A Reliability Study of the G-TELP Test, Level 2

Laura MacGregor

Introduction

This paper seeks to investigate the reliability of the G-TELP test, Level 2, Form 211, which was used to select eligible students for the Intensive English Course at Gakushuin University in April, 2002. Of the 237 students who took the test, 46 were admitted to the programme based on a cutoff score determined after the test.Using classical descriptive and item analyses, the Gakushuin data are examined here to assess how the items are performing, to determine the test's reliability, and finally, to assess the G-TELP's suitability as a placement instrument.

Overview of the Test

General information

The G-TELP is a 5-level criterion-referenced English proficiency test developed and maintained by San Diego State University. It is currently being used by 71 institutions (high schools, junior colleges, universities) in Japan, primarily as a placement instrument. Level 2, the subject of this study, is an 80-item test consisting of three sections: grammar (26 items), listening (26 items), and reading and vocabulary (28 items). Each of the section scores are reported out of 100 for a total score of 300. To be ranked at Level 2, test-takers must score 75 or higher in all three sections.

According to the test distributors, the G-TELP, which tests "authentic and modified English in normal communication," is designed to "measure language proficiency in two broad areas: functional ability and grammatical ability" (G-TELP Japan office, personal communication, April, 2002).

The general description for Level 2 states that:

This level assesses the ability of an examinee to use the language outside of classroom situations. This examinee is able to cope with some authentic English and has experienced contact with native speakers. Although his/her learning of the language has been classroom-based, the examinee is able to communicate with a native speaker within a wide range of tasks. (San Diego State University, 1998).

Item types

Since the actual test forms are reused by G-TELP and are therefore confidential, the contents of Form 211 cannot be disclosed. However, a study guide published by Kinseido (Morita, 1998) gives the following general information about the Level 2 test.

Level 2 addresses a finite number of skills, which are identified as follows according to test section. First, the grammar section includes items which address four grammar points: (i) participials (7 items); (ii) present unreal (7 items); (iii) past perfect tense (6 items); and (iv) future progressive tense (6 items). Four listening tasks address comprehension of: (i) a narrative; (ii)the description of a process; and conversations involving (iii) persuasion; and (iv) negotiation. The length of each listening passage in the sample guidebook averaged 2 minutes 51 seconds (ranging from 2 minutes 13 seconds to 3 minutes 35 seconds each). The reading and vocabulary section also has four sections which present four different text types in passages of 300-400 words: (i) a historical account; (ii) an article on a technical or social topic; (iii) an encyclopedia entry; and (iv) a business letter with descriptive and persuasive elements. There are three question types in this section: (i) reading comprehension (literal skills); (ii) inferential skills; and (iii) vocabulary knowledge.

Method

The data were first analyzed to generate descriptive statistics illustrating features of central tendency and dispersion. To determine how well the individual items performed on this administration of the test, the item facility (IF) and item discrimination (ID) values were calculated for each section. Finally, the reliability coefficient for the test was generated using the Kuder-Richardson 20 (K-R20) formula. The software used was Microsoft Excel for Mac (Microsoft, 2001).

Results

(a) Descriptive statistics

Descriptions of the central tendency (mean and median) are reported in Table 1 to identify the typical, or common scores for the test and for each section. Dispersion figures, which indicate how individual scores disperse around the central tendency are also reported (Table 1).

Table 1: Descriptive Statistics (N＝237)

	Central Tendency		Dispersion		
k	**Mean (%)**	**Median (%)**	**Low-High**	**Range**	**S**
Grammar					
26	60.9	62	15-100	86	15.417
Listening					
26	57.1	58	15-100	86	17.250
Reading and Vocabulary					
28	63.6	64	11-100	90	16.959
Total					
80	181.6	183	61-281	221	39.255

Key

N = number of test-takers　　k = number of test items

Median = point below which 50% of the scores fall and above which 50% of the scores fall

Range = high score - low score + 1　　S = standard deviation

The median scores are nearly the same as the mean scores, suggesting a nearly normal distribution. Central tendency alone however, does not give a complete picture of distribution as it works together with dispersion.

Dispersion indicates how individual scores disperse around the central tendency, and shows how widely the scores are spread out. The figures here (Table 1) indicate that a normal distribution is not likely for the grammar section or the reading and vocabulary section, and therefore for the whole test, since the distance between the low scores and the mean is different from the distance between the high scores and the mean. A look at the histograms for the three sections and for the total test (Figures 1-4 in the Appendix; score report data from G-TELP Japan office, April 8, 2002) confirms that no part of the test produces a normal distribution.

(b) Item facility and item discrimination

Next, item facility (IF) and item discrimination (ID) values are reported. Item facility represents the percentage of test-takers who answered the item correctly. Facility values are reported as factors of 1.00 and range from 0.00 to 1.00. Item facility shows how easy or difficult the test items are. An IF of .95 means that 95% of the test-takers got the item right, a very easy item. If the purpose of the test is to produce a wide range of scores, then items with IF values of around .50 should be used because they offer the greatest potential for variation among test-takers (Alderson, Clapham, and Wall, 1985, p. 81).Therefore, IF values can help control the

difficulty level of the test.

Item discrimination shows how well an item distinguishes between test-takers of different ability levels by examining the differences between the high scorers and the low scorers. ID values are calculated by first ranking test-takers according to their total scores, then comparing the proportion of correct answers (IF) in the top (or upper) group with those in the bottom (or lower)group. The formula for ID is:

$$ID = IF\ (upper) - IF\ (lower)$$

An ID of .40 or higher is considered to be acceptable; however, "there are no rules as to what [IDs] are acceptable, since the possibility of high [IDs] varies according to the test type and range of ability of the examinees" (Alderson, Clapham, and Wall, 1995, p. 82). Ebel (1979, quoted in Brown, 1996, p. 70) gives the following guide for items with IDs below .40:

.30- .39	Reasonably good but possibly subject to improvement
.20- .29	Marginal items, usually needing improvement
Below .19	Poor items, to be rejected or improved by revision

It is important to consider the relationship between IF and ID. As mentioned above, IFs close to .50, since they represent items of medium difficulty, are effective for the purpose of spreading the test scores out over a wide range. For the same reason, the IFs should be as close to .50 as possible to achieve high IDs, if the purpose of the test is to spread the test scores out as far as possible. A hypothetical table of IFs which are possible with certain ID values is presented as follows (Table 2, adapted from Alderson, Clapham, and Wall, 1995, p. 84):

Table 2: Maximum IDs for IF Values

IF	1.0	.93	.80	.70	.66	.50	.33	.30	.20	.06	0.0
max ID	0.0	.20	.60	.90	1.0	1.0	1.0	.90	.60	.30	0.0

Tables 3-5 present the IF data for the total population of test-takers for each of the three sections (using raw data provided by the G-TELP Tokyo office, April, 2002). Below the total IF figures, ID values are calculated using the upper 25% and lower 25% of the sample populations. The ID figures in bold, underlined type indicate items with ID values of below .19, which will be addressed in the following discussion section. However, since the actual contents of individual items are confidential, an examination of the possible factors which caused these items to perform poorly (i.e., choice of vocabulary or distractors, wording, and cultural appropriacy) cannot be made.

Table 3: IF and ID for Grammar Section (N＝26)

item #	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
total IF	**0.78**	**0.79**	**0.82**	**0.46**	**0.57**	**0.48**	**0.84**	**0.41**	**0.57**
IF upper 25%	0.78	0.94	0.87	0.70	0.76	0.58	0.90	0.43	0.70
IF lower 25%	0.68	0.71	0.77	0.24	0.42	0.37	0.75	0.39	0.54
ID	**0.10**	0.23	**0.10**	0.46	0.34	0.22	**0.15**	**0.04**	**0.15**
	easy		easy				easy	difficult	easy

item #	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*
total IF	**0.36**	**0.63**	**0.43**	**0.79**	**0.30**	**0.95**	**0.78**	**0.51**	**0.65**
IF upper 25%	0.53	0.71	0.54	0.85	0.42	0.97	0.84	0.73	0.72
IF lower 25%	0.19	0.54	0.33	0.70	0.19	0.92	0.73	0.28	0.54
ID	0.34	**0.16** easy	0.22	**0.15** easy	0.23	**0.05** easy	**0.10** easy	0.46	**0.18** easy

item #	*19*	*20*	*21*	*22*	*23*	*24*	*25*	*26*
total IF	**0.52**	**0.57**	**0.55**	**0.76**	**0.62**	**0.38**	**0.71**	**0.62**
IF upper 25%	0.65	0.77	0.67	0.85	0.80	0.58	0.90	0.72
IF lower 25%	0.37	0.39	0.32	0.61	0.43	0.18	0.54	0.52
ID	0.28	0.38	0.35	0.24	0.37	0.41	0.35	0.20

Table 4: IF and ID for Listening Section (N=26)

item #	*27*	*28*	*29*	*30*	*31*	*32*	*33*	*34*	*35*
total IF	**0.51**	**0.59**	**0.35**	**0.54**	**0.59**	**0.56**	**0.43**	**0.75**	**0.49**
IF upper 25%	0.67	0.75	0.62	0.70	0.76	0.78	0.66	0.85	0.73
IF lower 25%	0.41	0.41	0.16	0.38	0.41	0.37	0.23	0.57	0.32
ID	0.27	0.34	0.46	0.32	0.35	0.42	0.43	0.28	0.42

item #	*36*	*37*	*38*	*39*	*40*	*41*	*42*	*43*	*44*
IF upper 25%	**0.87**	**0.49**	**0.80**	**0.73**	**0.50**	**0.75**	**0.64**	**0.49**	**0.57**
IF lower 25%	0.96	0.77	0.92	0.85	0.71	0.95	0.78	0.51	0.75
IF lower 25%	0.77	0.23	0.66	0.62	0.37	0.66	0.41	0.41	0.35
ID	0.19	0.54	0.27	0.23	0.34	0.29	0.38	**0.10** difficult	0.39

item #	*45*	*46*	*47*	*48*	*49*	*50*	*51*	*52*
total IF	**0.34**	**0.34**	**0.46**	**0.56**	**0.69**	**0.39**	**0.85**	**0.57**
IF upper 25%	0.52	0.47	0.63	0.77	0.81	0.44	0.92	0.76
IF lower 25%	0.24	0.29	0.28	0.33	0.52	0.29	0.75	0.38
ID	0.28	**0.18** difficult	0.35	0.44	0.29	**0.15** difficult	**0.18** easy	0.38

Table 5: IF and ID for Reading and Vocabulary Section (N＝28)

item #	*53*	*54*	*55*	*56*	*57*	*58*	*59*	*60*	*61*	*62*
total IF	**0.93**	**0.79**	**0.53**	**0.89**	**0.77**	**0.89**	**0.54**	**0.72**	**0.64**	**0.73**
IF upper 25%	0.97	0.87	0.72	0.96	0.85	0.96	0.70	0.86	0.78	0.87
IF lower 25%	0.82	0.67	0.39	0.80	0.68	0.76	0.42	0.56	0.41	0.56
ID	**0.15** easy	0.20	0.33	**0.16** easy	**0.16** easy	0.20	0.28	0.30	0.38	0.32

item #	*63*	*64*	*65*	*66*	*67*	*68*	*69*	*70*	*71*	*72*
total IF	**0.53**	**0.84**	**0.68**	**0.68**	**0.74**	**0.41**	**0.70**	**0.45**	**0.35**	**0.34**
IF upper 25%	0.71	0.94	0.85	0.80	0.91	0.49	0.81	0.65	0.54	0.53
IF lower 25%	0.38	0.73	0.53	0.51	0.58	0.32	0.53	0.33	0.16	0.15
ID	0.33	0.20	0.32	0.29	0.33	**0.18** difficult	0.28	0.32	0.38	0.38

item #	*73*	*74*	*75*	*76*	*77*	*78*	*79*	*80*
total IF	**0.67**	**0.74**	**0.40**	**0.35**	**0.28**	**0.57**	**0.81**	**0.88**
IF upper 25%	0.94	0.82	0.57	0.48	0.46	0.82	0.95	1.00
IF lower 25%	0.41	0.65	0.22	0.24	0.14	0.37	0.68	0.71
ID	0.53	**0.18** easy	0.35	0.24	0.32	0.46	0.27	0.29

(c) Reliability

Reliability indicates how the scores on one test administration are likely to be very similar to those which would be obtained if it had been administered to the same students at another time. The greater the similarity between these two sets of scores, the higher the reliability of the test. While only one administration of the G-TELP occurred, the reliability can still be estimated, using the Kuder-Richardson 20 (K-R20) formula, the most accurate formula for measuring reliability. The formula for K-R20 is as follows:

$$\text{K-R20} = \frac{k}{k-1} \qquad \frac{\sum IV}{(1-S_t^2)}$$

Key

$\sum IV$ = sum of item variance for each item, where IV = IF(1-IF)

S_t^2 = variance for the whole test (standard deviation of the test scores squared)

The reliability of this form of G-TELP Level 2 was calculated to be .85, meaning that there is an 85% chance that a second administration of this test would produce very similar results. Unfortunately, there is no standard acceptable level of reliability, and opinions vary regarding minimum reliability coefficients according to the type of test. While there is no documentation on what the reliability of a criterion-referenced test like the G-TELP should be, Lado (quoted in Hughes, 1989, p. 20) suggested that good vocabulary, structure, and reading tests are usually in the .90 to .99 range.

Discussion

As mentioned above, the item types are finite, and the grammar items, in particular, are restricted to four grammar aspects which are repeated

over six or seven items each. It is not known how these four aspects were determined or whether indeed they reflect the skills and abilities that a successful Level 2 candidate should be able to demonstrate. Nor is it clear how the types of listening and reading passages were selected and how they contain discourse that reflects a Level 2 candidate.

The descriptive analyses of the test reveal that none of the sections produced normal test score distributions. That there are two peaks in the grammar and listening sections (Figures 1 and 2) indicates that these sections are not functioning well. While a normal distribution is not necessarily a requirement for a criterion referenced test, the distribution histogram for the total test scores (Figure 4) indicates that the majority of test-takers scored in the 150-210 range, clustering around the average of 181.6. This is irregular, since criterion referenced tests usually produce either positively skewed curves (where the scores tail off in a smooth curve towards the right end of the graph for an easy test) or negatively skewed curves (where the scores tail off in a smooth curve towards the left end of the graph for a difficult test. The fact that neither of those types of curves exist here indicates that there are problems with the test.

It is important to keep in mind that with a criterion reference test, the test-takers' scores are not being compared to each other (as in a norm referenced test like TOEFL and TOEIC), but are interpreted purely based on the number of items test-takers got right. The closer the score is to 100% (or at least 75% or higher for all three sections), the more closely the test-taker fits the profile of proficiency that the test aims to measure. Since the G-TELP was used at Gakushuin as a placement instrument to determine which students should enter the Intensive English course, a test which spread test-takers out over a normal distribution would have made it clearer who the most proficient students (according to the parameters of the test) were. If that is the case, then a norm referenced test, which

addresses general skills of language proficiency and covers a wider range of language and language use may be preferable. Further, the average scores for the reading and vocabulary section were much higher than that of the listening section (a 6.5% difference in the Gakushuin data and a 9% difference in national data for 1, 249 test-takers in Japan during the first half of 2002), suggesting that test level is not uniform across sections, at least for Japanese test-takers, and that the test may not be an accurate measure of the skills it directly addresses. Further, since only 14 of the 237 test-takers achieved a Level 2 score, the effectiveness of this test is questionable.Therefore, from a decision-making perspective, the G-TELP may not be the best instrument for placement purposes.

Item analyses revealed that 19 (24%) of the 80 test items were very weak (i.e., with ID values below .19), and therefore made no positive contribution to the test.These items are listed in Table 6, according to whether they were easy or difficult items (see also Tables 3-5).

Table 6: Summary of weak items with ID values below .19

Section	Easy items	Difficult items	Total (N)
Grammar	N=9 (#1, #3, #7, #9, #11, #13, #15, #16, #18)	N=1(#8)	10
Listening	N=1 (#51)	N=3(#43, #36, #50)	4
Reading and Vocabulary	N=4(#53, #56, #57, #74)	N=1(#68)	5

Clearly, there are many items which are too easy for this group of test-takers, suggesting that either the test is either entirely, or at least in part too easy, or that the test simply contains too many weak items that

need to be replaced with items that discriminate better between high and low scorers.

On the whole, the listening section discriminates most effectively, with only four very weak items, but still needs improving. The weakest section is the grammar section, which tests the following four grammar skills: participials, present unreal, past perfect, and future progressive. Half of the weak items in this section are participial items (N＝5: #1, #8, #11, #13, and #18). Therefore, participials are not really being tested here since five of the seven items which test this point are extremely weak. On the other hand, all of the future progressive items (#4, #6, #10, #14, #17, #24) are functioning fairly well, and can be said to be good or at least acceptable items.

The reliability coefficient of .85 is below Lado's .90-.99 range and is also below the .93 reliability figure of the TOEIC reading and listening sections (Woodford, 1991, p. 12), likely because roughly one quarter of the items are very weak. Therefore, the reliability of the G-TELP is questionable.

Conclusion

The G-TELP test was used as a placement instrument to screen applicants to the Intensive English Course at Gakushuin University. Using descriptive analyses and item analyses, the test was examined to assess its reliability. Descriptive analyses revealed that the none of the three sections of the test were functioning normally since they did not produce normal distributions. While a normal distribution is not necessarily a requirement for a criterion referenced test, the clustering of scores around the mean (Figure 4) and the appearance of double peaks in the grammar section and listening section histograms (Figures 1 and 2) indicate that the test is not functioning well, even for a criterion referenced test.

The item analyses showed that 24% of the items were extremely weak, and the majority of those items appeared in the grammar section. Thus, it is difficult to know if the results of the test indeed reflect the general description of what a Level 2 candidate can do.

The reliability coefficient of .85 is relatively low, suggesting that the test may not be a strong enough instrument for the important task of deciding who will be admitted to the Intensive English Course.

While there are a number of logistical factors to consider when selecting a placement test—cost, ease of administration, time required to generate and receive the results—the selection of an instrument which is appropriate to the task and which produces accurate results is crucial. Since there are obviously a number of problems with the G-TELP test, as evidenced by this study, it may be advisable to choose a more appropriate and reliable instrument for future test administrations.

Acknowledgements

1. Figures 1-4 are reproduced here with the permission of the G-TELP Tokyo office, October 2002. Raw item data provided by G-TELP was also used with permission.

2. Special thanks to Koarai Mikiya for his help with generating the statistical results.

References

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

Brown, J. D. (1996). *Testing in language programs.* Upper Saddle River, NJ: Prentice Hall Regents.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cam-

bridge University Press.

Microsoft Corporation.(2001). *Microsoft Excel 2001: mac* (Japanese version). Redmond, WA:Microsoft Corporation.

Morita, K. (1998). *国際英検 G-TELP,* 1, 2. Tokyo: Kinseido.

San Diego State University.(1998, April 1). *G-TELP Level Descriptors*. (booklet copied and distributed by the G-TELP Japan office)

Woodford, P. (1991, December 10). A historical overview of TOEIC and its mission. In TOEIC Steering Committee (Ed.), *The 35th TOEIC seminar* (pp. 10-15). Tokyo: The Institute for International Business Communication.

Appendix (Figures 1-4)

Figure 1: Grammar section score distribution

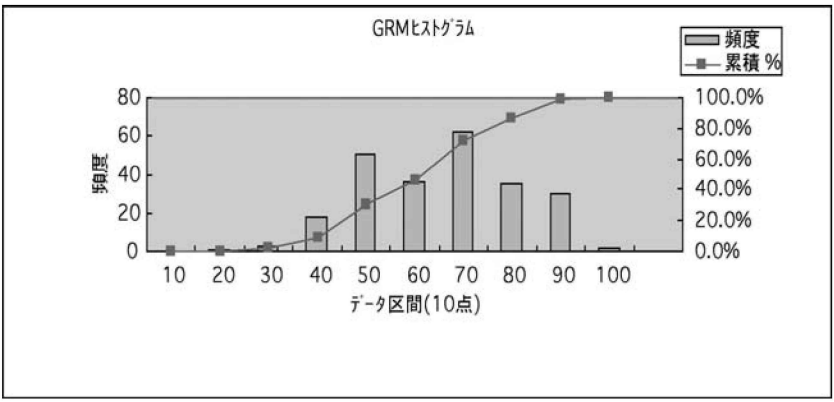


Figure 2: Listening section score distribution

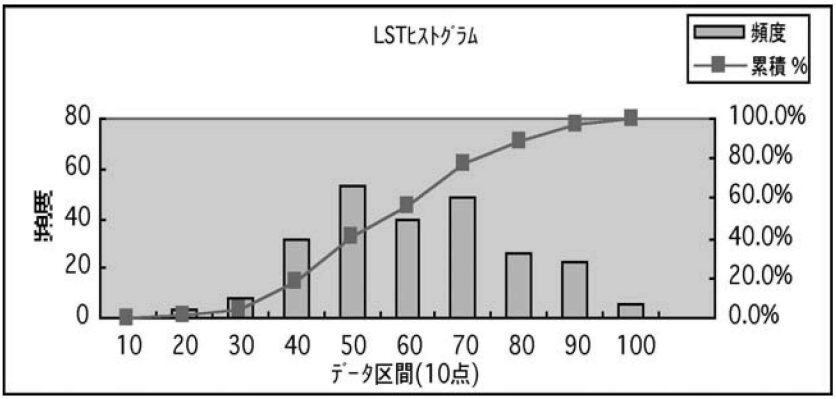


Figure 3: Reading and vocabulary section score distribution

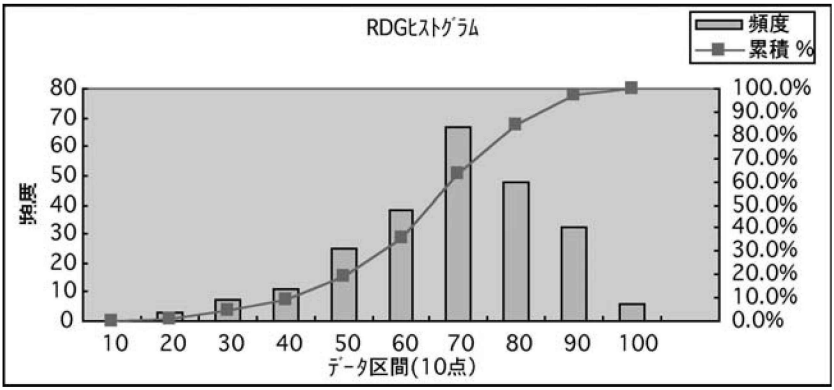
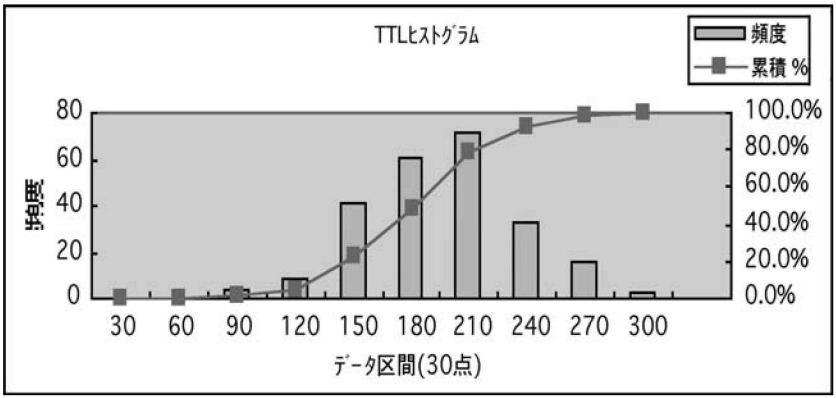


Figure 4: Total test score distribution



G-TELP Test Level 2 の信頼度に関する研究

ローラ　マクレガー

本稿では、G-TELP Test Level 2 の信頼度について調査を行う。G-TELP Test とは、サンディエゴ州立大学が開発し、維持・管理を行っている基準参照英語技能試験である。2002 年 4 月、学習院大学インテンシヴ・コースの選抜試験として、G-TELP Test が 237 名の学生に対して実施された。記述分析及び項目分析を通して学習院大学の試験データを調査することによって、項目の適切度及び試験としての信頼度の判定を行い、最後にプレイスメントテストとしての適性について結論を出したい。

記述分析では、どのセクションを見ても正規分布は認められなかったため、G-TELP Test が正常に機能していないということが明らかとなった。項目分析では、24%の項目（ほとんどが文法に関する項目）が極めて不適切であるということが明らかとなった。.85 という信頼度係数は、比較的低いものである。したがって、インテンシヴ・コースの履修者を決定するという重要な作業を行うためには、G-TELP Test は、不適切な手段であると言わざるを得ない。

プレイスメントテストを決定する際には、費用、手間、結果がでるまでにかかる時間等の様々な要因が関わってくるが、正確な結果が得られる適切な試験を選ぶことが肝腎である。本稿の研究で明らかなように、G-TELP Test には数多くの問題点があるため、今後の試験には、より適切で、信頼性の高いものを選ぶ必要があると思われる。