameworks that underpinned its conceptualisation, using datasets aggregated within each suite (separately for aural or written). Data analysis was carried out in R (R Core Team, 2022) and the code is accessible on the OSF.

We used omega estimators for reliability checks, responding to concerns that the use of Cronbach's alpha often violates certain statistical assumptions, leading to an underestimation of reliability (McNeish, 2018). Omega hierarchical (ω_h) estimated the reliability of the general factor (variance shared by all items) of a test after accounting for item-group factors (variance specific to subsets of items), and omega total (ωt) estimated the reliability of the total instrument (without taking into account item-group factors). As shown in Table 1, the results showed that all subtests rendered a satisfactory estimate of the total reliability indexed by ω_t. In addition, most datasets had results indicating a large part of the score variance was likely to be attributed to a common factor with acceptable ω_h. There were just three exceptions to this (Version A of TALL_VL in the aural suite, Version A and Version B of TALL_CST in the written suite) which had lower ω_h (range .31 to .35). In general, these findings suggest strong unidimensionality in the data collected by these instruments. We also included Cronbach's alpha estimator in Table 1, to facilitate comparison to reliability coefficients reported in other studies using alpha.

Table 1 Descriptive statistics, reliability, and test information of TALL subtests

Subtest	Dataset	*n*	*k*	Descriptive statistics			Reliability					Test information	
										95% CI		ability range	
				mean	sd	skew	ω_h	ω_t	α	lower	upper	lower (–6, 0)	higher (0, 6)
TALL_VL	Version A in aural suite	81	20	5.46	3.32	1.46	.31	.86	.70	.60	.79	3.90 (25.09%)	11.07 (71.24%)
	Version B in aural suite	84	20	7.14	3.98	0.82	.56	.90	.78	.70	.84	6.74 (33.68%)	13.09 (65.46%)
	Version A in written suite	84	20	8.23	4.47	0.58	.49	.91	.81	.75	.87	7.90 (39.49%)	11.96 (59.83%)
	Version B in written suite	81	20	9.95	4.68	0.03	.58	.93	.84	.78	.88	9.79 (48.96%)	10.06 (50.31%)
TALL_SD	Version A in both suites	165	30	21.80	6.50	-0.50	.62	.96	.89	.87	.92	33.25 (79.92%)	8.29 (19.93%)
	Version B in both suites	165	30	23.61	4.09	-1.84	.72	.92	.78	.73	.83	23.02 (76.78%)	5.81 (19.39%)
TALL_LA	Version A in aural suite	81	30	19.73	9.34	-0.41	.80	.98	.95	.94	.97	46.62 (77.69%)	13.39 (22.31%)
	Version B in aural suite	84	30	20.24	7.36	-0.58	.84	.97	.91	.88	.94	32.39 (72.07%)	12.53 (27.88%)
	Version A in written suite	84	30	24.73	7.55	-1.46	.74	.99	.96	.94	.97	65.80 (82.54%)	13.92 (17.46%)
	Version B in written suite	81	30	23.88	7.43	-1.44	.82	.98	.94	.93	.96	53.48 (90.91%)	5.34 (9.08%)
TALL_SNWR	Version A in both suites	165	17	5.70	2.56	0.74	.72	.89	.86	.82	.89	14.40 (27.02%)	37.62 (70.56%)
	Version B in both suites	165	17	5.71	2.69	0.68	.76	.91	.88	.85	.90	16.66 (26.71%)	44.65 (71.61%)
TALL_CST	Version A in aural suite	81	15	12.25	1.90	-0.87	.64	.87	.84	.78	.88	30.30 (77.32%)	7.30 (18.62%)
	Version B in aural suite	84	15	12.25	2.00	-0.95	.56	.86	.82	.76	.87	32.41 (80.47%)	6.82 (16.93%)
	Version A in written suite	84	15	13.32	1.20	-1.61	.35	.68	.64	.51	.74	15.67 (74.11%)	2.86 (13.53%)
	Version B in written suite	81	15	13.00	1.48	-1.63	.33	.79	.72	.62	.81	18.30 (77.08%)	3.55 (14.94%)

Note. Key to column headings: TALL_VL= Subtest of Vocabulary Learning; TALL_SD = Subtest of Sound Discrimination; TALL_LA = Subtest of Language Analysis; TALL_SNWR = Subtest of Serial Nonword Recall; TALL_CST = Subtest of Complex Span Task; *n* = number of participants (each did two rounds of tests, e.g., version A in aural suite and version B in written suite); *k* = number of test items or trials; sd = standard deviation; skew = skewness; ω_h= omega hierarchical; ω_t = omega total; α = Cronbach's alpha; 95 % CI = 95% confidence intervals of Cronbach's alpha with lower and upper bounds

For the item level analysis**,** Rasch models were applied to dichotomous data from the subtests TALL_VL, TALL_SD, and TALL_LA. The infit *t* statistics range of [-2, 2] is a rule of thumb for detecting potential misfitting items. The infit mean squares range of [0.50, 1.50] were also applied to check the item fit. Generalized Partial Credit models were used on the polytomous data from the two working memory subtests. Specifically, the proportion of correct recalls of nonwords (in TALL_SNWR) or English letters (in TALL_CST) in each trial constituted the item-level data. No clear evidence suggested that any items were of poor quality that may threaten the validity of the instruments, and so deletion of items was not necessary.

Test information was obtained to understand the overall precision of the subtests across the ability level they were designed to measure. In the Rasch models, the information contributed by each item in the subtests was summed to quantify test information (i.e., the sum of the item information that demonstrates the contribution items make to the estimation of ability in both the lower and upper half of the ability range). In the Generalized Partial Credit models, the information provided by each response category across all trials was summed. The results (in Table 1) indicated that test information varied between the subtests, which can be attributed to factors such as the number of items, the types of scores (dichotomous and polytomous), and the reliability of the subtests. Specifically, for the participants in this study, TALL_VL (measuring associative memory) provided the least information overall, whereas TALL_LA (measuring language analytic ability) provided most information. The results also indicated that TALL_SD, TALL_LA, and TALL_CST lacked challenge for the participants overall, providing less information about participants at higher ability levels. Conversely, TALL_VL and TALL_SNWR were found to be more difficult, offering less information about participants at lower ability levels.

Correlation matrices (in Table 2) showed that scores from most subtests in both suites had low positive correlations with one another at significance levels of $p < .05$, indicating that no subtests were redundant (as they did not display extremely high correlations). In general, these relationships among subtest scores suggested that the measures probably represent different constructs.

Table 2 Correlations (in Kendall's *τ*) between TALL subtests

	The aural suite				The written suite			
	1	**2**	**3**	**4**	**1**	**2**	**3**	**4**
1. TALL_VL	—				—			
2. TALL_SD	0.21*	—			0.12*	—		
3. TALL_LA	0.22*	0.25*	—		0.10*	0.20*	—	
4. TALL_SNWR	0.20*	0.16*	0.03	—	0.20*	0.16*	0.12	—
5. TALL_CST	0.15*	0.09	0.07	0.23*	0.13*	0.16*	0.19*	0.20*

Note. * indicates statistical significance at $p < .05$.

5 Conclusion

We present preliminary results from the initial validation of a new aptitude battery TALL—which in part draws on existing aptitude batteries but distinguishes itself from them through its commitment to more recent theoretical models of aptitude, namely, Skehan's (2016) Stages Approach and Wen's (2016) Phonological / Executive WM model. Our results show that, in general, the scores from the subtests not only demonstrate consistent performance among university participants but also effectively differentiate between varying levels of ability in the theorized subcomponents of a multi-faceted aptitude construct. In addition, TALL uniquely comprises two complete and parallel suites in the aural and the written modalities, each with two versions of test items, allowing us to compare the effects of modality and test session in measuring aptitude (Pan, 2023), although these comparisons are beyond the scope of this brief report.

Some limitations invite further research. We acknowledge that, as the developers of TALL, we are not independent. By making TALL openly accessible, we invite others to validate this battery. Future research may involve, for example, wider learner populations, especially those with low L1 literacy levels. This is needed to reduce our collective reliance on highly educated participants, a field-wide sampling bias that threatens generalizability. Checking divergent validity by comparing TALL to other aptitude measures is also necessary. Substantive research on the predictive validity of TALL in explaining learning outcomes should be included in future research agendas, for which we collected some data beyond the scope of this brief report. Additionally, TALL was design for participants with L1 Chinese, and amendments for use with other populations merits investigation.

We underline the potential of TALL as a shared open research infrastructure for data collection and data accumulation [1]. First, TALL—and its adaptations—can be used as a reliable measure of aptitude constructs that enables data collection to be carried out remotely for free. This may facilitate

better sampling practices and multi-site studies to obtain larger and more diverse samples. Second, after TALL has been adequately validated, it can constrain researcher degrees of freedom caused by methodological variation, thereby increasing the comparability of results. Third, TALL allows consumers of research about aptitude to access samples of the battery, and producers of research to access the full battery. This should facilitate the scrutiny of replicability and reproducibility of our findings. Finally, in the long run, TALL could amass a cumulative open data pool; that is, aggregated data collected by using a uniform battery, to promote high-quality syntheses of research findings.

In sum, TALL's development represents a first step towards addressing "the chicken-and-egg conundrum" that researchers face: the demand for openly available validated instruments, which, in itself, requires instruments to be openly available in the first place in order to then validate them multiple times across a range of contexts.

Note

1. Careful investigation of ethical issues around responsible international data sharing and adherence to the relevant policies are needed if we develop TALL into providing an open, cumulative data pool. Therefore, we have temporarily incorporated the facility for open data access in the current TALL website with a view to eliciting feedback from reviewers and editors, but we will not make data access live until open data protection issues are resolved.

Acknowledgements

We thank Mengqiu Qin for her help in managing the development team at Chongqing University, Dr Giulia Bovolenta for her input in designing the test items, and Mrs Daiva Judges for recording the stimuli. We are extremely thankful for the valuable comments from the anonymous reviewers and Dr Daniel Isbell as the Editor handling this paper of the special issue. Any remaining errors are solely our responsibility.

References

Bokander, L., & Bylund, E. (2020). Probing the Internal Validity of the LLAMA Language Aptitude Tests. *Language Learning, 70*(1), 11–47. https://doi.org/10.1111/lang.12368

Carroll, J. B. (1981). Twenty-five years of research on foreign language aptitude. In K. C. Diller (Ed.), *Individual Differences and Universals in Language Learning Aptitude* (pp. 83-118). Rowley, MA: Newbury House.

Carroll, J. B. & Sapon, S. (1959). *Modern Language Aptitude Test (MLAT)*. The Psychological Corporation.

Gass, S., Winke, P., Isbell, D. R., & Ahn, J. (2019). How captions help people learn languages: A

working-memory, eye-tracking study. *Language Learning and Technology, 23*(2), 84–104. https://doi.org/10125/44684

Isbell, D. R., & Kim, J. (2023). Developer involvement and COI disclosure in high-stakes English proficiency test validation research: A systematic review. *Research Methods in Applied Linguistics, 2*(3), 100060. https://doi.org/10.1016/j.rmal.2023.100060

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Marsden, E., & Morgan-Short, K. (2023). (Why) Are Open Research Practices the Future for the Study of Language Learning? *Language Learning*. https://doi.org/10.1111/lang.12568

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412–433. https://doi.org/10.1037/met0000144

Meara, P. M., & Rogers, V. E. (2019). *The LLAMA Tests v3.* Lognostics. https://www.lognostics.co.uk/tools/LLAMA_3/index.htm

Newman, A., Bavik, Y. L., Mount, M., & Shao, B. (2021). Data Collection via Online Platforms: Challenges and Recommendations for Future Research. *Applied Psychology, 70*(3), 1380–1402. https://doi.org/10.1111/apps.12302

Pan, J. (2023). *Developing and validating an internet-based battery of Tests of Aptitude for Language Learning (TALL)* [Unpublished doctoral dissertation]. University of York.

Pimsleur, P. (1966). *Pimsleur Language Aptitude Battery (form S)*. New York: Harcourt, Brace and World, Incorporated. https://lltf.net/aptitude-tests/language-aptitude-tests/pimsleur-language-aptitude-battery

Purpura, J. E., Brown, J. D., & Schoonen, R. (2015). Improving the validity of quantitative measures in Applied Linguistics research. *Language Learning, 65*(S1), 37–75. https://doi.org/10.1111/lang.12112

R Core Team (2022). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.*

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. Psychological Assessment, 31(12), 1395–1411. https://doi.org/10.1037/pas0000754

Skehan, P. (2016). Foreign language aptitude, acquisitional sequences, and psycholinguistic processes. In G. Granena, D. Jackson & Y. Yilmaz (eds.), *Cognitive individual differences in L2 processing and acquisition* (pp.15–38). Amsterdam: John Benjamins. https://doi.org/10.1075/bpa.3

Suzuki, Y. (2021). Individual differences in memory predict changes in breakdown and repair fluency but not speed fluency: A short-term fluency training intervention study. *Applied Psycholinguistics, 42* (4), 969 - 995. https://doi.org/10.1017/S0142716421000187

TALL (n.d.). *Tests of Aptitude for Language Learning*. https://www.tall-webtest.com

Wen, Z. (2016). *Working memory and second language learning: Toward an integrated approach*. Multilingual Matters. https://doi.org/10.21832/9781783095735

Supplemental materials

Supplemental materials, including subtest manuals, audio and visual stimuli for subtests design, data for analysis, and R code are available in the IRIS database:

Pan, J., & Marsden, E. (2024). *TALL subtests manuals. Instructional / Intervention / Teaching / Training materials from "Developing internet-based Tests of Aptitude for Language Learning (TALL): An open research endeavour"* [Text/Materials]. IRIS Database, University of York, UK. https://doi.org/10.48316/AZZWN-mJXhL

Pan, J., & Marsden, E. (2024). *Materials in TALL subtests. Instructional / Intervention / Teaching / Training materials from "Developing and validating an internet-based battery of Tests of Aptitude for Language Learning (TALL)"* [Text/Materials]. IRIS Database, University of York, UK. https://doi.org/10.48316/kYHn1-ndaUe

Pan, J., & Marsden, E. (2024). *Scores from TALL subtests. Data from "Developing and validating an internet-based battery of Tests of Aptitude for Language Learning (TALL)"* [Dataset], IRIS Database, University of York, UK. https://doi.org/10.48316/dd14t-dbNml

Pan, J., & Marsden, E. (2024). *R markdown files. Code from "Developing and validating an internet-based battery of Tests of Aptitude for Language Learning (TALL)"* [Software]. IRIS Database, University of York, UK. https://doi.org/10.48316/dkxbx-Kw6ji

All of the above are also available on the project's OSF page: https://osf.io/czqxt/