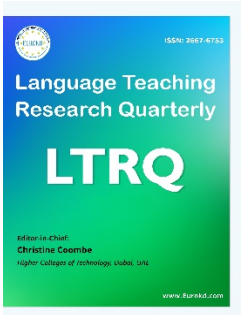


Language Teaching Research Quarterly

2021, Vol. 24, 23–43

Concurrent Validity of LLAMA_F: Measure of Language Analytic Ability as a Predictor of Morphosyntax Knowledge

Peter Kim

Teachers College, Columbia University, The United States of America

Received *16 June 2021* ***Accepted*** *18 October 2021*

Abstract

Foreign language aptitude is defined as one's potential to learn a second language. A language learner with higher aptitude is predicted to learn more, faster, and reach a higher level of proficiency. If this is the case, one way to validate the construct of aptitude and its measure is to conduct a validation study in which measures of aptitude is correlated with a learning outcome. This study aimed to conduct a concurrent validity of LLAMA_F, a foreign language aptitude test using grammaticality judgement test (GJT) as its concurrent criterion. This was done through dis-attenuated correlation, using reliability values obtained using classic test theory (CTT) and item analysis. The results show barely adequate reliability for LLAMA_F, high reliabilty values for GJT and a weak linear relationship between the two constructs. The findings of this study demonstrated that LLAMA_F might suffer from a lack of strong internal consistency. While the 200-item GJT was shown to be reliable, a few of its subcomponents were less than adequate. A further in-depth item analysis of GJT is needed in order to pare down the test and make it shorter for easier application and data collection. Finally, a nonlinear relationship between aptitude and measures of achievement is suggested for future research on language aptitude.

Keywords: *Concurrent Validity, CTT Reliability, Grammaticality Judgement Test, Language Aptitude*

Introduction

Foreign language (FL) aptitude is operationally defined as a special talent specific for learning a second language that exhibits variations among learners (Dörnyei & Skehan, 2003). John Carroll, one of the earliest and most influential FL aptitude researchers, defines it as "an individual's initial state of readiness and capacity for learning a foreign language, and probable

pk2505@tc.columbia.edu doi:10.32038/ltrq.2021.24.02

facility in doing so [given the presence of motivation and opportunity]" (Carroll, 1981, p.86). Variability in aptitude among individuals is manifested in the learning outcome, and those who have higher aptitude are predicted to reach a higher level of proficiency in the foreign language classroom and do so at a faster rate in the same fashion as one would imagine a musically talented or athletically gifted individual to outperform their less talented peers (Carroll, 1990; Stansfield, 1989). A seminal study by DeKeyser (2000) explored this direct relationship between aptitude and learning outcome. DeKeyser was motivated by two very important hypotheses in SLA: The Critical Period Hypothesis (CPH) and the Fundamental Difference Hypothesis (FDH). The former predicts that if CPH only applies to implicit language acquisition, then adults who are successful in L2 acquisition should have a high level of verbal ability, which affords them explicit learning of L2. The latter argues that with an increase in age, one's ability for implicit learning declines; therefore, adults must rely on explicit learning ability for language acquisition because there are fundamental differences between adult L2 acquisition and child L1 acquisition. The chief aim of DeKeyser's study was to test these hypotheses by examining the correlational relationships between variables of aptitude, age, and the learners' ultimate attainment. In his study, the correlation between GJT and aptitude score was .33 and significant. Results of the correlational analysis showed that adults who scored high on GJT also had a high verbal aptitude, which supports the notion that adults who have strong explicit learning skills are able to compensate for the loss of implicit skills as they mature past the age of the sensitive period. As a result, they are able to acquire L2 with greater success than those who have less aptitude.

Given the theoretical relationship between aptitude and language acquisition in adults, it is important for researchers to use reliable measures of foreign language aptitude while also pinpointing its nuanced relationship to specific aspects of language acquisition, i.e., morphosyntactic knowledge. The goal of this study was to examine the reliability of LLAMA_F, a freely available aptitude test of grammar inferencing ability, and the grammaticality judgement test (GJT) used by DeKeyser's 2000 study. In doing so, the current study will examine the psychometric properties of these instruments as well as explore the relationship between grammatical inferencing ability as measured by LLAMA_F and morphosyntactic knowledge as measured by grammaticality judgment test. Consequently, this study conducts a concurrent validity of LLAMA_F by analyzing its dis-attenuated correlation with GJTas the concurrent criterion.

Review of Literature

The relationship between grammar and aptitude has been arguably the most researched component of a learner's ultimate attainment with a consistent outcome of strong predictability (Li, 2015). For example, Bylund et al. (2010) showed that language aptitude was a reliable predictor of grammatical judgment test (GJT) score that resulted in a significant and positive correlation between the participants' aptitude and their performance on the GJT. In addition, Skehan's (2015) critical overview on the relationship between grammar and aptitude has shown that there is close proximity between the measure of grammatical sensitivity and the examinee's

intuition for metalinguistic awareness. This, in turn, had a positive relationship with higher L2 performance in general but to a lesser extent. Furthermore, examination of several grammatical focal structures (nominative-accusative case, clitic pronouns, direct object pronoun pseudo-cleft construction, S-V inversion, simple-complex morphosyntax) has shown that aptitude effects are related to the saliency and redundancy of the grammatical points. Citing de Graaff's (1997) large-scale study, Skehan postulated that aptitude works well for salient and redundant grammatical elements in the input because aptitude makes it more likely that the target point will be noticed. A meta-analysis by Li (2015) looked at 33 study reports in order to assess the association between language aptitude and L2 grammar acquisition. The study looked at 309 effect sizes and 3,106 L2 learners, and the results showed a moderate correlation between aptitude and L2 grammar learning (r=.31, 95% CI = .25 - .36).

One measure of language aptitude that has gained popularity due to being openly available as freeware is the LLAMA test. LLAMA (Meara, 2005) was developed as part of a research-training program at the University of Wales Swansea. It is a free aptitude test that is available to researchers, and its popularity has been increasing in recent years due to its accessibility with over 700 citations on Google scholar since 2013 (Rogers et al., 2017). LLAMA test consists of four components of vocabulary acquisition, sound recognition, sound-symbol correspondence and grammatical inferencing. These four components are assessed by four sub-tests that make up the LLAMA: LLAMA_B, LLAMA_D, LLAMA_E and LLAMA_F. The four components measure different aspects of language aptitude. Specifically, LLAMA_F was designed to measure an individual's ability to infer grammatical rules based on a limited number of exemplars.

The psychometric properties of LLAMA_F in the literature have shown that the internal consistency (Cronbach's alpha) based on a performance sample of 74 participants was $\alpha = .60$ (Granena, 2013). In a separate study with 135 college-level students, Cronbach's alpha was .66 (Granena, 2019). In addition, the internal validity of the LLAMA test battery has been examined by Bokander and Bylund (2019). Yet, a validation of LLAMA_F against a criterion of grammatical knowledge (morphosyntax) has not been addressed in the literature.

Regarding the criterion measure of morphosyntactic knowledge, timed GJTs have been cited in numerous studies as the instrument of choice used to assess learner's ultimate attainment of the target language grammar as well as their developing L2 proficiency (e.g., Abrahamsson & Hyltenstam, 2008; Birdsong, 1992; Schmid et al., 2014; Tsimpli et al., 2004; White & Genesee, 1996). GJT has been a standard instrument for measuring L2 learner's grammatical intuition, morphosyntactic knowledge and processing ability in SLA research. However, in recent years there has been some controversy surrounding the validity of GJT and what they actually measure. Studies using factor analysis demonstrated that timed GJT loads onto implicit knowledge factor while untimed GJT loads onto explicit knowledge factor (Bowles, 2011; Ellis & Loewen, 2007). Gutiérrez (2013), on the other hand, claimed that regardless of whether GJT is timed or untimed, grammatical items measure the test-taker's implicit knowledge, whereas ungrammatical items measure their explicit knowledge. Lately, the claim that GJT measures both

explicit and implicit knowledge has been challenged as studies by Ellis and Loewen (2007) and Bowles (2011) were heavily criticized for inappropriate use of factor analysis. Revalidation of GJT through confirmatory factor analysis concluded that GJTs are too coarse to be measures of implicit knowledge, and they are closer to measures of explicit than implicit knowledge (Vafaee et al., 2017). These findings in the literature regarding GJTs as measures of explicit knowledge aligns with DeKeyser's (2000) contention that adults who are successful in L2 acquisition should have a high level of aptitude linked to explicit learning of L2. That is, if explicit learning is to be the expected learning outcome of adult learners via language aptitude, GJTs are the appropriate instruments for analyzing such a relationship since both instruments are based on explicit cognitive processes.

One interesting finding from the analysis of aptitude studies on language achievement was that aptitude was only predicative in the initial stages of L2 grammar acquisition and less so during the latter stages of learning. That is, the effect of aptitude mostly operates in lower proficiency learners, and the same effect disappears once the learner reaches higher proficiency. For example, Li's (2015) meta-analysis on aptitude showed that high school students had higher correlations with aptitude than university students, and Li explains this not on the aptitude's effect on age but on aptitude's more pronounced effect on the initial stages of SLA. Findings like these have led Skehan to hypothesize that different components of aptitude operate during different stages of acquisition (Skehan, 2002). Li, on the other hand, hypothesized that explicit learning ability might be more relevant during the early stages of learning for certain types of salient linguistic features, while implicit abilities are more important during the latter stages of acquisition for non-salient features of the target language (Li 2015). Specifically, implicit aptitude is a set of cognitive abilities that allows the language user to make unconscious computations of the distributional and transitional probabilities of linguistic input, which in turn renders better acquisition of the target language (Li & DeKeyser, 2021).

In a most recent development regarding implicit aptitude, Li and Qian (2021) have argued that syntactic priming, which is the ability to reproduce linguistic structures based on the priming effect of previous exposure to the similar structure, is a valid measure of implicit language aptitude. When this was tested against the measure of explicit aptitude in LLAMA_F, syntactic priming was found to have divergent validity and thus distinct from explicit aptitude as measured by LLAMA_F but failed to converge with other measures of implicit aptitude. Li and Qian (2021)'s study, in conjunction with Granena's (2013) study, strongly suggest that LLAMA_F is a measure of more explicit aspects of language-learning aptitude. Therefore, if second language acquisition in adults is indeed differentially influenced by the type of aptitude being activated in the learner through varying proficiency levels, a correlation between aptitude and attainment by proficiency should indicate such disparity. Specifically, since LLAMA_F is a measure of explicit aptitude, as the proficiency of the participants increase, the correlation between aptitude score and achievement should decrease. The beginners are predicted to have the highest correlation between LLAMA_F and GJT, while the advanced proficiency learners are predicted to have the lowest correlation.

Lastly, it should be noted that the research on aptitude and language achievement is not only germane to contributing to the field of psychometrics and test validation but also relevant for classroom instruction. For example, Fu and Li (2021) have shown that the timing of corrective feedback on young EFL learners is associated with implicit and explicit aptitude and the predictive effectiveness of the corrective feedback. Therefore, in order to understand the nature of explicit and implicit aptitude, reliable and robust instruments for measuring the said constructs are needed. LLAMA_F certainly fits the bill, and due to the role of implicit versus explicit aptitude debate in L2 achievement, the changes in correlation strength with respect to proficiency must be considered. Based on these findings and the previous review of literature, the following research questions were proposed for the current study.

Research Questions

RQ_1: To what extent do LLAMA_F scores display satisfactory internal consistency, and to what extent are they composed of items covering an appropriate range of difficulty, supporting the scoring inference?

RQ_2: To what extent do GJT scores display satisfactory internal consistency, and to what extent are they composed of well-functioning items covering an appropriate range of difficulty, supporting the scoring inference?

RQ_3: Is there evidence of concurrent validity as measured by the dis-attenuated correlation between LLAMA_F and GJT, using reliability measures from RQ1 and RQ2, and do the correlations show differential effects among proficiency levels?

Method

Participants

One hundred and seventy-three (N=173) adult English learners (L1 Spanish) participated in this study. There were three inclusion criteria for selection based on the study's domain of generalizability. First, English was the participants' second language because the domain we wanted to generalize was for the L2 context. Second, the participants' first language was Spanish because the study aimed to control for the effect of participants' first languages by keeping them constant. Third, adults (over the age of 18) whose age of arrival to the United States was 12 or greater. This is because participants whose age of arrival is less than 12 are under the influence of a critical age period in which the acquisition of language may be under a different cognitive process. All participants were recruited through online advertisements.

Instruments

Measure of Linguistic Aptitude

For the current study, LLAMA_F, a grammar inferencing test, was used to measure the language analytic ability of the participants. According to the LLAMA manual, LLAMA_F was based on an earlier version of the test that was particularly effective at identifying outstanding analytical linguists. At the beginning of the test, examinees were given a series of pictures with a short

sentence in an artificial language that described each picture. Participants had five minutes to figure out the grammar of the unknown language. Then they were asked to match new picture prompts with sentences in the artificial language that correctly described them (Meara, 2005).

Measure of Grammar Knowledge

A timed grammaticality judgement test (GJT) was used to measure the participants' morphosyntactic knowledge of English. Specifically, the GJT in DeKeyser's 2000 study "*The robustness of critical period effects in second language acquisition"* (which in turn was adopted from Johnson and Newport's 1989 study) was used for the current study. DeKeyser found that adults who scored high on GJT also had high verbal aptitude scores, indicating that aptitude is positively correlated to one's performance on GJT. DeKeyser used a 200-item GJT that was an abridged adaptation of Johnson and Newport's 1989 instrument. In the study, DeKeyser reported a reliability coefficient of .91 for grammatical items and .97 for ungrammatical items on his GJT instrument. The GJT was comprised of 11 major categories of morphology and syntax, listed in Table 1.

Table 1

Rules Types Measured in GJT

11 Rule Types Tested in Grammaticality Judgment Test	
Past tense	Yes-no questions
Plural	Wh-questions
Third-person singular	Word order
Present progressive	Particle movement
Determiners	Subcategorization
Pronominalization	

There were exactly 100 grammatical and 100 ungrammatical items. Sentences were constructed with high-frequency words of one or two syllables in length, and only one violation of rule type was tested in ungrammatical/grammatical pairs. For example, past tense marking omitted in obligatory context has the following construction:

*Sandy fill a jar with cookies last night.**

Sandy filled a jar with cookies last night.

All items were randomized to ensure that they do not appear consecutively as a paired set of the same rule type. Scoring was done dichotomously with a point value of 0 for the wrong answer and 1 for the correct answer. GJT score for each participant was calculated as the total number of correctly marked items. The maximum total point possible was 200.

Data Collection Procedures

Data collection happened online due to COVID-19 restrictions on person-to-person contact, and it spanned three months from January 2021 to April 2021. All volunteers were asked to fill out a consent form followed by a brief background survey. The background survey was used to screen participants that met the selection criteria. The selected participants were invited to take the LLAMA_F test and the GJT test. The consent form, background survey, LLAMA_F and GJT

were made available online through Qualtrics. R was used for all descriptive statics and CTT measure of reliability and correlation analysis.

Results

Table 2
Descriptive Statistics and Reliability

	Number of items	Min	Max	Mean	SD	Reliability (Coefficient Alpha)
LLAMA_F	20	3	19	14.36	3.06	0.64
GJT	200	52	186	124.81	29.93	0.962

The table above shows that the internal consistency as measured by coefficient alpha for the twenty items of LLAMA_F was found to be 0.64. A value of Cronbach's alpha greater than 0.6 is considered barely acceptable according to some researchers (e.g., Griethuijsen et al., 2014; Wim et al., 2008). The alpha value of 0.64 is considered to meet the minimal threshold of "adequate" (Taber, 2018); however, others argue that the acceptable range of alpha begins at 0.7 (Tavakol & Dennick, 2011). The value of 0.64 agrees with the previously reported reliability values of LLAMA_F in literature – for example, $\alpha = .60$ (Granena, 2013) and $\alpha = .66$ (Granena, 2019). Nevertheless, 0.64 is considered to be a low-end of the range in terms of what is considered acceptable, and low values of alpha could be due to lack of internal consistency (poor inter-connection or relatedness between items), a low number of items, or a presence of heterogeneous constructs that are not related to each other (Tavakol & Dennick, 2011). On the other hand, the reliability of GJT was found to be excellent at 0.962. Their component reliabilities by 11 rule types are listed below.

Table 3
Descriptive Statistics and Reliability of GJT by Subcomponents

Rule type tested	Number of items	Min	Max	Mean	SD	Reliability (Coefficient Alpha)
Past tense	18	3	18	10.70	2.86	0.618
Plural	18	4	18	11.07	3.45	0.764
3rd person singular	16	2	16	10.04	3.22	0.749
Present progressive	12	2	12	8.89	2.68	0.78
Determiners	14	2	14	8.99	2.55	0.655
Pronominalization	16	1	16	9.76	3.34	0.745
Yes/no questions	24	4	24	14.49	4.47	0.784
Wh-questions	12	2	12	7.55	2.52	0.68
Word order	30	9	30	20.95	4.61	0.766
Particle movement	16	3	16	8.43	2.64	0.53
Subcategorization	20	5	18	11.54	3.00	0.611

Table 3 shows that if the subcomponents are judged by the criterion of alpha greater than .60, only one category, particle movement, fails to be satisfactory. A criterion of alpha greater than .70 makes five categories (past tense, determiners, wh-questions, particle movement, and

subcategorization) to be questionable regarding their internal consistency. Next, item analysis of LLAMA_F and GJT were carried out and reported below (Table 4). Due to a large number of items (200), GJT item analysis is reported in the appendix (Appendix A).

Table 4
LLAMA_F Item Analysis

Item	Item Mean	Corrected item-total pBis	Alpha if item deleted
1	0.761	0.339	0.614
2	0.872	0.319	0.619
3	0.883	0.186	0.632
4	0.854	0.331	0.617
5	0.645	0.093	0.645
6	0.680	0.274	0.621
7	0.837	0.272	0.623
8	0.808	0.327	0.616
9	0.872	0.234	0.628
10	0.709	0.333	0.614
11	0.720	0.175	0.634
12	0.750	0.279	0.621
13	0.558	0.116	0.643
14	0.616	0.318	0.615
15	0.750	0.411	0.604
16	0.500	0.119	0.643
17	0.761	0.155	0.636
18	0.761	0.174	0.634
19	0.715	0.201	0.631
20	0.308	0.016	0.654

Item analysis of LLAMA_F showed an acceptable range of item difficulty, which was identical to item mean for dichotomously scored items, ranging from 0.308 to 0.883. In general, no single item appeared to have been too difficult (mean less than 0.3) or too easy (mean greater than 0.9). However, looking at the point-biserial item-total corrected (without the item itself in total) correlation, thirteen items were found to be less than 0.3 (items 3, 5, 6, 7, 9, 11, 12, 13, 16, 17, 18, 19, 20). Still, only two items had "if alpha deleted" greater than 0.64, which was the reliability of the aptitude test. They were item 5 with 0.645 and item 20 with 0.654. Removing 13 items that have a point-biserial correlation less than 0.3 was considered impractical because that would eliminate more than half of the items in the current test. This would reduce the total number of items from 20 to 7 items. Instead, two items (item 5 and 20) were removed based on their "alpha if deleted" criterion of greater than 0.64, and this led to an improvement in the reliability of 0.661, an increase of 0.02.

Next, the analysis of GJT items showed that there were not any items that were too difficult (item mean less than 0.1), but there were 10 items that were considered hard (item mean less than 0.3). These were items 5, 36, 37, 82, 85, 115, 146, 153, 156, 194. In addition, sixteen items may have been too easy (item mean greater than 0.9). These were items 2, 21, 42, 50, 66, 90, 101, 133, 135, 143, 151, 158, 164, 167, 178, 196. Furthermore, the most serious violations of

good item values were found for point-biserial correlations less than 0.30. There were 89 items that fit the bill, a rather large number. One reason why so many items had such low correlations may be due to the fact that the test is comprised of 11 unique rule types, and how one does on one particular rule has no bearing on their performance on the other rules. For example, items in past tense may have no bearing on one's knowledge of particle movement or vice versa. In addition, a large number of beginners (about 50 in the current sample) who had little morphosyntactic knowledge might have contributed to making this correlation low since with beginners, their performance on one item is less likely to be related to their overall performance. Nevertheless, all items with bad item means (<0.1 or >0.9) or point-biserial correlations less than 0.30 were removed, which resulted in 108 items with a coefficient alpha of 0.974. If test users are interested in measuring English learners' overall morphosyntactic knowledge without too much concern for the exact rule types tested, then a shortened version of the GJT with 108 items is expected to function just as well in less time compared to the original GJT with a full set of 200 items.

Finally, in order to investigate the concurrent validity of LLAMA_F as a measure of grammar inferencing aptitude, LLAMA_F was correlated with GJT using a correction for attenuation. According to Bandalos (2018), the relationship between a predictor test (LLAMA_F) and a criterion (GJT) will be attenuated to the extent that both instruments are not measuring reliably. Because the correlation between the two scores is restricted by their reliabilities, correction for attenuation (equation 1 below) estimates how well the given instrument predicts the criterion score in spite of having less than perfect reliability.

$$\rho_{t_x t_y} = \frac{\rho_{XY}}{\sqrt{\rho_{XX`}\rho_{YY`}}} \quad (1)$$

Based on the reliabilities obtained from the current study, dis-attenuated correlations were calculated between LLAMA_F and GJT.

Table 5

Correlations between LLAMA_F and GJT

Rule type tested	Dis-attenuated correlation	Attenuated correlation
Past tense	0.26	0.213**
Plural	0.34	0.182*
3rd person singular	0.14	0.098
Present progressive	0.16	0.115
Determiners	0.16	0.107
Pronominalization	0.22	0.165*
Yes/no questions	0.23	0.166*
Wh-questions	0.078	0.052
Word order	0.257	0.180*
Particle movement	0.282	0.165*
Subcategorization	0.262	0.164*
Composite	0.228	0.181*

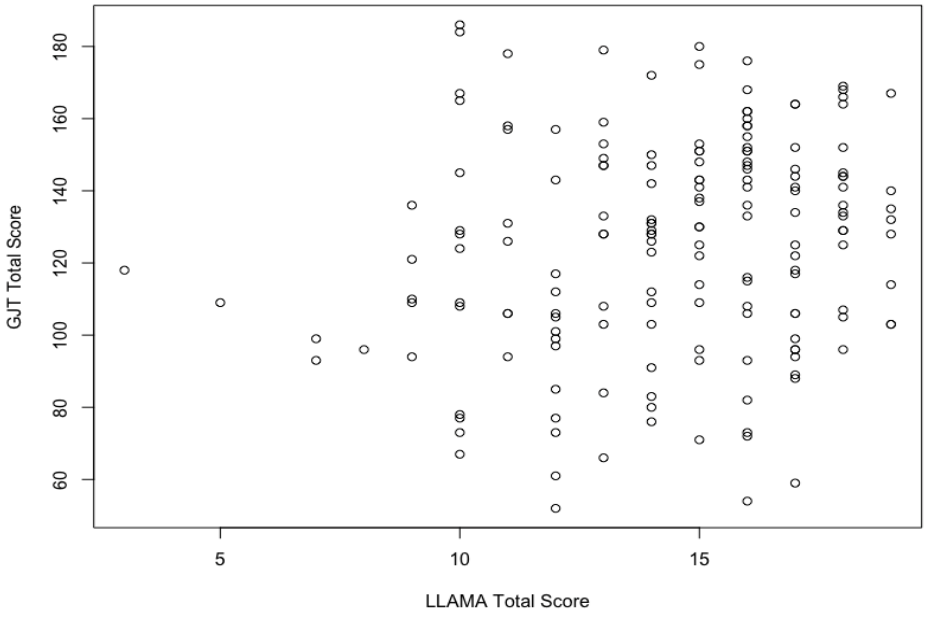
Note 1. * significant at the 0.05 level ** significant at the 0.01 level

Table 5 above shows that, as expected, the dis-attenuated correlation was higher than the Pearson product-moment correlation (attenuated) between LLAMA_F and GJT. For the attenuated correlations, all components, as well as the composite, were found to be significant at either the 0.05 or 0.01 level, except for present progressive, determiners, and wh-questions. The overall composite correlation between LLAMA_F and GJT was shown to be 0.228, with an R^2 value of 0.0519. This means that about 5% of the variance in GJT can be explained by one's LLAMA_F scores. This is considered a weak relationship (Akoglu, 2018) and lower than the moderate correlation found between aptitude and L2 grammar learning in the previous literature (r=.31, 95% CI = .25 - .36) (Li, 2015). However, the correlation of 0.31 was based on a meta-analysis that incorporated effect sizes of multiple measures of aptitude and grammar learning outcomes (Li, 2015), whereas, in the current study, one's grammatical inferencing ability was correlated to their morphosyntactic knowledge as measured by grammaticality judgment test. Therefore, the results of the study speak directly to how one's grammatical inferencing aptitude has a concurrent relationship to their actual knowledge of English morphosyntax. Still, the low correlation raises doubts about the validity of LLAMA_F as a measure of grammar aptitude, given that it was also found to have rather low reliability of 0.64. It may also question the validity of the construct of aptitude as a predictor of learning outcomes.

The low correlation between LLAMA_F and GJT is an indication that a simple linear relationship between the two constructs is not tenable, as the simple scatter plot between the two illustrates below (Figure 1).

Figure 1

Scatter Plot of LLAMA_F total Score vs GJT Total Score



This relationship was further explored by dividing the participants' GJT scores into three levels based on their percentile. The motivation behind this analysis was based on the conclusion

of previous research that the effect of aptitude mostly operates in lower proficiency learners, and the same effect disappears once the learner reaches higher proficiency (Li, 2015). In the current study, the low proficiency group was identified as being first quarter percentile or below (GJT <= 104.5); the intermediate group was identified as the interquartile range (GJT >104.5 and <147). The advanced proficiency group was identified as being in the third quarter percentile or above (GJT >147). The dis-attenuated correlation between the three groups of proficiency and LLAMA_F showed the following relationships: low proficiency was 0.113 (n=43), intermediate proficiency was 0.39 (n=84), and advanced proficiency was -0.326 (n=41). It appeared that LLAMA_F was most effective in predicting GJT for the intermediate group, and for the advanced group, the relationship was actually negative.

Discussion

Regarding research question one, LLAMA_F displayed a rather weak internal consistency of 0.64; however, no item appeared to have been too difficult or too easy, based on mean difficulty indices. Thirteen items were found to be less than 0.3 for point-biserial and item-total corrected correlations, and these items need to be further investigated in future research. Based on these two results, the unidimensionality of LLAMA_F may be questioned. According to Bokander and Bylund's (2019) examination of LLAMA, two-component principal component analysis of LLAMA_F produced one cluster of 10 items that were comprised of less complex grammatical rules. The second cluster of 7 items contained more complex rules. Therefore, the weak internal consistency of LLAMA_F may be due to its lack of unidimensional property based on the complexity of grammatical rules being tested. This does not necessary mean that LLAMA_F is not a valid test of explicit aptitude. Rather, the low covariance among the items could simply be due to the bi-dimensionality of two different types of grammatical rules being tested. In addition, the low alpha value of 0.64 should be considered within the broader context in which the test was administered. First, in the current study the level of sample heterogeneity may have contributed to the attenuation of LLAMA_F's estimate of internal consistency. The participants in the current study were homogenous L1 Spanish group, which may have contributed to less variance in true score compared to a heterogenous sample in which L1 backgrounds are more varied. However, a more serious factor affecting internal consistency is the ceiling effect of LLAMA_F. With a mean of 14.36 and a standard deviation of 3.06, 68 percent of the data falls above 10, or better than guessing on all 20 items with a 50% chance at getting the right answer for each item. One oblivious solution to rectify this is to increase the number of items so that the variance among the sample participants may increase. Specifically, to improve the reliability of LLAMA_F and its generalization inference, future studies should consider adding more items of both complex and simple rules through the use of D-studies under the G-theory framework (Shavelson & Webb, 1991). In short, with regard to reliability, the biggest weakness of LLAMA_F may be the low number of items. Fortunately, given that there are only 20 items in the test, adding more is the most direct and also the most expedient means of increasing the internal consistency of the test. As research on language aptitude continues to make advances, it

is important for key instruments of aptitude such as LLAMA to have robust and reliable psychometric properties. Due to the free and online availability of LLAMA, it has become one of the most popular test of language aptitude in recent years (Rogers et al., 2017). Thus, the future direction of aptitude research should first look to expand upon LLAMA_F so that the test's reliability falls within the acceptable range according to the best practices recommended by the field of language assessment and psychological measurement. Simply put, the test needs more items.

Research question two explored the psychometric properties of DeKeyser's (2000) GJT and found the test to have strong internal consistency with reliability of 0.962. When the 200-items were broken down by 11 rule types, reliabilities ranged from 0.53 to 0.766. Researchers interested in examining the differential effects of grammatical rule types should note these reliabilities when conducting correlational studies. From a more practical perspective, a 200-item GJT is rather burdensome for test-takers, and one way to remedy the possible effects of fatigue is to reduce the number of items. If a researcher is simply interested in the participants' overall knowledge of English morphosyntax, a shortened version of 108-items could provide them with a strong reliability of 0.974. The trade-off here is that the shortened version of the test would not be able to provide a nuanced analysis of the learner's morphosyntactic knowledge broken down by the specific rule types listed in Table 1. Furthermore, it is troubling that a highly reliable test with strong internal consistency is composed of elements (the rule types) that individually have weak reliabilities. This suggests that the strong reliability of GJT may be due to a large number of items in the test. In order to increase the reliability of the individual rule types, the number of items that test for these rule types could be added to the test. Adding more items per rule type, however, would increase the length of a test that is already quite long at 200 items. Attempts to achieve high reliabilty for all 11 rules in a single grammaticality judgement test that is also admissible in a short period of time may be untenable. The best practice is for researchers and practitioners to consider the purpose of the test use and judiciously choose between a shorter test with a single dimension of morphosyntactic knowledge versus 11 dimensions of rule types.

Finally, the findings of research question three did not fully support the claim of differentiated aptitude effects based on proficiency as stated in the literature. That is, the highest correlation between proficiency level (based on GJT scores) and aptitude was found for the intermediate group. In addition, the overall relationship between aptitude and GJT was found to be weakly correlated at 0.228. Clearly, the findings warrant more research as the relationship between grammatical inferencing aptitude and grammaticality judgement score was found to be rather nonlinear (Figure 1) and difficulty to decipher using correlational analysis alone. More research is needed in this area in order to understand the nuanced relationship between the two constructs and how they operate under second language acquisition. One interesting point to note is that when GJT was analyzed according to its 11 componential rule types, explicit and salient rules such as past tense and plural markings had the highest significant correlation with aptitude at 0.26 and 0.34, respectively. These findings provide support to the idea that different grammar structures show different correlations with age because not all structures are equally sensitive to

the CPH effect (DeKeyser, 2000). Specifically, in DeKeyser's (2000) study, present progressives, determiners, wh-questions, plurals and subcategorization were highly correlated with age of arrival, whereas word order, yes-no questions and pronoun gender did not show differential proficiency. DeKeyser explained the discrepancies in correlational significance by alluding to respective structure's perceptual salience and their interaction with implicit/explicit learning. While the saliency of explicit versus implicit learning discrepancy may have played a role, an alternative explanation for the differential effect of aptitude on rule type can be attributed to the test characteristics of the aptitude test. In LLAMA_F, the grammar of the artificial language featured in the test is highly inflectional. One of the key aspects of the aptitude test is to recognize the meaning of inflections and derivations, often expressed as suffixes in the target language and be able to extrapolate their meanings in new sentences. In English, plural markings and past tenses are inflectional and derivational morphemes found at the end of a word as suffixes. The overlap, therefore, between these two particular grammar points and the grammatical inferencing ability tested in LLAMA_F are explainable in terms of their structural similarity. What this implies is that grammatical inferencing ability tests such as LLAMA_F must expand their repertoire of rules to include syntax, prefixes, as well as interrogative and imperative sentences. Note that in Table 5, the wh-question had the lowest correlation with aptitude. Given that LLAMA_F does not feature any interrogative sentence types, the low correlation cannot be interpreted to mean that language aptitude is a weak indicator of L2 learners' ability to express wh-questions. Therefore, one way to improve the quality of the aptitude test is to add different types of sentences and grammatical structures in order to provide rigour to its generalizability in those contexts.

Conclusion

Research in aptitude and language attainment requires adequate psychometric instruments with strong measures of internal consistency. To this end, LLAMA_F and DeKeyser's (2000) GJT were analyzed separately for their reliability. The two tests were then correlated using dis-attenuated correlation in order to test the concurrent validity of LLAMA_F, a measure of language analytic ability against GJT as a test of morphosyntactic knowledge. The findings of this study demonstrated that LLAMA_F might suffer from a lack of strong internal consistency. One way to improve the reliability of LLAMA_F is to simply add more items. However, the new items should not be added as mere extensions of the current test. More grammar rules beyond suffixation should be included in the test in order to improve the generalizability of the test scores to other aspects of grammar. While the 200-item GJT was shown to be reliable, a few of its subcomponents were less than adequate. This creates a situation in which researchers interested in specific grammatical elements of the test must deal with their low reliabilities even though the test as a whole is highly reliable. Test users must choose between a single dimension of morphosyntactic knowledge with high reliability or choose the individual rule types that fit their research questions the best. In the case of the latter, more items are needed to increase the reliability of each rule type. In the case of the former, a further in-depth item analysis is needed

in order to pare down the test and make it shorter for easier application and data collection. Finally, the concurrent validity framework between aptitude and a criterion measure suffers from the fact that aptitude is theorized as a predictor of success in learning outcomes and not necessarily a concurrent criterion in which aptitude tests can be validated. Therefore, a more appropriate validity analysis in future research should involve a predictive validity study, in addition to a concurrent validity study. Beginners should be assessed for their aptitude, and learning outcomes should be measured at a time much later than when the aptitude test was first given.

References

Abrahamsson, N. & Hyltenstam, K. (2008). The robustness of aptitude effects in near-native second language acquisition. *Studies in Second Language Acquisition*, *30*(4), 481–509. https://doi.org/10.1017/S027226310808073X

Akoglu H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, *18*(3), 91–93. https://doi.org/10.1016/j.tjem.2018.08.001

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. Guilford Publications.

Birdsong, D. (1992). Ultimate attainment in second language acquisition. *Language, 68*(4), 706–755. https://doi.org/10.2307/416851

Bokander, L., & Bylund, E. (2019). Probing the internal validity of the LLAMA language aptitude tests. *Language Learning*, *70*(1), 11-47. https://doi.org/10.1111/lang.12368

Bowles, M. A. (2011). Measuring implicit and explicit linguistic knowledge: What can heritage language learners contribute? *Studies in Second Language Acquisition, 33*(2), 247–271. https://doi.org/10.1017/S0272263110000756

Bylund, E. & Abrahamsson, N. & Hyltenstam, K. (2010). The role of language aptitude in first language attrition: The case of prepubescent attriters. *Applied Linguistics*, *31*(3), 443–464. https://doi.org/10.1093/applin/amp059

Carroll, J. B. (1981). Twenty-five years of research on foreign language aptitude. In K.C. Diller (Ed.), *Individual differences & universals in language learning aptitude* (pp. 83–118). Newbury House.

Carroll, J. B. (1990). Cognitive abilities in foreign language aptitude: Then and now. In T. Parry and C. Stansfield (Eds.), *Language Aptitude Reconsidered* (pp. 11–29). Prentice-Hall.

de Graaff, R. (1997). The eXperanto experiment: Effects of explicit instruction on second language acquisition. *Studies in Second Language Acquisition, 19*(2), 249–276. https://www.jstor.org/stable/44488685

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, *22*(4), 499–533. https://doi.org/10.1017/S0272263100004022

Dörnyei, Z., & Skehan, P. (2003). Individual differences in second language learning. In C. J., Doughty, & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 589–630). Blackwell. https://doi.org/10.1002/9780470756492.ch18

Ellis, R., & Loewen, S. (2007). Confirming the operational definitions of explicit and implicit knowledge in Ellis (2005): Responding to Isemonger. *Studies in second language acquisition, 29*(1), 119–126. https://doi.org/10.1017/S0272263107070052

Fu, M., & Li, S. (2021). The associations between implicit and explicit language aptitude and the effects of the timing of corrective feedback. *Studies in Second Language Acquisition, 43*(3), 498–522. https://doi:10.1017/S0272263121000012

Granena, G. (2013). Cognitive aptitudes for second language learning and the LLAMA language aptitude test. In G. Granena & M. Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 105–130). John Benjamins.

Granena, G. (2019). Cognitive aptitude and L2 speaking proficiency: Links between LLAMA and Hi-LAB. *Studies in Second Language Acquisition, 41*(2), 313–336. https://doi.org/10.1017/S0272263118000256

Griethuijsen, R. A. L. F., Eijck, M. W., Haste, H., Brok, P. J., Skinner, N. C., Mansour, N., … & BouJaoude, S. (2014). Global patterns in students' views of science and interest in science. *Research in Science Education, 45*(4), 581–603. https://doi.org/10.1007/s11165-014-9438-6

Gutiérrez, X. (2013). The construct validity of grammaticality judgment tests as measures of implicit and explicit knowledge. *Studies in second language acquisition, 35*(3), 423–449. https://doi.org/10.1017/S0272263113000041

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, *21*, 60–99. https://doi.org/10.1016/0010-0285(89)90003-0

Li, S. (2015). The associations between language aptitude and second language grammar acquisition: A meta analytic review of five decades of research. *Applied Linguistics, 36*(3), 385–408. https://doi.org/10.1093/applin/amu054

Li, S. (2016). The construct validity of language aptitude: A meta-Analysis. *Studies in Second Language Acquisition*, *38*(4), 801–842. https://doi.org/10.1017/S027226311500042X

Li, S., & DeKeyser, R. (2021). Implicit language aptitude: conceptualizing the construct, validating the measures, and examining the evidence: introduction to the special issue. *Studies in Second Language Acquisition, 43*(3), 473-497. https://doi.org/10.1017/S0272263121000024

Li, S., & Qian, J. (2021). Exploring syntactic priming as a measure of implicit language aptitude. *Studies in Second Language Acquisition, 43*(3), 574–605. https://doi:10.1017/S0272263120000698

Meara, P. (2005). LLAMA language aptitude test manual. Retrieved from http://www.lognostics.co.uk/tools/llama/llama_manual.pdf

Rogers, V., Meara, P., Barnett-Legh, T., Curry, C., & Davie, E. (2017). Examining the LLAMA aptitude tests. *Journal of the European Second Language Association*, *1*(1), 49–60. http://doi.org/10.22599/jesla.24

Schmid, M. S., Gilbers, S., & Nota, A. (2014). Ultimate attainment in late second language acquisition: Phonetic and grammatical challenges in advanced Dutch-English bilingualism. *Second Language Research, 30*(2), 129–157. https://doi.org/10.1177/0267658313505314

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* SAGE Publications.

Skehan, P. (2002). Theorizing and updating aptitude. In Robinson, P. (Ed.) (2002*). Individual Differences and Instructed Language Learning* (pp.69–94). Philadelphia, NL: John Benjamins Publishing Company. https://doi.org/10.1075/lllt.2.06ske

Skehan, P. (2015). Foreign language aptitude and its relationship with grammar: a critical overview. *Applied Linguistics, 36*(3), 367–384. https://doi.org/10.1093/applin/amu072

Stansfield, C. W. (1989). *Language Aptitude Reconsidered*. *ERIC Digest*. ERIC Clearinghouse on Languages and Linguistics.

Taber, K. S. (2018). The use of Cronbach's Alpha when developing and reporting research instruments in science education. *Research in Science Education, 48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. 10.5116/ijme.4dfb.8dfd

Tsimpli, I., Sorace, A., Heycock, C., & Filiaci, F. (2004). First language attrition and syntactic subjects: A study of Greek and Italian near-native speakers of English. *International Journal of Bilingualism*, *8*(3), 257–277. https://doi.org/10.1177/13670069040080030601

Vafaee, P., Suzuki, Y., & Kachisnke, I. (2017). Validating grammaticality judgment tests: Evidence from two new psycholinguistic measures. *Studies in Second Language Acquisition, 39*(1), 59–95. https://doi.org/10.1017/S0272263115000455

White, L., & Genesee, F. (1996). How native is near-native? The issue of ultimate attainment in adult second language acquisition. *Second Language Research, 12*(3), 233–265. https://doi.org/10.1177/026765839601200301

Wim J., Katrien W., Patrick D. P., & Patrick V. K. (2008). *Marketing Research with SPSS*. Pearson.

Appendix A

Complete Item Analysis of GJT

Item	Item Mean	Corrected item-total pBis	Alpha is item deleted
1	0.55813953	0.35495317	0.96144348
2	0.98255814	-0.0111684	0.96171204
3	0.3255814	0.3569109	0.96144033
4	0.87209302	0.10612211	0.9617069
5	0.22093023	0.41869929	0.96136959
6	0.74418605	0.27214266	0.96155007
7	0.3255814	0.5310358	0.9612012
8	0.83139535	-0.0678488	0.96191415
9	0.49418605	0.48312947	0.96125597
10	0.65697674	0.13624389	0.96174534
11	0.54069767	0.5940819	0.96109391
12	0.37209302	0.24470701	0.9615985
13	0.35465116	0.38104129	0.96140649
14	0.86046512	0.2700845	0.96154694
15	0.54069767	0.29291612	0.96153383
16	0.88953488	0.04636143	0.96175324
17	0.66860465	-0.175299	0.96216637
18	0.88372093	0.23615317	0.96158027
19	0.62209302	0.33795528	0.96146715
20	0.30232558	0.22236148	0.96162083
21	0.97093023	-0.0797217	0.96175863
22	0.70348837	0.40113971	0.96138164
23	0.41860465	0.35375185	0.96144507
24	0.77906977	0.25699233	0.96156554
25	0.30813953	0.31767507	0.96149343
26	0.61046512	0.00824474	0.96193447
27	0.87209302	-0.1172503	0.96192306
28	0.89534884	0.08187039	0.96171774

29	0.64534884	0.54277484	0.96117958
30	0.8372093	0.29891064	0.96151569
31	0.3255814	0.35985422	0.96143629
32	0.43604651	0.34760043	0.9614541
33	0.77906977	0.47949602	0.96129569
34	0.80813953	0.33902959	0.96146854
35	0.68604651	0.65260371	0.96103752
36	0.18604651	0.16400829	0.96166764
37	0.26744186	0.40732796	0.96137631
38	0.88953488	0.0177548	0.96177925
39	0.60465116	-0.0961868	0.96208307
40	0.85465116	0.52725554	0.96128172
41	0.63953488	0.3603012	0.96143549
42	0.93604651	0.1033199	0.96168051
43	0.45930233	0.13517549	0.96176248
44	0.60465116	0.60171693	0.96108916
45	0.51162791	0.60929625	0.96107051
46	0.56395349	0.06243422	0.96186533
47	0.73837209	0.16458072	0.96168818
48	0.55232558	0.58241617	0.96111171
49	0.87209302	0.07105518	0.96174091
50	0.90116279	0.32778502	0.96150089
51	0.61627907	0.62084025	0.96106359
52	0.72674419	0.61732121	0.96110089
53	0.70930233	0.0623701	0.96182999
54	0.70348837	0.10024084	0.96178198
55	0.3255814	0.20716178	0.96164493
56	0.47674419	0.39436528	0.96138606
57	0.55232558	0.56564184	0.96113626
58	0.37790698	0.32056679	0.96149177
59	0.55813953	0.28754211	0.96154121
60	0.66860465	0.37608293	0.9614139
61	0.43604651	0.47784907	0.96126498
62	0.54651163	0.47209344	0.96127278
63	0.33139535	0.39209037	0.96139186
64	0.62209302	0.21551052	0.96164022
65	0.87790698	0.21630647	0.9615993
66	0.9244186	0.3693802	0.96148058

67	0.85465116	-0.0077472	0.96183133
68	0.4244186	0.48540376	0.96125454
69	0.50581395	0.53004792	0.96118707
70	0.49418605	0.57833416	0.96111604
71	0.53488372	0.29847148	0.96152585
72	0.74418605	0.52512313	0.96122709
73	0.79069767	0.44219935	0.96134424
74	0.62209302	0.34608147	0.96145564
75	0.53488372	0.68803563	0.96095536
76	0.53488372	0.53056431	0.96118686
77	0.8255814	0.25115476	0.96156793
78	0.81976744	0.08108576	0.96175849
79	0.69186047	-0.1316017	0.96209479
80	0.48837209	0.64681641	0.9610152
81	0.6627907	0.22530157	0.96162173
82	0.19767442	0.42211596	0.96137122
83	0.85465116	0.18556261	0.96163354
84	0.5872093	0.74233769	0.96088235
85	0.28488372	0.31532966	0.96149581
86	0.75	0.02689298	0.96185806
87	0.59302326	0.07642282	0.96184133
88	0.8255814	0.35773533	0.96144992
89	0.44186047	0.50709695	0.96122211
90	0.94186047	0.03890583	0.96172198
91	0.36627907	0.47710571	0.96127068
92	0.73255814	0.68732704	0.96101206
93	0.41860465	0.55178954	0.96115862
94	0.51744186	0.56624287	0.96113399
95	0.73837209	0.15433782	0.96170127
96	0.61046512	0.32115234	0.96149129
97	0.77906977	-0.0538089	0.96193969
98	0.34302326	0.3533693	0.96144516
99	0.4127907	0.27049643	0.96156465
100	0.60465116	0.14268173	0.96174544
101	0.9127907	-0.0817443	0.96184207
102	0.76744186	0.36706828	0.96143104
103	0.77325581	0.3467165	0.96145652
104	0.6627907	0.64816574	0.96103598

105	0.33139535	0.35021342	0.9614495
106	0.81395349	0.25113371	0.96156896
107	0.88372093	0.06543943	0.9617394
108	0.69767442	-0.2198211	0.96220791
109	0.8255814	0.03063609	0.9618111
110	0.73837209	-0.1038283	0.96202966
111	0.59883721	0.47828765	0.96126619
112	0.84302326	-0.079601	0.96191597
113	0.74418605	0.17596121	0.96167224
114	0.73837209	0.26405309	0.96156095
115	0.29651163	0.52964107	0.96120954
116	0.48255814	0.62068641	0.96105385
117	0.84302326	-0.1426223	0.96198219
118	0.38372093	0.24076407	0.96160509
119	0.46511628	0.39915347	0.96137911
120	0.77906977	0.23855187	0.96158783
121	0.4244186	0.47218743	0.96127371
122	0.44186047	0.46841852	0.9612785
123	0.43604651	0.53698607	0.96117884
124	0.52325581	0.63266993	0.96103637
125	0.58139535	0.06631803	0.9618575
126	0.84302326	0.59021259	0.96120583
127	0.43604651	0.47904663	0.96126324
128	0.79651163	-0.0900853	0.96196914
129	0.8372093	0.34305938	0.96146811
130	0.76162791	0.28977292	0.96152668
131	0.65697674	0.66832518	0.96100603
132	0.62209302	0.2851874	0.96154182
133	0.90697674	0.32800814	0.96150323
134	0.51162791	0.37669096	0.96141188
135	0.98837209	0.11752708	0.96166508
136	0.5	0.27519163	0.96156005
137	0.65116279	0.7089376	0.96094712
138	0.65116279	0.5304752	0.9611978
139	0.46511628	0.57224996	0.96112569
140	0.63372093	0.63461177	0.96104747
141	0.44767442	0.41853221	0.96135102
142	0.78488372	0.08500276	0.96177041

143	0.97674419	0.05604491	0.96168928
144	0.47674419	0.48938073	0.96124697
145	0.36046512	-0.0077558	0.96194899
146	0.14534884	0.17407358	0.96164532
147	0.84883721	-0.0537564	0.96188358
148	0.78488372	0.51429206	0.96125574
149	0.80232558	0.5054973	0.96127391
150	0.5	0.27243889	0.96156407
151	0.9244186	0.2071526	0.9616056
152	0.87790698	0.12134457	0.96168973
153	0.12209302	0.17973999	0.96163414
154	0.81395349	-0.1064852	0.9619726
155	0.87209302	0.05236407	0.96175902
156	0.22674419	0.11561889	0.96173832
157	0.6627907	0.34464827	0.9614572
158	0.90116279	0.17754774	0.96163155
159	0.75581395	0.55653335	0.96119143
160	0.6744186	0.51124065	0.96122845
161	0.63953488	0.54640875	0.96117352
162	0.44186047	0.47001251	0.96127617
163	0.89534884	0.19547526	0.96161664
164	0.95348837	0.03042517	0.96171857
165	0.45930233	-0.2642924	0.96233619
166	0.70930233	0.20496221	0.96164203
167	0.91860465	0.35991694	0.96148363
168	0.49418605	0.60218308	0.96108092
169	0.53488372	0.47415727	0.96126948
170	0.87209302	0.34883772	0.9614708
171	0.31395349	0.43319075	0.96133685
172	0.35465116	0.54264956	0.96117976
173	0.54069767	0.49454745	0.96123979
174	0.6744186	0.31580132	0.96149659
175	0.56395349	0.63517645	0.96103545
176	0.55813953	0.45197377	0.96130245
177	0.86627907	0.49897018	0.96131949
178	0.96511628	0.08295206	0.96168142
179	0.40697674	0.34439363	0.96145842
180	0.37790698	0.11943208	0.96177554

181	0.55232558	0.62279306	0.96105255
182	0.80813953	0.01939525	0.9618346
183	0.3255814	0.36195676	0.96143341
184	0.37209302	-0.0669479	0.96203498
185	0.63372093	0.65154006	0.96102341
186	0.78488372	0.54893304	0.96121395
187	0.59883721	0.54053679	0.96117658
188	0.38953488	0.58795077	0.96110982
189	0.85465116	0.2222193	0.96159593
190	0.41860465	0.51762137	0.96120818
191	0.52325581	0.69569114	0.96094343
192	0.6744186	0.72107371	0.96093874
193	0.3255814	0.1799723	0.96168198
194	0.24418605	0.19193218	0.96164941
195	0.38953488	0.00302212	0.96194183
196	0.93023256	0.3040641	0.96153405
197	0.43023256	0.50458415	0.96122637
198	0.79069767	0.5557968	0.96120868
199	0.64534884	0.52661749	0.9612023
200	0.68604651	0.49490684	0.96125286

Acknowledgements
Not applicable.
Funding
Not applicable.
Ethics Declarations
Competing Interests
No, there are no conflicting interests.
Rights and Permissions
Open Access