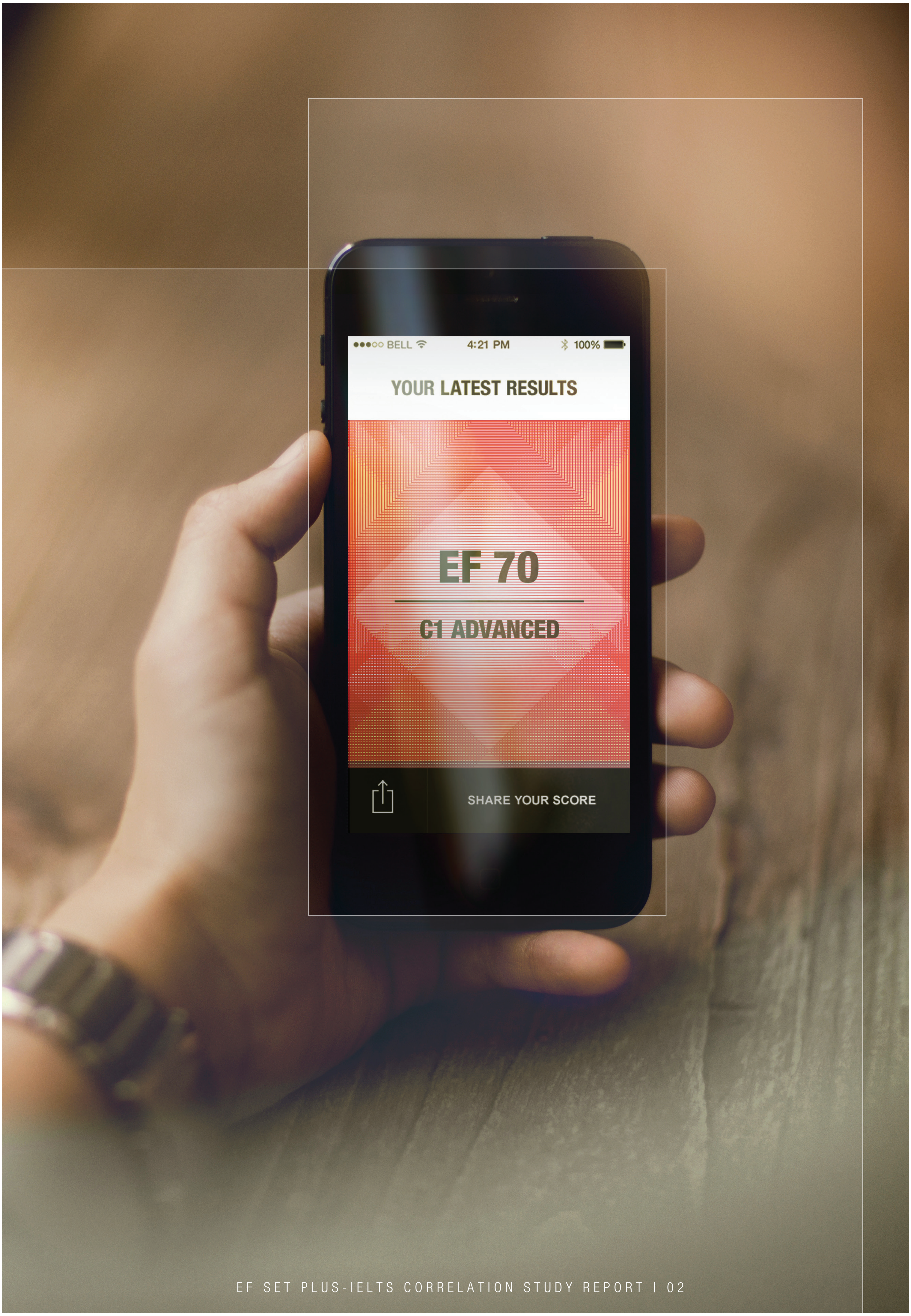
EF SET

EF STANDARD ENGLISH TEST

EFSET PLUS-IELTS CORRELATION STUDY REPORT

SEPTEMBER 2015

SIGN IN
EF STANDARD ENGLISH TEST
EF 50
B1 INTERMEDIATE
Your EFSET PLUS score out of 100 is
EF 50
Congratulations on completing both the EFSET Reading and Listening test.
in Add to profile
Not happy with your results? Take the test again
YOUR RESULTS IN DETAIL
EFSET
Reading
EF 1
EF 1
EF 50
EF 49
EF 100
EF 100


●●●○○ BELL
4:21 PM
100%
YOUR LATEST RESULTS
EF 70
C1 ADVANCED
SHARE YOUR SCORE

EXTERNAL VALIDITY OF EF SET PLUS AND IELTS SCORES

Richard M. Luecht

Abstract

This study was carried out to explore the statistical association between EF SET PLUS™ and IELTS scores. Four-hundred and six examinees were included in the study. There were reasonably strong positive correlations between EF SET PLUS™ and reported IELTS reading and listening scores. Average performance of the examinees within the IELTS performance categories further indicated a very solid linear pattern of association between reading and listening scores on the two examinations. Scale reliability was demonstrated to be very good for the EF SET scores given the adaptive nature of the test. These results suggest that EF SET PLUS examinations are very reliable, demonstrate expected positive associations with IELTS scores, but also may be getting at a somewhat unique set of English language reading and listening.

TABLE OF CONTENTS

INTRODUCTION

This report describes a validation study carried out in summer 2014 for the new EF Standard English Test (EF SET PLUS™). The purpose of this report is to present empirical, external validity evidence regarding the relationship between EF SET PLUS proficiency scores and reported International English Language Testing System (IELTS) scores. The IELTS modules measure reading, listening, writing and speaking skills (general training versus academic varieties). As a system, IELTS is widely recognized as one of the premier tests of English language proficiency in the world. IELTS test scores are reported using a nine-band scale, where Band 1 refers to individuals who essentially exhibit no language proficiency beyond possibly several isolated words and Band 9 denotes an expert user who demonstrates complete and fluent operational command of the language across various language contexts.

EF SET and EF SET PLUS are free, online tests designed to provide separate measures of English language reading and listening proficiency. The tests are professionally developed and administered online with a computer interface that is standardized across computer platforms. The reading and listening sections of both tests are adaptively tailored to each examinee's proficiency, providing an efficient and accurate way of assessing language skills. As an interpretive aid, performance scores on EF SET and EF SET PLUS are directly aligned with six levels (A1 to C2) of the Council of Europe's Common European Framework of Reference for languages. For more information on EF SET's score scale, visit: www.efset.org/english-score/cefr.

In this study, an international sample of non-native English language learners were recruited and screened over a period of 4 months. Four-hundred and six examinees who met the study eligibility requirements were administered both an EF SET PLUS reading and listening test. As part of the eligibility requirements, the examinees were required to upload a digital copy of their IELTS test score report. Their scores on EF SET PLUS and their reported IELTS scores were then analyzed to investigate the degree of statistical correspondence between the tests. The study results confirm that the EF SET PLUS scores are quite reliable across the corresponding reading and listening score scales and that there is reasonable statistical correspondence with IELTS scores. Overall, this provides important evidence about the validity of the EF SET PLUS reading and listening scores.

METHODS

This section of the paper describes the EF SET PLUS examinations and scoring process. It also describes the participant sample used for the validation study. Analysis and results are covered in the subsequent section.

Description of the EF SET PLUS Tests and Score Scales

Separate reading and listening test forms which were statistically equivalent to the EF SET PLUS were used for this study. This was to ensure that there was no learning effect of the publicly available EF SET PLUS. The EF SET tests employ various types of selected-response item types, including multiple-selection items. A set of items is associated with a specific reading or listening stimulus to comprise a *task*. In turn, one or more tasks are assembled as a unit to prescribed statistical and content specifications; these are called *modules*. The modules can vary in length, depending on the number of items associated with each task. Because of the extra time needed to listen to the task-based aural stimuli, the listening modules tend to have slightly fewer items than the reading modules. In general, the reading modules for this study had from 16 to 24 items. The listening modules each had between 12 and 18 items. In aggregate, each examinee was administered a three-stage test consisting of one module per stage.

The actual test forms for EF SET and EF SET PLUS are administered using an adaptive framework known as *computerized adaptive multistage testing* or ca-MST (Luecht & Nungester, 1998; Luecht, 2000; Zenisky, Hambleton & Luecht, 2010; Luecht, 2014a). Ca-MST is a psychometrically powerful and flexible test design that provides each examinee with a test form customized for his or her demonstrated level of language proficiency. For this study, each EF SET examinee was administered a three-stage 1-3-4 ca-MST *panel* with three levels of difficulty at stage 2 and four levels of difficulty at stage 3 as shown in Figure 1. The panels are self-adapting. Once assigned to an examinee, each panel has internal routing instructions that create a statistically optimal pathway for that examinee through the panel. The statistical optimization of the routing maximizes the precision of every examinee's final score.

As Figure 1 demonstrates, all examinees assigned a particular panel start with the same module at Stage 1 (M1, a medium difficulty module). Based on their performance on the M1 module, they are then routed to either module E2, M2 or D2 at Stage 2. The panel routes the lowest performing examinees to E2 and the highest performing examinees to D2. All others are routed to M2. Combining performance from both Stages 1 and 2, each examinee is then routed to module VE3, ME3, MD3, or VD3 for the final stage of testing. This type of *adaptive* routing has been demonstrated to significantly improve the precision of the final score estimates compared to a fixed (non-adaptive) test form of comparable length (Luecht & Nungester, 1998). The cut scores used for routing are established when the panel is constructed to statistically optimize the precision of each pathway through the panel.

METHODS

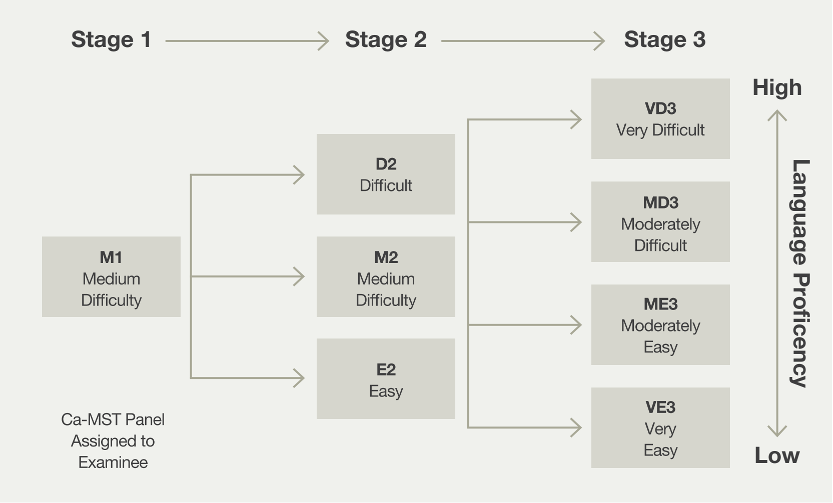


Figure1. An Example of a 1-3-4 ca-MST Panel

All EF SET items are statistically calibrated to the EF reading and listening score scales. The calibration process employs *item response theory* (IRT) to determine the difficulty of each item relative to all other items. The IRT-calibrated items and tasks for the reading and listening panels used in this study were previously administered to large samples of EF examinees and calibrated using the Rasch calibration software program WINSTEPS (Linacre, 2013). This software is used world-wide for IRT calibrations. The IRT model used for the calibrations is known as the *partial-credit model* or PCM (Wright & Masters, 1982; Masters, 2010). The partial-credit model can be written as follows:

$$P\left(x = X \middle| \theta; b_i, \mathbf{d}_i\right) \equiv P_{ix}(\theta) = \frac{\exp\left[\sum_{k=0}^{x} \theta - \left(b_i + d_{ik}\right)\right]}{\sum_{j=0}^{m} \exp\left[\sum_{k=0}^{j} \theta - \left(b_i + d_{ik}\right)\right]}$$

Equation 1

where θ is the examinee's proficiency score, b_i denotes an item difficulty or location for item i, and d_{ik} denotes two or more threshold parameters associated with separations of the category points for items that use three or more score points (k=0,…,xi). All reading items and tasks for the EF Standard Setting (conducted in 2014 - see Technical Background Report for more details) were calibrated to one IRT scale, θ_R. All listening items and tasks were calibrated to another IRT scale, θ_L.

Using the calibrated PCM items and tasks, a language proficiency score on either the θ_R or θ_L scale can be readily estimated regardless of whether a particular examinee follows an easier or more difficult route through the panel (i.e. the routes or pathways denoted by the arrows in Figure 1). The differences in module difficulty within each panel are automatically managed by a well-established IRT scoring process known as *maximum likelihood estimation* (MLE).

METHODS

Type of Language User	Level	Code	Description
Basic	Beginner	A1	Understands familiar everyday words, expressions and very basic phrases aimed at the satisfaction of needs of a concrete type
	Elementary	A2	Understands sentences and frequently used expressions (e.g. personal and family information, shopping, local geography, employment)
Independent	Intermediate	B1	Understand the main points of clear, standard input on familiar matters regularly encountered in work, school, leisure, etc.
	Upper Intermediate	B2	Understands main ideas of complex text or speech on both concrete and abstract topics, including technical discussions in field of specialisation
Proficient	Advanced	C1	Understands a wide range of demanding, longer texts, and recognises implicit or nuanced meanings
	Mastery	C2	Understands with ease virtually every form of material read, including abstract or linguistically complex text such as manuals, specialised articles and literary works, and any kind of spoken language, including live broadcasts delivered at native speed

Figure 2. Six CEFR Language Proficiency Levels. Visit: www.efset.org/english-score/cefr

The content validity of the EF SET ca-MST modules and panels is well-established and follows state-of-the-art task and test design principles established by world experts on language and adaptive assessment design. The EF SET Technical Background Report (EF SET, 2014) provides a comprehensive overview of the test development process. It should be noted that the EF SET and EF SET PLUS alignment to the CEFR levels was established through a formal standard-setting process (Luecht, 2014b; EF SET, 2014).

Validation Study Sample

Examinees were recruited to participate in the online EF validation study. The primary eligibility requirements were: (a) having a valid email address and (b) being able to provide by digital upload an official IELTS score report showing recent reading and listening scores. "Recent" was operationally defined as having taken the IELTS modules within the past 18 months. All potential examinees completed a brief survey to establish their eligibility and then uploaded a digital copy of their IELTS score report. Only eligible candidates were allowed to proceed to the next phase and actually take the EFSET PLUS reading and listening forms. The validation study testing was carried out during the summer of 2014.

The examinees were administered and completed both an EF SET PLUS reading and listening panel. Every examinee that completed both EF SET PLUS panels within the testing window and whose performance demonstrated reasonable effort[1] was compensated with a voucher for £50. Ultimately, there were 406 participants with complete data.

[1] *Examinees who left entire modules blank, took unreasonably little time to complete the tests, or who otherwise exhibited an obvious lack of effort were excluded. The application process carefully explained the study participation "rules" to each examinee.*

METHODS

Demographically, the sample was comprised of 199 (48.9%) women and 207 (50.9%) men. Ages of the examinees ranged from 16 to 38 years; the average age was 22.8 with a standard deviation of 3.2 years. The majority of the study participants (123 or 30.3%) listed their nationality as Vietnamese. Other relatively high-percentage nationalities listed were Hong Kong (14.3%), India (13.3%), China (10.8%), and Brazil (5.7%). The remaining 104 participants (25.6%) were from other Asian countries, as well as various European, African and South American nations.

Education and English as a second language (ESL) experience of the sample are jointly summarized in Table 1. In general, the sample was comprised primarily of well-educated, young Asian/Asian Indian adults with somewhat extensive ESL experience. The gender mix was about equal.

Table 1. Language Experience and Educational Information for the Sample (N=406)

Language Experience	**Frequency**	**Percent**
Less than 1 yr.	13	3.2%
1-3 years	36	8.9%
4-6 years	65	16.0%
7-9 years	79	19.5%
More than 9 yrs.	213	52.5%
Degree	**Frequency**	**Percent**
Did not finish high/secondary school	5	1.2%
High/secondary school	85	20.9%
Further education: some college	35	8.6%
Bachelor’s degree	239	58.9%
Master’s degree	32	7.9%
Other degrees	10	2.5%
Major Area of Study	**Frequency**	**Percent**
Sciences, engineering or medicine	95	23.4%
Business	78	19.2%
Politics & social science	57	14.0%
Other	55	13.5%
Languages	43	10.6%
Mathematics	29	7.1%
Art and design	22	5.4%
Arts & science, humanities	16	3.9%
Missing Responses	**11**	**2.7%**

IELTS reading, listening and combined performance band scores for the 406 sample participants are summarized in Table 2. As noted earlier, none of the volunteer examinees had IELTS reading or listening scores below 4.0 on the IELTS scale. This was an unexpected sampling limitation.

METHODS

In terms of the present study, it may have introduced some statistical variance restriction of the IELTS score distributions—an issue further compounded by the potential ranges of different reading and listening proficiencies collapsed into each IELTS score band. That is not to imply any limitations on the utility of the IELTS scores. Rather, as demonstrated further on, this sampling limitation, combined with the variance restriction of the IELTS score bands may have statistically suppressed the magnitude of the potential correlations between IELTS and EF SET PLUS.

Table 2. Summary of IELTS Performance (N=406)

Statistics	Reported IELTS Score		
	Reading	Listening	Total
Mean	7.151	7.069	6.773
Std. Deviation	1.181	1.179	0.819
Minimum	4.0	4.5	4.5
Maximum	9.0	9.0	9.0

The EF SET PLUS reading and listening records were matched and then rescored using the IRT-based scoring tables for the two panels as a score-confirmation step. All EF SET PLUS reading and listening scores were reconfirmed to a high degree of estimation precision. The descriptive statistics on the key proficiency-related variables, estimated reliability coefficients, correlations (observed and disattentuated), and some auxiliary performance comparisons between their EF SET PLUS listening and reading scores and IELTS scores are presented in the next section.

ANALYSIS AND RESULTS

Descriptive statistics for the EF SET PLUS scores are shown in Table 3 for the 406 examinees that participated in this study. The variables "Reading θ_R" and "Listening θ_L" are the two EF SET PLUS proficiency scores. By IRT convention, proficiency scores estimates are often denoted by the Greek letter θ ("theta"). Note that in practice, these IRT scores are rescaled to a more convenient and somewhat more interpretable set of scale values (0 to 100). For various technical statistical reasons, that rescaling was not applied for purposes of this study. Here, it is sufficient to note that the score estimates of θ_R and θ_L can be negative or positive[2], where higher positive numbers denote better language proficiency as measured by the EF SET ca-MST panels. The "SE(θ_R)" and "SE(θ_L)" variables denote the IRT standard errors for the corresponding estimated EF SET θ scores. All test scores contain some degree of error. The computed SE(θ) values merely help to quantify the magnitude of the score estimation errors. In general, smaller errors of estimate denote more precise scores. The average standard errors were used to compute what are termed marginal reliability coefficients for purposes of computing disattentuated correlations (i.e. correlations corrected for unreliability of the scores).

Table 3. EF SET Descriptive Statistics for EF SET PLUS IRT Proficiency Scores (N=406)

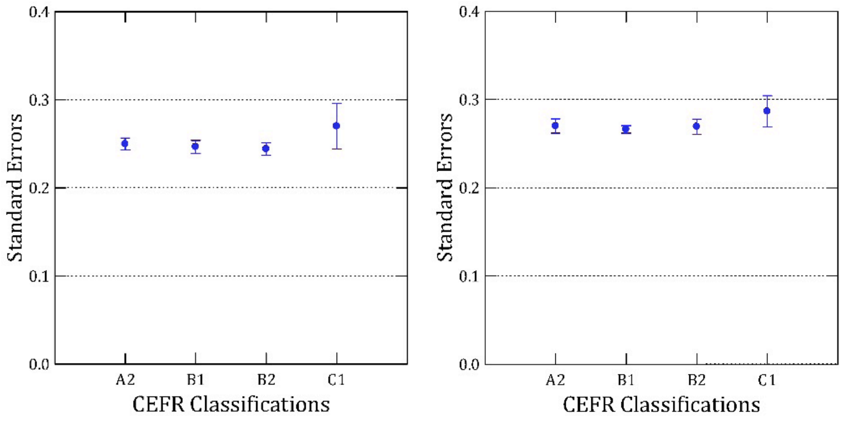
Statistics	Reading θ_R	SE(θ_R)	Listening θ_L	SE(θ_L)
Mean	0.773	0.246	1.032	0.271
Std. Deviation	0.795	0.006	0.775	0.008
Minimum	-1.684	0.240	-0.815	0.256
Maximum	3.296	0.300	2.884	0.316

An important benefit of the multistage test design used for EF SET PLUS is evident when considering the magnitude of the standard errors. Achievement or placement tests that employ test forms comprised of a fixed set of items—that is, *non*-adaptive tests—typically have smaller standard errors of estimate near the population mean or near a particular cut score (e.g. for placement) to ensure maximally precise score estimates at the point along the score scale. However, there is a trade-off. The same fixed set of items will tend to have large errors of estimate nearer the tails. The adaptive EF SET panels (see Figure 1) are specifically designed to provide somewhat more uniform precision ACROSS the entire the score scale—providing the best possible precision of the estimates of θ_R and θ_L. Figure 3 displays two summary plots of the average standard errors, SE(θ), by CEFR[3] level for the 406 examinees' reading (left) and listening (right) panels. The error bands are 95% confidence bands on the distribution of standard errors.

[2] *The IRT calibration software, WINSTEPS (Linacre, 2013) scales the EF SET tests' item banks to have a mean item difficulty parameter estimates (scale locations centers) of zero. The examinees scores are not centered or otherwise standardized to zero and should not be interpreted as "z-scores" or other normal-curve equivalents.*

[3] *The CEFR "C2" level (see Figure 2) was not applied for this validation study version of EF SET. Following standard setting in 2014 (see EF SET, 2014), the C2 level classification has been added.*

ANALYSIS AND RESULTS



(Reading) **(Listening)**

Figure 3. Distributions of IRT Standard Errors of Estimate by CEFR level

Clearly, Figure 3 confirms that the IRT-based standard errors of estimate are relatively small across the entire score scale. The nominally higher errors within the "C1" level for both reading and listening reflect limitations in the current item banks near the highest CEFR levels. Even those standard errors will be reduced by the panel assembly process as EF SET matures and expands its banks of reading and listening tasks.

Referring to the means in Table 3, the 406 participants included in this study are substantially more proficient than the typical EF examinee. This sample was normatively compared to extremely large samples of examinees (N > 37,000) who took EF SET and EF SET PLUS reading and listening forms over the past two years. This validation sample placed on average near the 79th percentile for reading and near the 85th percentile for listening. Similar results were obtained by comparing the participants' reported IELTS scores to the corresponding published large-sample 2013 IELTS results (www.ielts.org). This finding is consistent with the previously noted sampling limitation where all of study participants have IELTS scores of 4.0 or higher. In a practical sense, that same sampling limitation rather naturally yielded a group of competent and motivated examinees. The corresponding sampling trade-off is that some censoring or restriction of variances of the scores probably suppressed the correlations between the studied variables. That issue is addressed next.

Pearson product-moment correlations were computed between four score variables: (i) IELTS reading scores; (ii) IELTS listening scores; (iii) EF SET PLUS IRT score estimates for θ_R (reading); and (iv) EF SET PLUS IRT score estimates for θ_L (listening). Correlations denote the degree of statistical linear association between pairs of variables. Values near 1.0 indicate an almost perfect linear relationship between the variable pair. Values near zero indicate almost no linear association and values near --1.0 indicate a nearly perfect inverse relationship (i.e. increasing values on one variable are strongly associated with decreasing values on the second variable). Validity studies such as this often result in "moderate", positive correlations (e.g. 0.4 to 0.7). The computed correlations between the scores for the 406 study participants are shown in the lower "triangle" of the correlation matrix in Figure 4 (i.e. in the unshaded cells below the diagonal of the matrix).

ANALYSIS AND RESULTS

The most relevant correlations from a validity perspective are the two correlations between the IELTS reading and estimated EF SET PLUS θ_R scores (0.52) and between the IELTS listening and estimated EF SET PLUS θ_L scores (also 0.52). They suggest reasonable positive alignment between IELTS and EF SET PLUS scores. However, two factors can conspire to reduce the magnitude of a correlation coefficient between any two variables. One factor is the amount of measurement error present in the scores—that is, the reliability of the observed or estimated scores. The second factor is a sampling consideration—namely any censoring or restriction of the variance of the scores. Both factors are considered below.

Score Variables	IELTS Reading	IELTS Listening	IRT θ_R Estimates	IRT θ_L Estimates
IELTS Reading	**0.91**	0.81	0.57	0.61
IELTS Listening	0.73	**0.90**	0.48	0.58
IRT θ_R Estimates	0.52	0.44	**0.90**	0.72
IRT θ_L Estimates	0.55	0.52	0.64	**0.88**

Table 4. Correlations Between IELTS and EF SET PLUS Scores
(Disattenuated Correlations Above the Diagonal, Reliability Coefficients on the Diagonal of the Matrix

Reliability coefficients indicate the magnitude of score precision near the mean of the score scale. The most commonly reported type of reliability coefficient is called Cronbach's α ("alpha"). Cronbach's α provides a somewhat conservative estimate of the average consistency of scores across the scale (Haertel, 2006). Values above 0.9 are considered to be very good. Because of the adaptive nature of the EF SET panels, traditional reliability coefficients can only be approximated using what is termed a marginal reliability coefficient. It is computed as

$$\rho^2\left(\hat{\theta},\theta\right)=1-\frac{E\left[\sigma^2\left(\hat{\theta}\middle|\theta\right)\right]}{\sigma^2\left(\hat{\theta}\right)}$$

Equation 2

where the numerator of the rightmost term is the average error variance of estimate (or square of the mean of the corresponding standard errors from Table 3) and the denominator of the rightmost term is the variance of the estimated IRT θ scores (Lord & Novick, 1968). Provided that the data fit the IRT model used for calibration and scoring—the PCM in the case of EF SET and EF SET PLUS—this marginal reliability is usually quite comparable to Cronbach's α coefficient. The reliability coefficients reported for IELTS are α coefficients (IELTS, 2013).

ANALYSIS AND RESULTS

The disattenuated correlations in the upper section of the matrix in Table 4 estimate the *true-score correlations*—that is, the statistical relationships between the four scores if measurement errors were eliminated all-together. The disattenuated correlations are computed by simply dividing the Pearson product-moment correlation between each pair of score variables by the square root of the product of the reliability coefficients for each score (Haertel, 2006, p.85). Because the reliability coefficients for the IELTS and EF SET scores are all relatively high, the magnitude of increase in the true-score [disattentuated] correlations is not overtly larger than the observed correlations in the lower section of the matrix. It should be nonetheless be apparent that the EF SET PLUS scores are at a comparable level of reliability to the IELTS scores[4]. Also, as noted earlier the errors of estimate are fairly uniform across the EF SET PLUS score scales (see Figure 3).

Figures 5 and 6 respectively show the scatter plots for the observed reading and listening scores. The IELTS reading and listening scores are plotted relative to the horizontal axis in each plot. The EF SET PLUS scores are plotted relative to the vertical axis. The best-fitting regression line is also shown for each pair of score variables. It should be apparent that the EF SET scores are substantially more variable than the reported IELTS scores.

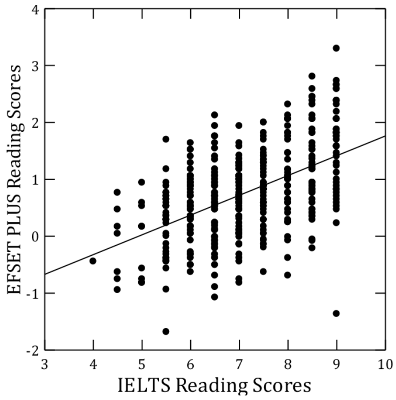


Figure 5. Scatterplot of EF SET PLUS Reading Scores (Vertical Axisby IELTS Reading Scores

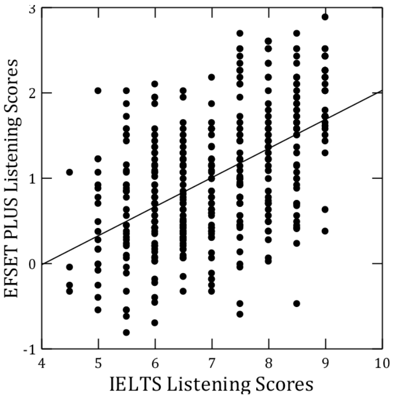


Figure 6. Scatterplot of EF SET PLUS Listening Scores (Vertical Axisby IELTS Listening Scores

ANALYSIS AND RESULTS

Dealing with the inherent variance restriction of either or both the IELTS and the EF SET PLUS scores presents a more complicated challenge. Any restriction of the sampling variance of any distribution of scores will tend to reduce the correlation of those scores with any other variable. As noted earlier, the variances of all scores for the 406 participants in this study were restricted because all of the volunteers had IELTS scores or 4.0 or higher. That characteristic of the volunteer sample therefore restricted the sampling of variances of all of the scores so that a sufficiently able and motivated sample could be ensured within the practical timeline and resource allocation limitations of the study. Also, the IELTS score bands may have induced an additional type of variance restriction due to what is essentially “rounding” within those proficiency groupings. When both types of variance restriction are considered, it should not be surprising that the correlations between EF SET PLUS and IELTS scores are moderate.

However, there is another way to evaluate the association between EF SET PLUS and IELTS. Figure 7 provides a very insightful plot of the nature of the association between IELTS and EF SET PLUS when the variation within each of the IELTS score bands is essentially removed. For purposes of this analysis, each participant’s reported IELTS score was rounded to the nearest integer value. The dot symbols in Figure 7 represent the mean or average performance on the EF SET PLUS reading and listening tests for all examinees with combined IELTS scores of 5, 6, 7, 8 or 9. (Note that none of the participants had combined IELTS scores of 4.) The error bars for each plotted reading or listening score mean reflect the empirical sampling error of the EF SET PLUS means within that corresponding IELTS band. The noticeably wider error bars for examinees in the more extreme IELTS bands (5 and 9) are due to having smaller numbers of examinees in those bands.

[4] Note that the EF SET PLUS reliability results reported here are specific to the score variances for this sample of 406 participants. If the large-sample EF SET PLUS 2013-14 variances were used—that is, over 37,000 examinees not restricted by the eligibility requirement of an IELTS score of 4.0 or higher—the marginal reliabilities for reading and listening would be 0.95 and 0.94, respectively.

ANALYSIS AND RESULTS

The visually strong and positive linear trend evident in Figure 7 provides rather compelling evidence of a solid correspondence between reported IELTS scores and EF SET PLUS scores, on average. The associated polynomial trend lines, using the rounded IELTS total score groups as the independent variable were also confirmed using analysis of variance (F1,401= 54.19, p(F)<0.001 for reading and F1,401= 59.679, p(F)<0.001 for listening). A polynomial trend analysis sequentially fits incrementally more complex patterns to model the change between two variables. The simplest trend, which is a linear trend that shows more or less a consistent increase in the independent variable—in this case, the IELTS total scores—is usually preferred to a more complex trend. Here, the polynomial trend analysis confirms the likelihood of very strong linear trend between the EF SET PLUS reading and listening test scores and IELTS composite score grouping (again, rounded to the nearest integer). The interpretation is that as we move up the IELTS scale, EF SET PLUS scores likewise increase, on average. Any non-linear differences between IELTS and EF SET PLUS performance therefore seem to occur within the broader IELTS categories.

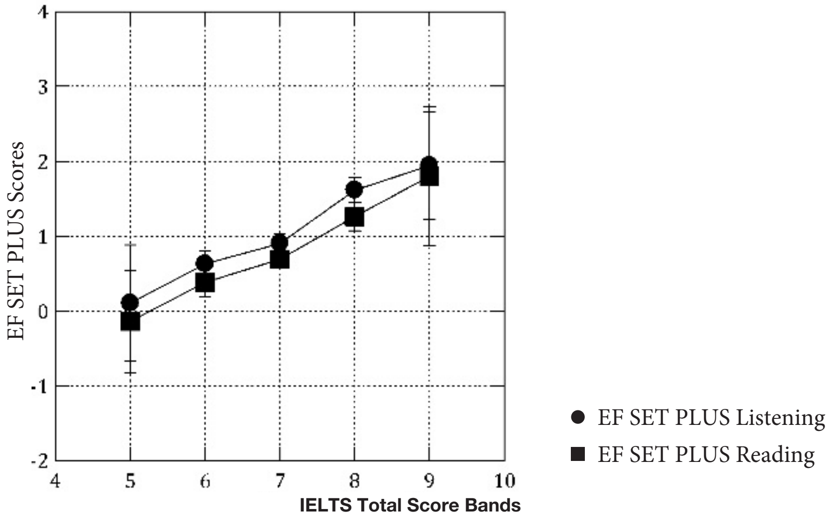


Figure 7. EF SET PLUS Reading and Listening Scores Grouped by IELTS Bands (N=406)

It is important to understand that there is no absolute "gold standard" for language assessments. IELTS is a mature test; EF SET PLUS is relatively new. The EF SET PLUS scores are reported on a scale with more detail than IELTS. The published data on IELTS and results presented here and elsewhere regarding EF SET PLUS seem to confirm that both tests yield highly reliable scores of reading and listening. The demonstration of only moderate positive correlations between IELTS and EF SET PLUS scores may be because each test is getting at different traits—which is always possible—or due largely due to rather subtle variance restriction issues in this study. The findings from this study merely suggest the need for additional and ongoing validity evidence gathering as a responsible measurement practice clearly supported by the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014) and the *International Test Commission Guidelines* (ITC, 2008).

DISCUSSION

Validity studies of this type are interesting but also challenging. Obviously, a relatively new English language testing like EF SET PLUS would like to show strong, positive correspondence with an established program like IELTS. At the same time, if the EF SET PLUS scores had demonstrated too high of a correlation with IELTS, it might question whether there is anything useful to be gained by having the new testing program.

Despite the potential complications of subtle variance restrictions described in this report, these results are encouraging from both reliability and validity perspectives. They suggest that the EF SET examinations are very reliable and also demonstrate expected positive associations with IELTS. At the same time, the EF SET examinations may indeed be getting at something “different” than IELTS by design. The moderately small sample size and other factors recommend over-interpretation of this results, or conclusive claims as to whether IELTS or EF SET tests are measuring “truth” in English language reading and listening proficiency. That claim is impossible to substantiate.

A final point of comment concerns establishing a concordance relationship between IELTS and EF SET PLUS. At present, there is no direct alignment or concordance between IELTS scores and EF SET scores. In fact, given the only moderate positive correlations reported between EF SET PLUS and IELTS (Figure 4), any attempt to establish direct concordance between those two score scales is probably not psychometrically appropriate[5].

[5] *Score or classification concordance tables are sometimes created to show the approximate equivalence of scores on two scales that measure similar—but not necessarily the same—constructs. An example would be the well-known concordance between college admissions tests like the ACT Assessment (Act, Inc.) and the Scholastic Aptitude Test (SAT) in the US. Basing concordance on tests with only moderate correlations can lead to misuse of the scores if some users consider the scores to actually be exchangeable. Concorded scores are not exchangeable (Kolen & Brennan, 2014). A policy decision was therefore made NOT to provide concordanc information between IELTS and EF SET examinations until additional evidence is gathered.*

REFERENCES

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014).
Standards for Educational and Psychological Testing. Washington, DC: AERA.

EF. (2014). *EF SET Technical Background Report*. London, U.K: www.efset.org.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.).
Educational Measurement, 4th Edition, pp. 65-110.
Washington, DC: American Council on Education/Praeger Publishers.

International English Language Testing System. (2013). *IELTS | Researchers - Test performance 2013*. Author: International English Language Testing System, www.ielts.org.

International Test Commission. (2008). International Test Commission Guidelines. Website: www.intestcom.org/guidelines/

Kolen, M. J. & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices, 2nd edition*. New York: Springer.

Linacre, M. (2013). *WINSTEPS Rasch Measurement* Version 3.74).
[Computer program]. Author: www.winsteps.com.

Lord F.M. & Novick, M. (1968). *Statistical theories of mental test scores*.
Reading, MA: Addison-Wesley.

Luecht, R. M. (2000, April). *Implementing the computer-adaptive sequential testing (CAST) framework to mass produce high quality computer-adaptive and mastery tests*. Symposium paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Luecht, R. M. (2014a). Computerized adaptive multistage design considerations and operational issues (pp. 69-83). In D. Yan, A. A. von Davier & C. Lewis (Eds.)
Computerized Multistage Testing: Theory and Applications. New York: Taylor-Francis.

Luecht, R. M. (2014b). Setting Standards for the *EF SET Reading and Listening Assessments*. EF SET Technical Background Report commissioned by EF.
London, U.K: Author.

Luecht, R. M. & Nungester, R. J. (1998). Some practical applications of computerized adaptive sequential testing. Journal of Educational Measurement, 35, 229-249.

Zenisky, A.; Hambleton, R. J.; & Luecht, R. M. (2010). Multistage Testing: Issues, Designs, and Research. In W. J. van der Linden and C. E. W. Glas (Eds). Elements of Adaptive Testing, pp. 355-372. New York: Springer.

ABOUT THE AUTHOR

Richard M. Luecht, PhD, Professor of Educational Research Methodology at the University of North Carolina at Greensboro (UNCG), is the chief psychometric consultant for the EF SET team. He is also a Senior Research Scientist with the Center for Assessment Research and Technology, a not-for-profit psychometric services division of the Center for Credentialing and Education, Greensboro, NC.

Ric has published numerous articles and book chapters on technical measurement issues. He has been a technical consultant and advisor for many state department of education testing agencies and large-scale testing organizations, including New York, Pennsylvania, Delaware, Georgia, North Carolina, South Carolina, New Jersey, Puerto Rico, The College Board, Educational Testing Service, HUMRRO, the Partnership for Assessment of Readiness for College and Career (PARCC), the National Center and State Collaborative (NCSC), the American Institute of Certified Public Accountants, the National Board on Professional Teaching Standards, Cisco Corporation, the Defense Language Institute, the National Commission on the Certification of Physicians Assistants, and Education First (EF SET).

He has been an active participant previously at the National Council of Measurement in Education (NCME), American Educational Research Association (AERA) and Association of Test Publishers (ATP)meetings, teaching workshops and giving presentations on topics such as assessment engineering and principled assessment design, computer-based testing, multistage testing design and implementation, standard setting, automated test assembly, IRT calibration, scale maintenance and scoring, designing complex performance assessments, diagnostic testing, multidimensional IRT, and language testing.

Before joining UNCG, Ric was the Director for Computerized Adaptive Testing Research and Senior Psychometrician at the National Board of Medical Examiners where he oversaw psychometric processing for the United States Medical Licensing Examination (USMLE) Step and numerous subject examinations, as well being instrumental in the design of systems and technologies for the migration of the United States Medical Licensing Examination programs to computerized delivery.
He has also designed software systems and algorithms for large-scale automated test assembly and devised a computerized adaptive multistage testing implementation framework that is used by a number of large-scale testing programs. His most recent work involves the development of a comprehensive framework and associated methodologies for a new approach to large-scale formative assessment design and implementation called assessment engineering (AE).

WWW.EFSET.ORG

Education First