Articles

Validating a Questionnaire on Confidence in Speaking English as a Foreign Language

Dale T. Griffee
Seigakuin University

Despite repeated calls for reliability and validation of data elicitation instruments, research continues to be published based on questionnaires which do not report reliability or validation. The purpose of this paper is to examine the process by which a questionnaire, in this case one designed to measure confidence in speaking English as a foreign language (CSEFL), can be created, revised, and validated. Special attention is given to content validity, criterion validity, and construct validity. The concept and definition of validity is discussed and specific steps and procedures for the validation process are given. A pilot study is briefly summarized followed by the results of the present study. It is concluded that while the majority of the questionnaires used in ESL classroom research in Japan are not valid, the present study provides the necessary steps and procedures by which teacher-researchers can construct valid and reliable research questionnaires.

データ収集の道具の信頼性と妥当性の必要性が繰り返し叫ばれるにもかかわらず、信頼性と妥当性に関する言及がなされていないアンケートの結果にもとづいた研究が出版されつづきている。この論文の目的は、外国語として英語を話す自信（CSEFL)を測定するアンケートが作成され、変更され、妥当性をもつようになる過程を検討する。特に、内容的妥当性、基準関連妥当性、構成概念妥当性に焦点があてられる。妥当性の概念と定義が論じられ、妥当性を達成するための具体的な段取りが提案される。初めにパイロット・スタディが簡単に言及され、その後、本研究の結果が示される。結論としては、日本における英語の授業研究において使われているアンケートの大半は妥当ではないが、ここで示される段取りに従うことで、調査・研究を行なう教師は、妥当で信頼性のあるアンケートを作成することができるということである。

For some time interest in research has been growing among teachers of English as a foreign language (Nunan, 1992, p. xi). As a result of this interest, many classroom teachers have been taking a more active role in conducting and publishing research based on their own classroom observations (van Lier, 1988) and much of this classroom data is being gathered through teacher-designed instruments such as questionnaires and various forms of tests. Many of these instruments, however, are reported with little or no mention of either validity or reliability, which weakens any research based on them (Benson, 1991; Greer, 1996; Keim, Furuya, Doye & Carslon, 1996; Kobayashi, 1991; Teweles, 1996).

First I will begin by discussing the concepts and definitions of validity and reliability, next I will describe the steps and procedures involved in validating a questionnaire, and finally I will report a study aimed at creating a valid and reliable questionnaire. My major purpose is to argue for the role of validity and reliability tests in creating and reporting questionnaire research.

Definitions of Validity and Reliability

In validation, we are interested in arguments which show the degree to which an instrument measures what its author claims it to be measuring (Hatch & Lazaraton, 1991, p. 37; Most & Zeidner, 1995, p. 493). Although it is common to talk about instrument validation, validity is not a quality that belongs in some special way to an instrument. We cannot say that an instrument itself is valid or invalid, but rather that the instrument scores are valid for certain purposes (Cronbach, 1990, p. 145). For example, a proficiency test such as the TOEFL might be considered valid for approximating English proficiency but not for indicating ability to adapt to and live in an English speaking culture. In this sense, validity refers not to the instrument, nor to the scores, but to the use of the scores. More specifically, validity refers to inferences one makes using the scores of a certain test (Pedhazur & Schmelkin, 1991, p. 80).

The notion that there are different types of validity is controversial. Some researchers (Hatch & Lazaraton, 1991) state that there are different types of validity while others (Bachman, 1990; Most & Zeidner, 1995; Pedhazur & Schmelkin, 1991) claim that the notion of different types of validity is mistaken. Either way, it is thought important to report more than one type of validation process. As Bachman says, "it is only through the collection and interpretation of all relevant types of information that validity can be demonstrated" (1990, p. 237).

Important aspects of validation are content validation, criterion validation, and construct validation. Content validity has to do with how well an instrument measures what it says it is measuring (Brown, 1988, p. 102). Brown says that the first step is to establish what the instrument is measuring and the second step is to gather a panel of experts to judge the match between the individual items and the subsections of the instrument. To the extent the panel agrees, one can claim content validity. Criterion-related validity has to do with the extent to which a relationship exists between a high or low score on an instrument and an external criterion believed to indicate the ability being tested or measured. The most common type of criterion-related validation is to compare a new instrument against an established, reliable, and validated instrument. The problem is finding a criterion that is generally accepted and therefore valid because, as Kline (1995, p. 512) states, "the vast majority of psychological tests are not valid." Construct validity, considered central to the validation process (Bachman, 1990, p. 254), is the degree to which the instrument measures the construct under consideration. Construct validation is demonstrated through an argument that the construct, which we cannot see or measure, is indirectly being measured by questionnaire items, which can be seen and measured.

Reliability, on the other hand, is a statistical procedure that indicates how dependably an instrument measures what it claims to be measuring (Brown, 1988; Griffee, 1996a, 1996b; Hatch & Lazaraton, 1991). For any research instrument, including those created by teachers for the purpose of classroom data gathering, one should report both validity and reliability (this, of course, does not include questionnaire forms used only for pedagogical purposes within the classroom). Without such reporting, the reader cannot know how to interpret the inferences made on the basis of the data (Bachman, 1990, p. 24). To put it more bluntly, reliability is a necessary, but not sufficient precondition for validity. If a questionnaire is not reliable, it cannot be valid (Pedhazur & Schmelkin, 1991, p. 81).

It is not the responsibility of the reader to assume reliability (or validity, for that matter); both must be reported. There have been repeated calls for reporting of both validity and reliability (Chaudron, 1988; Kasper & Dahl, 1991; Long, 1990; Luppescu & Day, 1990); these calls apparently are not having much effect among classroom practitioners as evidenced by a check of *The Language Teacher*, a monthly classroom teacher journal published in Japan. From 1976 to 1996, not one of the 13 articles employing questionnaire data in their findings reported instrument reliability or offered any evidence of validation. For the same period, of the

12 articles *JALT Journal* which used data from questionnaires, none reported reliability and nine made no mention of validity. In considering how to construct a questionnaire instrument for research, the literature below suggests five stages of development: the before-writing stage; the writing stage; the piloting stage; the reliability determination stage; and the validation stage.

The Before-Writing Stage—Psychological Constructs

To understand validation, it is necessary to understand what a psychological construct is. A psychological construct is "a theoretically existing (but unobservable) variable" whose existence can be inferred from a variety of sources (Slavin, 1992, p. 244). In the language teaching profession, teachers commonly discuss such psychological constructs as intelligence, aptitude, motivation, confidence, and proficiency. Questionnaires ask specific questions in an attempt to measure such constructs.

Recall that validity is the degree to which inferences can be made about what an instrument claims to be measuring (Ary, Jacobs, & Razavieh, 1990, p. 256; Brown, 1996, p. 231). While validity is not proof, it is an argument on the basis of which researchers hope to convince their readers that the instrument is being used in the situation for which it was designed. In the case of a new instrument, validity is demonstrated through an argument that the instrument is correctly designed for the purposes the researcher has in mind. In order to argue that an instrument is measuring what the researcher states it is measuring, the researcher must make clear what construct is being measured by the instrument. It is for this reason that Bachman (1990) suggests that a first step in instrument creation is to examine theories that discuss what we intend to measure. If no relevant theory exists, Bachman suggests that we could at least create a definition of what we are trying to measure and list the content areas. These content areas can then become the subtests of our instrument (Brown, 1988). For example, suppose that a researcher wants to measure the construct "confidence." He or she examines the theoretical literature on the subject and perhaps finds a paper that defines the term and argues that confidence is composed of qualities X and Y. It is not possible for researchers to directly examine or measure the construct of confidence in students. Nor is it possible to directly measure qualities X and Y. But qualities X and Y are more specific than the construct, and items can be devised that infer the existence of quality X and quality Y. In this way, X and Y have become the

basis for the subsections of the instrument. The instrument will have two sections, a section with items purporting to measure quality X and another section composed of items purporting to measure quality Y.

In addition to a serious consideration of the construct, it is also necessary to think about such issues as the requirements for classroom use. For example, how many pages will the instrument contain? Will negative questions be allowed? And what is the type of data desired? (e.g., Likert scales, cloze passages, or open-ended questions) (Tullock-Rhody & Alexander, 1980). In thinking about items which might be included in a questionnaire, Allen (1995) suggests brainstorming items from researcher intuition as well as gathering items from the literature. Another way to elicit items is to ask students similar to those for whom the questionnaire is being developed for items (Horwitz, 1988). For example, in describing a reading questionnaire designed to distinguish good readers from poor readers, Tullock-Rhody and Alexander (1980) report sessions in which they asked elementary school children to describe someone they knew who was a good reader and someone they knew who was a poor reader. Students' views were incorporated into their questionnaire using the students' own language as much as possible.

The Writing Stage

Brown (1996, p. 233) suggests arranging the content areas previously identified and deciding how many items would be needed in each category. Brown also suggests asking colleagues to help in writing items and writing one-third more items than deemed necessary. If some items are not adequate, they can be eliminated. Logically analyze your scoring procedures (Pedhazur & Schmelkin, 1991). Can your construct be measured by "yes" or "no" questions or do you require a greater range of possible responses? If you select a Likert scale, ask a knowledgeable colleague if your scale logically covers all responsible responses in an equal fashion. Try to avoid conflating categories in your instructions to respondents. An example of a conflating or confusing category would be asking respondents if they "believe and approve of" certain practices because it is possible to believe X without approving of X. For example, it is possible to believe that persons should be allowed to smoke cigarettes without approving of smoking. After items have been written, ask expert judges, persons who might be expected to be interested in and experienced with the construct your instrument is attempting to measure, to evaluate your items against the construct. In our imaginary example above, expert judges would be asked to evaluate each item in the

subsection against the quality that subsection is attempting to measure. The issue could be stated, do the items in the X section actually measure quality X? If a number of judges object to a given item, serious consideration should be given to either revising or eliminating the item. When all items have been vetted, show them to students similar to the ones for whom the instrument is designed. Ask these students to check each item for comprehensibility and to indicate any vocabulary item they do not understand. It may be necessary to substitute easier vocabulary items or to paraphrase certain items, but a higher level of understanding on the part of respondents will result in less guessing, which in turn will result in higher instrument reliability.

The Piloting Stage

Pilot the instrument on the same type of students for whom the instrument is being designed. In the pilot study, consider writing similar items, placing them in random order, and then correlating student answers to these paired items to see if students answered them in the same way (Reid, 1990; see also Griffee, 1996a). A high correlation between paired items indicates that students interpret the items in a similar way. A low or negative correlation indicates that students are not answering the items in a similar way, which becomes a source of randomness or unreliability. As an alternative, you can correlate each item with the total test scores and keep only the items with high correlations (Cronbach, 1990, p. 170). Revising or eliminating items having low correlation will tend to have the effect of making questionnaire items more consistent and thus more reliable.

The Reliability Stage

With the results of the pilot study, calculate descriptive statistics, reliability coefficients, and the standard error of measurement (Brown, 1996; Griffee, 1996a). What constitutes an adequate reliability coefficient depends on at least six factors: the type of decision, the importance of the decision, the type of reliability estimate, the construct being measured, the instrument medium, and the amount of error the researcher is willing to accept (Griffee, 1996b). The type of decision refers to whether the instrument is being used to measure individuals or to compare groups. Making decisions about individuals demands higher reliability than comparing groups (Pedhazur & Schmelkin, 1991, p. 109). Importance of decision refers to how serious the decision is

and how irrevocable the decision is (e.g., acceptance into or rejection from a program). Serious, irrevocable decisions demand higher reliability because of the effect of the decision on individual lives. The type of reliability refers to the formula being used to calculate the coefficient or to the type of reliability calculation (e.g. test-retest, internal consistency). For example, the Kuder-Richardson 21 formula tends to underestimate reliability compared with the Kuder-Richardson 20 formula. The construct being measured refers to whether the construct is easy to measure or difficult to measure (e.g. a mood, feeling, or trait). We may tolerate lower reliability for a difficult-to-measure construct than we will for an easy-to-measure construct. The instrument medium refers to whether the instrument is paper-and-pencil or an interview. An interview might be allowed lower reliability than a paper-and-pencil test. Finally, a researcher may accept lower reliability in an early phase of the research than at a later phase. Table 1 summarizes these comments. There is no hard and fast rule on what constitutes acceptable reliability. Although some writers (Vierra & Pollock, 1992, p. 62) suggest .70 as a cutoff point, others (Ary, Jacobs, & Razavieh, 1990, p. 282) would allow lower levels of reliability, from .30 to .50, for decisions about groups. Finally, Pedhazur & Schmelkin (1991, p. 104) discuss various formulas for determining the reliability coefficient and conclude that Cronbach's alpha is the coefficient of choice when measuring constructs.

Table 1: Factors to consider in determining adequate reliability

Factors	Operationalized as	Reliability could be: higher	lower
1. The type of decision	Who/what being measured?	Individual	Group
2. The importance of the decision	What is being decided?	Serious	Not serious
3. The type of reliability	Which formula is used?	KR-20	KR-estimate
4. The construct being measured	Is it difficult or easy to measure?	Easy	Difficult
5. The instrument media	Paper & pencil or interview?	Written	Interview
6. The amount of error the researcher is willing to accept	What stage is the research at?	Late	Early

Table 2: Steps in creating a valid and reliable questionnaire

Stages and procedures
Before writing
1. Investigate available theories that describe your construct.
2. Review all instruments purporting to measure your construct.
3. Define the construct you are trying to measure.
4. List classroom requirements and type of data you want.
5. Brainstorm items from self and literature.
6. Interview colleagues and students for items.
Item writing
7. Decide how many items are required for each subtest or content area, then write more items than are needed.
8. Ask your colleagues for help in item writing.
9. Logically analyze the scoring procedures.
10. Ask expert judges and students to review items.
Piloting
11. Consider pairing and correlating items. Correlate matched pairs, or correlate NS and NNS pairs, or correlate each item with the total, and eliminate or revise low correlating pairs, and pilot again.
12. Pilot the instrument with students similar to those for whom the test is intended.
Reliability determination
13. Calculate descriptive statistics and reliability coefficient.
Validation
14. Explore content validity by convening a panel of experts to judge the match of questionnaire items to construct content.
15. Explore construct validity by conducting a differential group experiment or an intervention experiment.
16. Explore criterion-related validity.

The Validation Stage

It is traditional to consider three types of validity: content validity, construct validity, and criterion-related validity. Bachman (1990, p. 236) suggests that validation is a unitary concept and argues that all three types of validity must be investigated and reported. Content validity can be explored by convening a panel of experts to judge the degree to which the instrument items actually represent the elements being tested (Ary, Jacobs & Razavieh, 1990; Brown, 1996). Construct validity can be

explored by differential group experiments, intervention experiments (Brown, 1996), or factor analysis (Boyal, Stankov, & Cattell, 1995; Kline, 1994). A differential groups experiment compares the performance of two groups on a test, one group which obviously has the construct and another group which obviously does not have the construct. An intervention experiment is similar but uses only one group, for example, first year students at the beginning of the school year and the same students at the end of the school year. If the students score higher with each subsequent instrument administration, a researcher can argue that the construct is being acquired. Construct validity can also be explored by statistical procedures such as factor analysis which seek to locate and identify various factors underlying the construction of an instrument. Criterion-related validity can be explored by demonstrating a relationship between test scores of a pilot group similar to those for whom the instrument is designed and some other criterion instrument which is believed to measure the construct being tested, such as: ability as defined by group membership, a recognized test of the same ability, or success on a task that involves the ability being tested (Bachman, 1990, p. 248).

Table 2 lists and summarizes the general stages and specific steps in creating and validating a questionnaire. While in practice it might not be possible or even desirable to realize all 16 procedures, they are listed here for the sake of completeness.

Pilot Study

A pilot study was conducted (Griffee, 1996c) which formed the background of the present study. Two test sources (Mitchell, 1983; Sweetland & Keyser, 1991) were searched for questionnaires measuring confidence and none were found. It was determined that a questionnaire measuring confidence in speaking English would be constructed. Twenty items were brainstormed and administered to 25 university students. Reliability was calculated using the Cronbach alpha formula and paired items were correlated. A factor analysis was calculated looking for roots greater than one using the oblique transformation method. Three factors were identified with eigen values greater than one suggesting that there are possibly three factors of interest. Two factors were identified as a combination of ability (Factor 1) and willingness to engage with others (Factor 3). Factor two was identified as outgoingness or low anxiety.

The Present Study

The primary purpose of this paper is to explain and demonstrate how a questionnaire can be constructed and validated. The purpose of reporting the present study on the creation and validation of a questionnaire on confidence in speaking English as a foreign language (CSEFL) is to illustrate the steps that were taken. The specific research questions addressed in this study are:

1) What is the degree of content validity of the CSEFL?
2) What is the degree of criterion validity of the CSEFL?
3) What is the degree of construct validity of the CSEFL?

Method

Subjects: There were 250 subjects in this study drawn from four small, private colleges in Saitama, Japan. For the most part, the students were in their first or second year, were in their early 20s, and had a variety of majors. Approximately half of the students were males and approximately half were females. Proficiency scores were not available for all students. The entire sample of convenience consisted of each student in 10 intact classes. See Table 3 for group size, school, and school year.

Materials: Version one of the CSEFL questionnaire from the pilot was taken as the base document. Six items having low correlations were eliminated and a panel of experts which consisted of two English native speaker (ENS) males, two ENS females, two Japanese native speakers (JNS) males,

Table 3: Groups, Schools, School Year of Subjects & Alpha Reliability

Group/College	Number	Year	alpha reliability
1. S. Junior College	20	first	.84
2. M. University	26	second	.88
3. T. I. University	21	first	.90
4. S. University	25	first	.92
5. S. University	16	second	.85
6. T. I. University	25	third	.92
7. T. I. University	39	third	.94
8. S. Junior College	21	first	.70
9. S. University	24	first	.86
10. S. University	33	second	.92

and two JNS females was convened to judge the adequacy of the remaining items. The eight panel members, equally divided by gender and ethnic group to reduce possible bias, were interviewed and as a result, six items were dropped. In addition, one item from the factor analysis did not load on any factor and was cut, leaving nine items from the original questionnaire.

A theoretical model of the construct "confidence" was created which hypothesized three aspects of confidence: ability, assurance, and willing engagement. By ability what is meant a command of grammar, vocabulary, and pronunciation. By assurance what is meant that the speaker has a feeling of security and comfort in speaking English. By willing engagement what is meant the speaker is glad to speak in English with native speakers of English.

To create additional items, five colleagues (one JNS female, two ENS females, one JNS male, and one ENS male) were interviewed asking two questions each: think of a person you know who can speak (English/Japanese) with confidence; what are some specific things they do that make you think they are confident? The JNSs were asked about persons who could speak English confidently and the ENSs were asked about persons who could speak Japanese confidently. Twenty-four items were gathered from the interviews. In addition, as a class exercise, 16 second-year students were asked the same questions and given time to write their answers. Twenty-three items were collected and combined with the 24 colleague answers and the nine original questionnaire items creating a pool of 56 items. From this pool, 30 items were selected for inclusion in the revised questionnaire: 10 under the ability category, 11 under the assurance category, and nine under the willing engagement category. An additional panel of 12 experts was convened to review the pool of 30 items and make recommendations for exclusion or inclusion in the questionnaire.

Procedures: The questionnaire was given to five teachers at the four schools. After teachers were instructed on the nature and purpose of the questionnaire, they administered the questionnaire in their classes and returned the questionnaire to the researcher, who scored it. To help establish criterion-related validity, teachers were asked to select one or two persons in each class who the teacher believed would score high on the confidence questionnaire and one or two students who would score low. Selection was to occur before the questionnaire was administered.

Analysis: The alpha level was set at .05 and all statistics were calculated using StatView 4.5 statistical program for the Macintosh (Abacus Concepts,

1995). The statistical procedures used were Factor Analysis (FA) and Pearson Product-Moment Correlation. In the FA oblique rotation was used and the factor extraction method was the Iterated Principal Axis method using the squared multiple correlation for estimating the initial commonalties. The number of factors to extract was determined by the number with eigen values greater than one. All data sets were independent and, given the large N size, the assumptions of factor analysis e.g. normal distribution are assumed to have been met.

Results

To investigate the first research question on content validity, a 12-member panel (three ENS women, three ENS men, three JNS women, and three JNS men) was convened. An expert was defined as a person who, because of vocation and professional interest, might reasonably be considered as having both interest and knowledge of the subject area under consideration. The panel was asked to rate all items as to validity on a five-point Likert scale of strongly agree, agree, undecided, disagree, and strongly disagree. Five items received five or more negative votes and were eliminated. From the remaining items, a second CSEFL questionnaire was created with 24 items in the three categories of ability, assurance, and willing engagement, and these were randomly ordered.

The CSEFL questionnaire is designed for typical Japanese university students in Japan. Since this group can be comprised of students from intermediate proficiency to rather low proficiency, it was felt that exposing low-level students to the items would yield useful feedback. Six students (three males and three females) typical of the lower proficiency student who would take the questionnaire were individually asked to read the new 24 item CSEFL and indicate any item or word which was not clear. The students did not reject any item as a whole, but did indicate several specific words which they did not understand. One such vocabulary item was the word "argue" (I can argue in English with native speakers) and another word was "willing" (I am willing to speak to many foreigners). "Argue" was changed to "discuss" and "willing" was changed to "I hope to." After these changes were made another eight students (four males and four females) were interviewed in a similar manner and these eight students did not indicate any difficulty with the revised items.

To investigate the second research question on criterion validity, all teachers were asked to nominate one or two students in each class who they believed would score high on the CSEFL questionnaire, and one or two students they believed would score low. The teachers nominated

Table 4: Teacher Nominations of High & Low Confidence Scorers

Class	*N*	Students			
		Nominated to score high	Actually scored high	Nominated to score low	Actually scored low
1	20	2	1	2	1
2	26	2	2	1	0
3	21	2	2	1	1
4	25	2	2	2	1
5	16	1	1	1	0
6	25	1	1	2	2
7	39	3	3	2	1
8	21	3	1	3	1
9	24	3	1	3	1
10	33	2	1	2	0
Total	250	21	15	19	8
Percent		0.71		0.42	

21 students they believed would score high, and 19 students they believed would score low. Students nominated to score high were judged to have actually scored high if their scores were in the top one-third of the class and those nominated to score low were considered to have actually scored low if their percentage correct was in the bottom one-third of the class scores. Table 4 shows the results. The CSEFL agrees with teacher ratings 71% at the higher end and 42% at the lower end.

To investigate the third research question on content validity, first a Principle Components Analysis (PCA) was used followed by Factor Analysis (FA). Hatch & Lazaraton (1991, p. 493) suggest using oblique rotation

Table 5: Factors and Variance Proportions for the PCA

Factors	Magnitude	Variance Proportion
Factor 1	7.724	.322
Factor 2	2.159	.090
Factor 3	1.178	.049
Factor 4	1.129	.047
Factor 5	1.113	.046

factor analysis (FA) to confirm PCA because FA looks at only common variance and ignores error variance and variance not shared by all the factors. The PCA revealed 12 factors with five factors having eigen values over one. Table 5 shows the five factors, their magnitude, and how much of the total variance they account for.

During data inputting, it appeared that some of the items had been rated by students in a contradictory way. For example, many respondents who consistently circled "undecided," "disagree," and "strongly disagree" for most items circled "agree" or even "strongly agree" for item 15 (At a party, I often talk to someone I don't know in English).

Table 6: Correlations of Each Item with the Total Minus Itself

Item number as it appeared in the original brainstorm list	Item number as it appeared in the questionnaire version 2	Correlation
1	5	.568
2	11	.669
3	17	.458
4	1	.545
5	9	.543
6	21	.486
7	20	.474
8	8	.550
9	19	.467
10	4	.656
11	12	.668
12	13	.573
13	22	.545
14	14	.426
15	6	.499
16	24	.455
17	18	.531
18	15	.033•
19	7	.505
20	10	.544
21	16	.496
22	3	.368
23	23	.435
24	2	.551

Note. • = non-significant correlation, all others significant at $p < .05$.

Table 7: Factors and Variance Proportions for the FA

Factors	Magnitude	Variance Proportion
Factor 1	5.440	.363
Factor 2	1.332	.089
Factor 3	.556	.037

Why would students who consistently indicate that they do not like speaking English suddenly indicate that at parties they would talk to a stranger in English? Perhaps a construct other than confidence is being tapped. Kline (1995) suggests using item analysis to remove bad items and factor the reduced set. Each item was correlated against the total minus itself which resulted in the correlations in Table 6.

Table 6 shows questionnaire items 1-8, which were the items hypothesized to measure factor one (ability), items 9-16, factor two (assurance), and items 17-24, factor three (willing engagement). The five highest correlations in each of the three groups were selected and refactored.

Table 8: Factor Loadings: Oblique Solution Primary Pattern Matrix

	Factor 1	Factor 2	Factor 3
Item 1	.520*	.147	-.036
Item 2	-.804	.731*	-.001
Item 4	.178	.691*	.009
Item 5	.546*	.202	.044
Item 6	.198	.396*	.081
Item 7	.396*	.307*	-.077
Item 8	.714*	-.065	-.006
Item 9	.717*	-.038	-.048
Item 10	.576*	.076	-.001
Item 11	.779*	-.018	.053
Item 12	.043	.379*	.442*
Item 13	.012	.138	.992*
Item 16	-.024	.611*	.104
Item 18	.348*	.214	.080
Item 22	.593*	-.048	.114

Note: * = factor loadings at .30 or higher

FA shows three factors, two of which have eigen values over one. The magnitude and proportion of the variance of all factors can be seen in Table 7. The oblique solution primary pattern matrix, Table 8, shows nine items load on factor one, six items load on factor two, and two items load on factor three.

It was hypothesized that five items would load on each of three factors. Results show that all five of the items predicted to load on ability, did so (items 5, 11, 1, 9, and 8), that three out of five predicted items loaded on assurance (items 4, 12, and 6), but that none of the predicted items loaded on willing engagement. In addition, four items loaded in ways which were not predicted (items 7, 10, 18, and 22). Items 7 and 12 load at significant levels on two factors and were cut from CSEFL version three as well as item 13 which loaded only on factor three. This left 12 items for the working version of the questionnaire which appears in the Appendix as version three.

Discussion

The first question is, what is the degree of content validity of the CSEFL? To bring about content validation, two steps must be taken. One, it must be decided what the instrument is claiming to measure and two, it must be decided how to measure the representativeness of each part of the instrument. Condition one has been met in that a model of confidence in speaking English as a foreign language was created which hypothesized three content areas. Condition two has been met in that a panel of experts rated each item in each of the three content areas. All items in the CSEFL have a high degree of panel approval, thus content validly can be claimed.

The second question is, what is the degree of criterion validity of the CSEFL?

Since there are no known reliable or valid measures of confidence in speaking English as a foreign language, this paper uses teacher judgment as a criterion. Criterion response as measured by teacher judgments of students who will score high and students who will score low was mixed. Teachers were generally able to identify students who would score high, but less able to identify students who would score low. One possible explanation is that the CSEFL questionnaire is valid for identifying speakers who are confident, but is not valid for identifying speakers who are not confident. Another possible explanation is that teachers cannot adequately judge certain types of students who appear as not having confidence when in fact, they do. This researcher marked one

female student as being low in confidence whereas her score placed her in about in the middle of the class. In a subsequent class exercise, this student declared herself to be an analytic learner who likes solitary tasks such as reading (Nunan, 1988, p. 91). It may be possible that her learning style was interpreted as lack of confidence. It may be necessary to include learning style in addition to the results of a questionnaire such as the CSEFL in compiling a student profile. Against teacher judgment of high achievement, the CSEFL has a relatively satisfactory rating and thus at least partial criterion-related validity can be claimed.

The third question is, what is the degree of construct validity of the CSEFL? The results of the factor analysis are not as clear as we might wish. The high Cronbach alpha reliability coefficients might indicate that high internal consistency in fact reflects item redundancy in which items are little more than paraphrases of each other (Boyal, Stankov, & Cattell, 1995, p. 436). On the other hand, two of the factors have high predicted loadings, which tends to support the validity of the hypothesized construct. The loadings on the third factor are so low as to indicate not only that is particular factor is not supported but also that no additional factor can be substantiated. It may be the case that the lack of a full theoretical model accounting for and describing the construct of confidence leaves us in ignorance as to additional factors. Finally, Boyal, Stankov, and Cattell (1995, p. 421) indicate that while FA provides evidence as to construct validity, which is important, such evidence alone is insufficient. They maintain that predictive evidence alone is essential, and future research may be necessary along those lines. However, the construct validity, criterion validity, and construct validity obtained in the present study suggest that we can argue for partial construct validation. Taking all three types of validation procedures into consideration, it can be argued that the CSEFL is a valid instrument for purposes of researching groups while maintaining some reservations when it comes to individuals keeping in mind the warning of Bachman and Palmer (1996, p. 22) that "it is important for test developers and users to realize that test validation is an on-going process and that the interpretations we make of test scores can never be considered absolutely valid."

Conclusion

This paper has pointed out that the vast majority of the questionnaires used in ESL and EFL classroom research offer no evidence of validation and that conclusions based on the results of such questionnaires are problematic. There might be at least three reasons for this

state of affairs. One is that teacher-researchers do not believe it is necessary to report validity or reliability. Second, validity is seen as residing in the instrument. If the instrument was considered valid in another country for another student population, then it must be valid in this country for our students. Third, and closely related, is the idea that if an instrument has been judged valid once, then it must be valid for all time. None of these assumptions are correct and their combined effect is the continued use of invalid and unreliable instruments which results in flawed research. The present study indicates some of the necessary steps and procedures teacher-researchers can take to promote valid and reliable research instruments.

Acknowledgments

The author wishes to thank J. D. Brown for advice and help on earlier forms of this paper as well as a perceptive, unnamed JALT Journal *reviewer.*

Dale T. Griffee, Associate Professor at Seigakuin University, is author of several ESL textbooks. He is editor of the *JALT Applied Materials* (JAM) series and co-editor, with David Nunan, of *Classroom Teachers and Classroom Research*, JALT, 1997. His major research interests are testing, evaluation, and assessment as well as classroom research.

References

Abacus Concepts, StatView 4.5 (Computer Software). (1995). Berkeley, CA: Abacus Concepts.

Allen, A. (1995). Begging the questionnaire: Instrument effect on readers' responses to a self-report checklist. *Language Testing, 12*(2), 133-156.

Ary, D., Jacobs, L., & Razavieh, A. (1990). *Introduction to research in education* (4th ed.). Harcourt Brace.

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford University Press.

Benson, M. (1991). Attitudes and motivation towards English: A survey of Japanese freshmen. *RELC Journal, 22*(1), 34-48.

Boyal, G. J., Stankov, L., & Cattell, R. B. (1995). Measurement and statistical models in the study of personality and intelligence. In D. H. Saklofske & M. Zeidner (Eds.), *International handbook of personality and intelligence* (pp. 417-446). New York: Plenum Press.

Brown, J. D. (1988). *Understanding research in second language learning*. Cambridge: Cambridge University Press.

Brown, J. D. (1996). *Testing in language programs.* Prentice Hall-Regents.

Chaudron, C. (1988). *Second language classrooms: Research on teaching and learning.* Cambridge: Cambridge University Press.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper Collins.

Greer, D. (1996). Gardner and Lambert in the classroom. *The Language Teacher, 17*(1), 10-14.

Griffee, D. T. (1996a). Reliability and a learner style questionnaire. In G. van Troyer, S. Cornwell, & H. Morikawa (Eds.), *On JALT 95 Curriculum & Evaluation: Proceedings of the JALT 1995 International Conference on Language Teaching/Learning* (pp. 283-292). Tokyo: The Japan Association for Language Teaching.

Griffee, D. T. (1996b). Classroom research, instrument reliability, and instrument revision. *Temple University Japan Working Papers in Applied Linguistics, 9,* 25-41.

Griffee, D. T. (1996c). Validation and classroom research instruments: Process and product. *The Journal of Seigakuin University, 9*(1), 53-70.

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics.* New York: Newbury House.

Horwitz, E. K. (1988). The beliefs about language learning of beginning university foreign language students. *The Modern Language Journal, 72*(3), 283-294.

Kasper, G., & Dahl, M. (1991). Research methods in interlanguage pragmatics. *Studies in Second Language Acquisition, 13*(2), 215-247.

Keim, B., Furuya, R., Doye, C., & Carlson, A. (1996). A survey of the attitudes and beliefs about foreign language learning of Japanese university students taking communicative courses. *JACET Bulletin, 27,* 87-106.

Kline, P. (1994). *An easy guide to factor analysis.* London: Rutledge.

Kline, P. (1995). A critical review of the measurement of personality and intelligence. In D. H. Saklofske & M. Zeidner (Eds.), *International handbook of personality and intelligence* (pp. 505-524). New York: Plenum Press.

Kobayashi, J. (1991). Cross-cultural differences in classroom management: Coping with student silences and communication failures. *The Language Teacher, 15*(6), 17-19.

Long, M. (1990). Second language classroom research and teacher education. In Brumfit, C. & Mitchell, R. (Eds.) *Research in the language classroom* (pp. 161-170). ELT Documents 133. Modern English Publications in association with The British Council.

Luppescu, S., & Day, R. (1990). Examining attitude in teachers and students: The need to evaluate questionnaire data. *Second Language Research, 6*(2), 125-134.

Mitchell, J. V. (Ed.) (1983). *Tests in print III.* Lincoln, NE: University of Nebraska Press.

Most, R., & Zeidner, M. (1995). Constructing personality and intelligence instruments. In D. H. Saklofske & M. Zeidner (Eds.), *International handbook of*

personality and intelligence (pp. 417-446). New York: Plenum Press.
Nunan, D. (1988). *The learner-centered curriculum*. Cambridge: Cambridge University Press.
Nunan, D. (1992). *Research methods in language learning*. Cambridge: Cambridge University Press.
Pedhazur, E. S., & Schmelkin, L. P. (1991). *Measurement, design, and analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Reid, J. (1990). The dirty laundry of ESL survey research. *TESOL Quarterly, 24*(2), 323-338.
Slavin, R. E. (1992). *Research methods in education* (2nd ed.). Boston: Allyn and Bacon.
Sweetland, R. C., & Keyser, D. J. (Eds.). (1991). *A comprehensive reference for assessment in psychology, education, and business* (3rd ed.). Austin, TX: Pro-ed.
Teweles, B. (1996). Motivational differences between Chinese and Japanese learners of English as a foreign language. *JALT Journal, 18*(2), 211-228.
Tullock-Rhody, R., & Alexander, J. E. (1980). A scale for assessing attitudes toward reading in secondary schools. *Journal of Reading, 23*(7), 606-614.
van Lier, L. (1988). *The classroom and the language learner: Ethnography and second-language classroom research*. London: Longman.
Vierra, A., & Pollock, J. (1992). *Reading educational research* (2nd ed.). Scottsdale, AZ: Gorsuch Scarisbrick Publishers.

(Received January 7, 1997)

Appendix: Version 3 of Confidence in Speaking Questionnaire

Confidence in Speaking English v.3

Name _____ Student # _____

How confident are you in speaking English?
Circle your best answer for each statement.

For example:
- I like ice cream.

Strongly agree Agree Undecided Disagree Strongly disagree

1. I can be interviewed in English.
Strongly agree Agree Undecided Disagree Strongly disagree

2. I would like to study in an English speaking country.
Strongly agree Agree Undecided Disagree Strongly disagree

3. I like speaking English.
Strongly agree Agree Undecided Disagree Strongly disagree

4. I can discuss in English with native speakers.
Strongly agree Agree Undecided Disagree Strongly disagree

5. When I speak English I feel cheerful.
Strongly agree Agree Undecided Disagree Strongly disagree

6. I can speak English easily.
Strongly agree Agree Undecided Disagree Strongly disagree

7. I can show an English speaking visitor around the campus and answer questions.
Strongly agree Agree Undecided Disagree Strongly disagree

8. I say something to other people in English everyday.
Strongly agree Agree Undecided Disagree Strongly disagree

9. I can give my opinion in English when talking to a native speaker.
Strongly agree Agree Undecided Disagree Strongly disagree

10. I look for chances to speak English.
Strongly agree Agree Undecided Disagree Strongly disagree

11. I will speak to a group of people in English.
Strongly agree Agree Undecided Disagree Strongly disagree

12. I am relaxed when speaking English.
Strongly agree Agree Undecided Disagree Strongly disagree