Research Issues and Language Program Direction

L. Kathy Heilenman
Editor

HH Heinle & Heinle Publishers
an International Thomson Publishing company
ITP Boston, Massachusetts 02116 U.S.A.

Manufactured in the United States of America.

ISBN: 0-8384-1023-5

10 9 8 7 6 5 4 3 2 1

Investigating the Properties of Assessment Instruments and the Setting of Proficiency Standards for Admission into University Second Language Courses

Micheline Chalhoub-Deville
The University of Iowa

Introduction

The introduction of the Proficiency Guidelines (ACTFL 1986) and the corresponding Oral Proficiency Interview (OPI) have placed heavy emphasis on the assessment of oral language ability. As Teschner (1991) writes, in the early 1980s the notion of proficiency "often appeared to focus solely on speaking in general and the oral interview in particular" (p. ix). Notwithstanding the contribution of the proficiency movement to assessing speaking, the assessment of the other modalities is also critical in order to get a rich and more complete picture of learners' second language (L2) abilities. Consequently, in recent years educators have increasingly encompassed a broader conceptualization of proficiency which includes all four modalities. The 1990s have accordingly witnessed increased activity focusing on the development of proficiency-based instruments for assessing speaking as well as the other modalities. Institutions, such as The Ohio State University (Birchbichler, Corl, and Deville 1993; Corl, Harlow, Macián, and Saunders 1996; Robinson 1996), the University of South Carolina (Mosher 1989; Fleak 1991), and The University of Iowa (Wherritt, Druva-Roush, and Moore 1990) have assumed a prominent role in developing proficiency-based assessments in the various modalities to be used in undergraduate language programs.

Since the mid 1980s, the University of Minnesota has also been one of the leading institutions in setting up college-wide entrance and graduation L2 proficiency requirements and developing instruments based on the ACTFL Guidelines for assessing all four modalities (Barnes, Klee, and Wakefield 1990; Lange 1987; Lange, Prior, and Sims 1992). Today, it continues its commitment to proficiency-based assessment and, as part of the Minnesota Articulation Project (MNAP), has recently developed new assessments to replace the initial entrance proficiency instruments. (The University of Minnesota is also working to revamp the graduation proficiency instruments.) These MNAP instruments have been described in detail in Chalhoub-Deville (1997).

While Chalhoub-Deville (1997) presents detailed information about the design, content, and format of the MNAP instruments, the present paper documents their psychometric properties. More specifically, the present paper describes the standard setting process that preceded the administration of the various instruments and reports the results of field testing at the University of Minnesota. A brief description of the MNAP and its assessments follows.

The MNAP Assessment Instruments

MNAP is a statewide effort that includes the University of Minnesota, various public and private schools, community colleges, private colleges, and a state university. The MNAP agenda includes the development of an operational model and a corresponding battery of assessment instruments for coordinating L2 outcomes across levels of instruction and educational systems in Minnesota. MNAP has produced assessment instruments in French, German, and Spanish for assessing students' L2 proficiency as they move from the secondary into the postsecondary levels. Specifically, the purpose of these instruments is to help determine if students can perform in the various modalities at the Intermediate Low (IL) level, the proficiency standard required for students to enroll in second-year French, German, or Spanish programs at the postsecondary level.

In the remainder of this section and article, the discussion will focus on the MNAP reading and writing assessments that were field tested at the University of Minnesota in the summer of 1996. (Instruments exist for all four modalities, but present space constraints permit discussion of only these two.)

The reading and writing instruments developed in French, German, and Spanish are thematically based. Each assessment instrument has a theme deemed appropriate to incoming students (e.g., an exchange program, pen pal, etc.). Each of the thematically oriented reading instruments, for example, includes several segments. Each segment consists of a *situation*, a *text*, and a set of related *items*. The *situation*, which serves as an advance organizer, prepares test-takers to approach the reading text by helping them access relevant schemata. The *text* refers to the passage that students are asked to read. Selected texts deal with topics such as descriptions of famous persons, vacations, dining, daily routines, etc., and various text types (such as notes, biography articles, advertisements, etc.). The *items* developed for the reading texts are multiple-choice in format and presented in English. (See example in Appendix A.) Each reading instrument includes 35 items, with each item weighted one point. The total possible score on each of the reading assessment instruments, therefore, is 35.

Similar to the reading items, each of the thematically connected writing instruments includes six segments. Each segment has a *situation*, a *warm-up*, and a *task*. The *situation* provides detailed description of the immediate context within which test-takers are asked to compose their responses. Following the *situation*, test-takers are provided with a *warm-up* activity to help them organize their thoughts and language before they start writing. These *warm-up* activities are not scored. The *task* points out what test-takers need to write about. The *task* specifies the relevant content and the required length of the response. The entire writing assessment instrument is presented in English. (See example in Appendix B.) The reader is referred to Chalhoub-Deville (1997) for additional information concerning these instruments.

The rating scheme used to assess students' performance on the writing instruments is included in Appendix C. Raters are presented with a detailed scoring scheme that includes four principal aspects: comprehensibility, task fulfillment, vocabulary, and discourse. Brief descriptions are provided for each of these criteria. Such descriptions are given not only for the required IL level but also for the Novice High and Intermediate Mid levels. The description of the IL and its two adjacent levels is meant to clarify and help raters better focus on the features characteristic of the intended IL level. Raters are also provided with language-specific rating criteria and sample performance examples. In assessing students' L2 writing performance, raters are asked to make dichotomous ratings (1, 0) on each of the above

listed four criteria. The total possible score on each of the writing instruments, therefore, is 24—four points for each of the six writing segments.

Setting Passing Scores

While the reading items and the writing tasks are selected to reflect intermediate level properties according to the ACTFL Guidelines, it is the test-takers' performance on these items and tasks that determines whether examinees have achieved the designated IL level. The question that arises, therefore, is what score on each of the reading and writing instruments is equivalent to an IL performance? More specifically, what scores out of 35 for reading and out of 24 for writing are deemed appropriate for the test-taker to be judged as performing at the ACTFL IL level in each of these two modalities?

Often it is not stated how cut or passing scores are decided upon on L2 tests. Typically, test developers rely on their knowledge and experience to decide on the score or the number of tasks that test-takers must complete successfully in order to pass. Another popular approach is to allow a given percent of students to pass, e.g., the top 20%. In situations where passing rate restrictions are not a factor to consider, such decisions are not appropriate. Why allow the top 20% and not the top 15%, 25%, or 35% of the test-takers to pass? What is to be inferred about the test-takers' abilities when passing scores are decided upon in a relatively arbitrary fashion? A systematic approach to setting passing scores that takes into account issues such as the purpose and content of the instruments and involves the potential test score users is likely to produce meaningful scores and afford a more appropriate use of those scores.

Pass score decisions can be made based on systematic procedures. The manual by Livingston and Zieky (1982), for example, is a classic guidebook that provides practical explanations and descriptions for setting passing score standards relevant to various educational tests. In setting passing scores, several issues need to be considered, including the judges, the standard setting method and corresponding process, and the appropriateness of the derived passing scores. These issues are addressed in the following sections.

The Judges

French, German, and Spanish educators at the University of Minnesota were asked to participate in the standard setting sessions. The educators

represented both the faculty members and graduate teaching assistants who typically teach the language courses. As such, these educators are well-acquainted with the performance expected of students at this level. Additionally, given the University of Minnesota's long tradition with regard to proficiency-based testing, these educators have extensive experience with the ACTFL Guidelines.

Eleven judges, including both faculty and graduate teaching assistants, participated in the standard setting process. Four judges participated in each of the French and German sessions and three in the Spanish one. Typically, in standard setting sessions, the more judges that can be included the better. The rationale for selecting eleven judges in the present study is the similarity in terms of orientation between this group of judges and those they represent. Also, given that the judges in each language group are asked to perform their ratings independently, it can be argued that any derived passing score has been cross-validated.

The Standard Setting Method

Several approaches can be used for setting scores based on evaluations of assessment instrument items. A common approach is based on the borderline test-taker concept where Subject Matter Experts (SMEs) are asked to hypothesize regarding the performance of borderline test-takers on each item in a given assessment instrument. There are different methods that follow this approach. The most common methods are the Nedelsky method (Nedelsky 1954), the Ebel method (Ebel 1972), and the Angoff method (Angoff 1984). The Nedelsky method is used with multiple-choice items only. In this method, the SMEs identify the distracters in each item that a borderline test-taker can clearly recognize as not plausible, and it is assumed that the test-taker would *randomly* identify the correct answer from the remaining plausible options. The Ebel method is more elaborate. It is a two-stage process where SMEs first classify test items into categories, based on the importance and difficulty of each item, and then estimate the proportion of test items a borderline test-taker can answer correctly. The Angoff procedure requires SMEs to provide the probability of a borderline test-taker (or the proportion out of 100 borderline test-takers) being able to respond to every test item correctly. Both the Ebel and the Angoff procedures can be used with non-multiple-choice items. (For more information about these procedures see Livingston and Zieky 1982.)

In determining the method for setting the pass scores on the French, German, and Spanish assessment instruments, three interrelated factors were paramount. First, it was important that the method be easily explained to the judges. In order for the judges to be able to apply the method appropriately and to be confident about using the ensuing pass score, it was critical that the method be readily comprehensible. The second factor pertained to the ease of and the speed by which the procedures can be completed. These procedures can be quite involved and take an inordinate amount of time, and so it was decided to choose a method that permits the completion of the procedures in a relatively short amount of time. Third, given that the instruments include both multiple-choice and detailed scoring criteria, it was necessary to choose a method that can be used for setting pass scores for both types of items. Given that any method requires significant effort on the part of the judges, it was critical not to overburden the judges by requiring them to learn two different methods. Given these considerations, the Angoff method was chosen. As Livingston and Zieky (1982) write, "Angoff's method is the easiest of the three methods to explain and the fastest to use" (p. 54). Additionally, the Angoff method can be used with both multiple-choice items and detailed scoring criteria.

The Standard Setting Process

Independent standard setting sessions were held for each of the French, German, and Spanish groups. The participating judges in these sessions were provided with copies of the assessment instruments. The judges were asked to imagine 100 borderline IL students they are likely to encounter in their L2 classes and to decide on the proportion who are likely to perform successfully on each of the reading items and writing criteria. After a practice exercise, all participating judges independently provided their percentages. Nevertheless, when the researcher noted variation in the judges' percentages of more than ten to fifteen points on any given item, the judges were asked to discuss their reasons for providing such diverse ratings. Consequently, the judges were allowed to change their ratings if they deemed it appropriate.

The passing score was calculated by adding the average proportions and ratings provided by the judges for each item and computing the mean of those averaged ratings. The results of the French, German, and Spanish standard setting sessions for each of the writing and reading assessment

Table 1
Proposed Passing Scores Based on the Angoff Method

	French	German	Spanish
Reading (total possible 35 points)	29.36	28.70	27.01
Writing (total possible 24 points)	20.44	20.72	21.07

instruments are presented in Table 1 above. As Table 1 shows, the passing scores computed for the reading instruments (where the total possible score on each of the three language reading instruments is 35) are 29.36 in French, 28.70 in German, and 27.01 in Spanish. As for the writing instruments (where the total possible score on each of the three language writing instruments is 24), the passing scores are 20.44 in French, 20.72 in German, and 21.07 in Spanish.

Although the three language group judges worked independently, they arrived at quite comparable passing scores for both the reading and writing instruments. Such score comparability may be attributed to the similarity of the instruments in terms of purpose, content, and format across the three languages. Additionally, the background information obtained from the test-takers (see Test-Taker Samples section) yields similar profiles of students in the three language groups, which may also have contributed to the judges' comparable passing score decisions.

While there were no a priori intentions to set identical passing scores across the three languages, the judges, based on the closeness of the derived scores, decided to adopt the same passing scores for each of the reading and writing instruments. (These scores are reported in the following section.) It was reasoned that, given the judgmental nature of the process, it would be more meaningful to present students with the same passing scores than to have to explain the small differences. Nevertheless the appropriateness of such a decision remains to be seen with regard to actual student performances.

Passing Scores and Standard Error of Measurement

When setting passing scores, it is recommended that the standard error of measurement (SEM) be considered (APA 1985). The SEM provides information about potential fluctuations in test scores due to measurement error. Ebel (1979) defines SEM as "an estimate of the standard deviation

of the errors of measurement associated with the test scores in a given set" (p. 379). In the present context, SEM is computed to estimate the variations around the proposed passing scores. Given the criterion-referenced nature of the MNAP instruments, i.e., the focus on IL performance, it was decided to use Berk's (1984 in Bachman, 1990) formula for computing the SEM indices:

$$\sqrt{\frac{x_i\,(n - x_i)}{n - 1}}$$

where x_i is the proposed passing score, and
n is the number of items on the test

Tables 2 and 3 (page 185) present the SEM for various passing scores on the reading and writing assessment instruments respectively. The figures in each table apply to the three language groups, given that each of the French, German, and Spanish modality instruments include an equal number of items and the same proposed passing scores, which is the information needed to compute SEM using Berk's formula. Tables 2 and 3 report estimates for one SEM and two SEM. The one SEM indicates 68% probability that the proposed passing scores fall within the range computed. The two SEM provides with 95% confidence the range expected for the proposed passing scores. Although the tables report the band scores for both one and two SEM, i.e., the 68% and 95% probability, it is argued that given the nature of the test, 68% is sufficiently stringent. Consequently, the present discussion will focus on the 68% probability figures.

With regard to the reading instruments, the figures in Table 2 show that the one SEM at the proposed score 29 is 2.26, which, when rounded to the nearest whole number, yields a band score of 27–31. This indicates that we are approximately 68% confident that the proposed passing score of 29 is between 27–31. Given the lower band of the score, it would be recommended, therefore, that the passing score for the reading instruments be set at 27. Indeed, the language experts in French, German, and Spanish, after some discussion, agreed to set the passing score at 27.

With regard to the writing instruments, Table 3 shows that the one SEM at the proposed score of 21 is 1.66. When rounded to the nearest whole score, the 1.66 SEM gives a band score of 19–23. Although the recommendation would be to set the passing score at 19, the language experts in all three languages felt very strongly that a score of 19 was not equivalent to

IL on these writing instruments. The language expert judges agreed to set the passing score at 21.

Again, the appropriateness of these reading and writing passing scores, empirically derived from expert judgment, must be further examined based on the test-takers' actual performance. The following sections report on the results of the analyses performed on the data obtained from administering the reading and writing instruments to students seeking enrollment in the French, German, or Spanish programs at the University of Minnesota.

Table 2

Confidence Intervals for Passing Scores on the French, German, and Spanish Reading Instruments

Passing Score	1 SEM	Band with 68% Probability	2 SEM	Band with 95% Probability
29	2.26	26.74–31.26	4.52	24.48–33.52
28	2.40	25.60–30.40	4.80	23.20–32.80
27	2.52	24.48–29.52	5.04	21.96–32.04
26	2.62	23.38–28.62	5.24	20.76–31.24
25	2.71	22.29–27.71	5.42	19.58–30.42
24	2.79	21.21–26.79	5.58	18.42–29.58
23	2.85	20.15–25.85	5.70	17.30–28.70

Table 3

Confidence Intervals for Passing Scores on the French, German, and Spanish Writing Instruments

Passing Score	1 SEM	Band with 68% Probability	2 SEM	Band with 95% Probability
21	1.66	19.34–22.66	3.32	17.68–24.32
20	1.87	18.13–21.87	3.74	16.26–23.74
19	2.03	16.97–21.03	4.06	14.94–23.06
18	2.17	15.83–20.17	4.34	15.83–22.34
17	2.27	14.73–19.27	4.54	12.46–21.54

Test-Taker Samples

The reading and writing MNAP instruments were administered to incoming students at the University of Minnesota during summer orientation, 1996. Information regarding the number of students who took the various assessment instruments is summarized in Table 4. The figures indicate a much bigger sample size in Spanish as compared to French and German, which reflects the typically larger enrollment in Spanish programs. Additionally, Table 4 reports two sets of sample sizes for the French and Spanish reading instruments because those who took the reading instruments in August and September received slightly different ones. Based on the results of item analyses performed on the data from the initial set of test administration in June and July, minor revisions were made to six items on the French and three items on the Spanish reading instruments for the August-September testing period.

Table 4
French, German, and Spanish Student Samples

	French	German	Spanish
Reading: June–July	157	na	482
Reading: August–September	85	na	255
Reading: June–September		208	
Writing	240	206	737

Responses to a background questionnaire administered to these test-takers indicate that over 90% of the students were 18 years old or younger. Gender breakdowns indicate that the French sample includes 74% females and 26% males; the German comprises 47% females and 52% males; and the Spanish sample consists of 66% females and 34% males. Over 96% of all the test-takers report that their parents attained at least high school degrees. With regard to academic achievement, over 91% of the test-takers in all three language samples report high school GPAs between 3.00–4.00. Similarly, over 91% of the test-takers in the three samples indicate that the last grade received in their L2 classes was a "B" or better. Finally, approximately 98% of the students in all three languages report the last grade received in an English test to be at least a "B."

As for instruction in L2, 41% of the test-takers in the French sample indicate having had two to three years of French and 55% indicate having

studied it four years or more. In the German sample, 47% state that they had two to three years of the language and 50% had four years or more. In the Spanish sample, while 51% of the students report having had two to three years of the language, 47% indicate having had four years or more. In terms of time elapsed since last enrolled in the L2 class, responses show that for 79% of the French and German and 83% of the Spanish students it had been one year or less. Approximately 15% of the students in all three languages report that two years had gone by since they were last enrolled in an L2 class. Finally, students in the three language groups report having spent minimal time in a community where the L2 is the primary language of communication.

Regarding motivation factors such as the likelihood of taking a second language if not required, students in all three languages were evenly distributed on the five-point scale for each of the three languages (1 = very unlikely, 5 = very likely). A similar trend in responses is observed for students' likelihood of studying the L2 past the language requirement.

Results

Descriptive Statistics

Reading Instruments

For each of the three languages, the total possible score on the reading instrument is 35 points. The means for the two French reading instrument administrations are 24.87 for June–July and 27.06 for August–September with standard deviations of 4.42 and 5.52 respectively. The mean for the German reading instrument is 26.96 with a standard deviation of 4.72. As for the Spanish reading instrument, the mean is 26.79 for June–July and 27.99 for August–September with standard deviations of 4.96 and 4.78 respectively. These figures show that the distribution of scores in all three languages tends to be slightly negatively skewed, as would be expected in criterion referenced assessment where the majority of the students are expected to perform successfully.

The internal consistency reliability, Cronbach's alpha, indices for these reading instruments are as follows: French (June–July) .70 and (August–September) .83; German .77; and Spanish (June–July) .81 and (August–September) .77. These indices show adequate reliability for criterion referenced assessment instruments.

Writing Instruments

Ratings provided a range from 0–24 for French, German, and Spanish with means of 17.69, 18.04, and 19.44 and standard deviations of 6.21, 6.17, and 6.31 respectively. These statistics point out that scores on the writing instruments are negatively skewed with the majority of the students performing successfully on the writing tasks.

Internal consistency indices have also been computed. Cronbach's alpha for each of the French and German samples is .93, and reaches .95 for the Spanish sample, which is quite high. With regard to inter-rater reliability, a random sample of 30 writing performances for each of French, German, and Spanish have been re-rated by a second independent set of raters. Inter-rater reliability for the French sample is .96; .91 for German; and .93 for Spanish, indicating a high level of agreement.

Analysis of Variance Results

In addition to the descriptive analyses, inferential statistics are employed to further investigate the properties of the reading and writing instruments. Analysis of variance (ANOVA) was used to investigate whether the instruments differentiate among test-takers' language proficiency as measured in terms of the number of years of high school study. The following sections summarize the results of the ANOVA analyses.

Reading instruments. One-way ANOVA analyses are performed on each of the three language groups to examine whether there is a significant difference in mean scores on the reading instruments by years of L2 study in school (2 years, 3 years, and ≥ 4 years). In French, analyses are performed on the scores from the 29 items that appeared in the instrument on the test administrations from June through September. In German, analyses are performed on the scores from the 35 items. Spanish analyses are performed on the scores from the 33 items that appeared in the instrument on the test administrations of June through September. (See reduced number explanation under Test-Taker Samples). ANOVA results show significance for each of the three language samples (French: $F(2, 188) = 7.30$; German: $F(2, 161) = 11.09$; Spanish: $F(2, 578) = 81.48$) at $p < .001$.

The Scheffé post-hoc analysis, a very stringent and conservative procedure, is used to examine the statistical significance of all possible pairwise comparisons. With regard to the French group, there is a significant difference between test-takers who have 2 years versus ≥ 4 years of L2

Table 5
Means and Standard Deviations (in parentheses) for the Reading Instruments According to Years of L2 Study in School

	French (n=189; 24 points)	German (n=162; 35 points)	Spanish (n=579; 33 points)
2 years	20.10 (4.11)	24.29 (4.44)	22.36 (5.00)
3 years	20.93 (4.02)	26.30 (4.30)	25.71 (4.46)
4 yrs or >	22.82 (3.70)	28.17 (4.01)	27.96 (3.47)

study and those who have 3 years versus ≥ 4 years. The 2 years versus 3 years pair-wise comparison is not significant. In German the only significant difference is between those who have 2 years of German and ≥ 4 years. Finally, all three pair-wise comparisons of Spanish are significant. Table 5 reports the means and standard deviations for each level of the three languages. It is important to note that the means indicate that the higher the number of years of L2 study, the better the test-takers' performance on the reading instruments, providing evidence to support the desired function of these instruments.

Writing instruments. One-way ANOVA analyses are also performed to examine whether there is a significant difference in mean scores on the writing instruments by years of L2 study in the school. ANOVA results are significant for each of the three language samples (French: $F(2, 164)=28.83$; German: $F(2, 155)= 8.34$; Spanish: $F(2, 586)= 131.44$ at $p < .001$. Scheffé post-hoc analyses indicate a significant difference among all three pair-wise comparisons in French. In German, similar to the pattern observed on the reading assessment instrument, the only significant difference is between 2 years and ≥ 4 years. Also similar to the pattern noted on the Spanish reading instrument, the three Spanish writing pair-wise comparisons are significant. Finally, as observed in Table 6, the higher the number of years of L2 study, the better the test-takers' performance on the writing instruments, again supporting the intended function of these instruments.

In summary, the descriptive analyses provide evidence to support the quality of the instruments. Additionally, the ANOVA and post-hoc analyses indicate that these instruments are, in general, discriminating

Table 6
Means and Standard Deviations for the Writing Instruments According to Years of L2 Study in School

	French (n = 165; 24 points)	German (n = 156; 24 points)	Spanish (n = 579; 24 points)
2 years	10.59 (6.20)	15.29 (6.42)	13.59 (8.07)
3 years	15.18 (6.88)	17.26 (7.19)	19.21 (5.50)
4 yrs or >	19.93 (4.22)	20.01 (4.91)	22.38 (2.59)

among students with varying years of L2 study in school. Finally, test-takers who have more years of L2 study exhibit better performance on these instruments.

Passing Rates

Given the judgmental nature of the procedure for setting cut scores, it is critical to also examine these scores based on the actual performance of the test-taker groups for which the instruments are intended. Table 7 shows the passing rates at various cutoff scores, including the proposed Angoff-based passing score. The figures indicate that 55% of the French and German students and 64% of the Spanish students would pass the reading assessments at the proposed passing score of 27. As would be expected, the lower the passing score, the greater the students' passing rate.

With regard to the writing instruments, Table 8 shows that 45%, 47%, and 66% of the French, German, and Spanish students respectively would pass the writing assessments at the proposed passing score of 21. Also, similar to the reading figures, the percentage of passing increases with

Table 7
Passing Rates on the Reading Assessment Instruments

	Various Proposed Passing Scores				
	27	26	25	24	23
French (n = 73)	55%	67%	70%	73%	81%
German (n = 199)	55%	64%	72%	79%	83%
Spanish (n = 215)	64%	72%	77%	80%	86%

Table 8
Passing Rates on the Writing Assessment Instruments

	Various Proposed Passing Scores				
	21	20	19	18	17
French (n = 240)	45%	52%	60%	64%	68%
German (n = 206)	47%	54%	59%	64%	70%
Spanish (n = 737)	66%	70%	73%	77%	79%

lower passing scores as cutoff points. In deciding on the appropriateness of the passing score, an important variable to consider further is the number of years of studying the L2 in high school. This variable, as the data below shows, has proven to be more important regarding passing rates than the different passing scores.

Passing Rates and Years of Studying the L2

The percentages of test-takers passing the reading and writing instruments according to the number of years of studying L2 in school are presented in Tables 9 and 10 respectively. For both reading and writing, figures show relatively minimal change in passing rates across the various passing scores, especially at the first two levels of L2 study. This trend is observed across the three language groups. The striking change in passing percentage occurs when looking across the different years of studying the L2, and in particular at four years or more.

Table 9
Passing Rate on the Reading Assessment Instruments: Years of L2 Study in School by Various Cutoff Scores

Passing Score	French (n = 55)			German (n = 175)			Spanish (n = 175)		
	2 yrs	3 yrs	4 yrs or >	2 yrs	3 yrs	4 yrs or >	2 yrs	3 yrs	4 yrs or >
27	2%	9%	40%	6%	14%	32%	6%	15%	38%
26	6%	9%	47%	8%	15%	37%	10%	15%	42%
25	6%	9%	49%	11%	17%	39%	12%	19%	43%
24	6%	9%	53%	13%	20%	43%	13%	20%	43%
23	6%	16%	55%	13%	21%	46%	15%	21%	45%

Table 10
Passing Rate on the Writing Assessment Instruments: Years of L2 Study in School by Various Cutoff Scores

Passing Score	French (n = 198)			German (n = 162)			Spanish (n = 607)		
	2 yrs	3 yrs	4 yrs or >	2 yrs	3 yrs	4 yrs or >	2 yrs	3 yrs	4 yrs or >
21	1%	11%	32%	4%	13%	28%	6%	17%	41%
20	1%	13%	36%	6%	15%	32%	7%	18%	43%
19	1%	13%	43%	7%	15%	36%	8%	19%	44%
18	1%	14%	47%	9%	15%	38%	9%	20%	45%
17	2%	14%	49%	11%	18%	41%	10%	21%	46%

The figures in both the reading and writing tables show prominent increases in the percentage of passing for those students who have studied the L2 for three years versus two. The dramatic increase in the percentage of passing occurs, however, for those who have studied the L2 for four years or more. In short, the figures in Tables 9 and 10 show that passing scores, while important, are not as critical for passing the reading and writing assessment instruments as the number of years the test-takers have studied the L2. The findings of the present study send a strong message to teachers, students, counselors, administrators, and parents regarding students' preparation in L2 in the schools. Based on the current findings, students are more likely to be judged as performing at the IL level on the reading and writing instruments if they have studied French, German, or Spanish for at least four years.

Two issues need to be considered with regard to the appreciable increase in passing rates for those who have studied L2 for four years or more. First, at an advanced level of instruction self-selection becomes a confounding variable. In other words, it could be that the more proficient students are more likely to continue their L2 study. Second, it is also important to consider a potential threshold effect. It may be that the performance of students who have had four or more years of L2 instruction reflects a significant increase in proficiency level because students with two or three years of L2 study are invested more in restructuring and consolidating their L2. It may be that only after four years of study that progress in L2 proficiency becomes evident.

In conclusion, the passing score, although an important factor in determining passing rates, is not as critical a factor in substantially increasing the percentage of students likely to perform well on these instruments. The number of years of L2 study in school proves to be a more crucial variable in raising the passing percentage. Finally, with regard to the instruments themselves, the higher passing rates for those students who have had more years of L2 study provide additional evidence to support the adequate functioning of the present assessments.

Conclusion

The purpose of this paper is to report the results of the analyses performed to investigate the properties of the MNAP reading and writing instruments currently used at the University of Minnesota to admit students into second-year French, German, and Spanish language courses. The present findings provide evidence to support the appropriateness of these instruments for their intended use. Nevertheless, more work is needed, especially in the following areas.

First, support for the continued refinement of these instruments is needed. Bernhardt and Deville (1991) argue forcefully that language departments are mandated to maintain and continue the development of their testing programs by allocating the necessary monetary and human resources. Bernhardt and Deville write that "without such an investment, a testing program does not exist. What does exist is a set of trials for students to survive" (p. 58). Similarly Cumming and Berwick (1996) argue that validation is a long-term process that leads to "ongoing modifications of test instruments, the construct, and the conceptual framework" underlying those instruments (p. 5). In short, the initial work performed on these assessment instruments is not sufficient to ensure their continued validity and reliability properties. A financial and human resource commitment is required to continue the research and development agenda necessary to monitor and document how these instruments are functioning. Additionally, such research can inform our understanding of the L2 proficiency constructs operationalized in these instruments.

Second, the passing scores need to be revisited. These instruments are intended to be used not only by the University of Minnesota, but also by all the MNAP institutions. Therefore, the University of Minnesota proposed passing scores need to be revisited with the MNAP members. Given the relatively diverse student populations that the various MNAP

members deal with in their institutions (secondary, postsecondary, private, public), it is possible that the members may arrive at slightly different passing scores. Additionally, with the more diverse student population, it would be interesting to examine how various passing scores compare to the number of years of L2 study with regard to passing rates. Such data have been collected and are currently being analyzed.

Also related to passing standards is the designated IL level. An issue that the University of Minnesota and other MNAP L2 educators will have to address is whether to employ the same IL standards over time. It is important to note here that given the changes envisioned in terms of student L2 preparation in the various educational systems, secondary and postsecondary, it is reasonable to assume that over time students taking the MNAP instruments will likely attain higher L2 proficiency. As a result, changes not only in the passing scores but also in the designated IL level may need to be revisited.

Third, a frequently encountered conception among MNAP educators, including those at the University of Minnesota, is that three years of language study at the secondary school level is on average equivalent to one year—two semesters or three quarters—at the postsecondary level. Such a belief is not restricted to MNAP members. Other educators and institutions have advanced variations on this rule (see Lange et al. 1992; Wherritt, Druva-Roush, and Moore 1990).

Additionally, according to an article in *Education Week* (Hendrie, November 26, 1997), New York state voted on November 14, 1997 to include a three year foreign language requirement (and passing a state exam) to its high school diploma. Other states, according to the same article, including Indiana, Louisiana, South Carolina, Texas, and Virginia, also require three years of an L2 study for a "special, more advanced diploma" (Hendrie, November 26, 1997, p. 13). In a subsequent article (Hendrie, December 3, 1997), the NY Regents decided to "rescind its earlier action and mandate that level of study [three years] only for students receiving an advanced diploma" (p. 3). The chancellor of the state board commented that the message from the schools was that the Regents "had really gone too far, too quickly" (p. 3). Notwithstanding, the three-year L2 requirement in the schools is quite popular and is believed by some to be equivalent to one year at the postsecondary level. The results of the present study, however, do not support such beliefs or findings. The present findings indicate that, in general, students need at least four years of L2 instruction at the school

level in order to enroll in a second-year L2 class at a postsecondary institution such as the University of Minnesota.

While it is beyond the scope of the present study to compare incoming student performance with that of students currently enrolled at the University of Minnesota, future research is planned to address this question. Again, such research is necessary as part of the continued investigation of the qualities of the present instruments.

Finally, Suen (1990) and Kane (1992) emphasize the importance of examining the validity of the decisions made based on assessment instruments with set passing scores. In the present context, this means documentation is needed to validate the decisions of admission or no admission into second-year French, German, and Spanish language courses based on the passing scores. Although data is not currently available to investigate the appropriateness of these decisions, anecdotal evidence from the Directors of Language Instruction (DLI) in the three language programs and from the Testing Office indicates that minimal migration has occurred since the implementation of these assessment instruments. In other words, students have, for the most part, been accurately admitted or denied admission into the second-year L2 courses. The DLIs and Testing Office personnel expressed satisfaction with the way the instruments have functioned. Nonetheless, further research is needed to investigate this issue.

Works Cited

American Council on the Teaching of Foreign Languages (ACTFL). 1986. *ACTFL Proficiency Guidelines.* Hastings-on-Hudson, NY: ACTFL.

American Psychological Association (APA). 1985. *Standards for Educational and Psychological Testing.* Washington, DC: American Psychological Association.

Angoff, William H. 1984. *Scales, Norms, and Equivalent Scores.* Princeton, NJ: Educational Testing Service. (Originally published in Robert L. Thorndike ed., *Educational Measurement,* 2nd ed. (Washington, DC: American Council on Education, 1971).

Bachman, Lyle. 1990. *Language Testing in Practice.* Oxford: Oxford University Press.

Barnes, Betsy, Carol Klee, and Ray Wakefield. 1990. A Funny Thing Happened on the Way to the Language Requirement. *ADFL Bulletin* 22: 35–39.

Bernhardt, Elizabeth, and Craig Deville. 1991. Testing in Foreign Language Programs and Testing Programs in Foreign Language Departments: Reflections and Recommendations. In *Assessing Foreign Language Proficiency of Undergraduates,* edited by Richard Teschner, 43–59. AAUSC Issues in Language Program Direction. Boston, MA: Heinle & Heinle Publishers.

Birchbichler, Diane, Kathryn Corl, and Craig Deville. 1993. The Dynamics of Placement Testing: Implications for Articulation and Program Revision. In *The Dynamics of Language Program Direction,* edited by David P. Benseler. AAUSC Issues in Language Program Direction. Boston, MA: Heinle & Heinle Publishers.

Chalhoub-Deville, Micheline. 1997. The Minnesota Articulation Project and Its Proficiency-Based Assessments. *Foreign Language Annals* 30: 492–502.

Corl, Kathryn, Linda Harlow, Jan Macian, and Donna Saunders. 1996. Collaborative Partnerships for Articulation: Asking the Right Questions. *Foreign Language Annals* 29: 111–24.

Cumming, Alister, and Richard Berwick. 1996. The Concept of Validation in Language Testing. In *Validation in Language Testing,* edited by Alister Cumming and Richard Berwick, 1–14. Bristol, PA: Multilingual Matters Ltd.

Ebel, Robert L. 1972. *Essentials of Educational Measurement.* 2nd ed. Englewood Cliffs, NJ: Prentice-Hall.

_____. 1979. *Essentials of Educational Measurement.* 3rd ed. Englewood Cliffs, NJ: Prentice-Hall.

Fleak, Ken. 1991. Using an Exit Requirement to Assess the Global Performance of Undergraduate Foreign Language Students. In *Assessing Foreign Language Proficiency of Undergraduates,* edited by Richard Teschner, 115–134. AAUSC Issues in Language Program Direction. Boston, MA: Heinle & Heinle.

Hendrie, Caroline. "N.Y. Students Must Master Second Language." *Education Week XVII* November 26, 1997: 1,13.

_____. "N.Y. Regents to Drop Foreign-language Requirement." *Education Week XVII* December 3, 1997: 3.

Kane, Michael. 1992. An Argument-Based Approach to Validity. *Psychological Bulletin* 112: 527–535.

Lange, Dale. 1987. Developing and Implementing Proficiency-Oriented Tests for a New Language Requirement at the University of Minnesota: Issues and Problems for Implementing the ACTFL/ETS/ILR Proficiency Guidelines. In *Proceedings of the Symposium on the Evaluation of Foreign Language Proficiency*, edited by Albert Valdman, 275–90. Bloomington, IN: Indiana University.

Lange, Dale, Paul Prior, and William Sims. 1992. Prior Instruction, Equivalency Formulas, and Functional Proficiency: Examining the Problem of Secondary School-College Articulation. *The Modern Language Journal* 76: 284–94.

Livingston, Samuel A., and Michael J. Zieky. 1982. *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests.* Princeton, NJ: Educational Testing Service.

Mosher, Art. 1989. The South Carolina Plan for Improved Curriculum Articulation Between High Schools and Colleges. *Foreign Language Annals* 22: 157–62.

Nedelsky, Leo. 1954. Absolute Grading Standards for Objective Tests. *Educational and Psychological Measurement* 14: 3–19.

Robinson, Deborah W. 1996. *The Collaborative Foreign Language Articulation and Assessment: Training Manual.* Columbus, OH: The Ohio State Foreign Language Center.

Suen, Hoi K. 1990. *Principles of Test Theories.* Hillsdale, NJ: Lawrence Erlbaum.

Teschner, Richard. 1991. Introduction. In *Assessing Foreign Language Proficiency of Undergraduates*, edited by Richard Teschner, ix-xii. AAUSC Issues in Language Program Direction. Boston, MA: Heinle & Heinle Publishers.

Wherritt, Irene, Cynthia Druva-Roush, and Joyce Moore. 1990. The Development of a Foreign Language Placement System at the University of Iowa. In *Assessing Foreign Language Proficiency of Undergraduates,* edited by Richard Teschner, 79–92. AAUSC Issues in Language Program Direction. Boston, MA: Heinle & Heinle Publishers.

Appendix A

French

Sample 1.

A classmate shows you this postcard that her family received from some French-speaking friends vacationing in Mexico.

Salut les amis!

Un petit bonjour du Mexique. En ce moment, on est à Mérida, une petite ville sympa. Cet après-midi on a visité une cité Maya. C'était très beau. Ils étaient extraordinaires, les Toltèques. Sinon, tout va bien, sauf un petit coup de soleil. Demain, on va reprendre le bus pour aller voir une réserve d'oiseaux au bord de la mer. Ensuite, je pense qu'on va revenir vers les plages de l'autre côté du Yucatan. Bon, vous voyez, ça se passe bien.

Grosses bises,

Ramon et Félicité

1. Where were Ramon and Félicité when they wrote this postcard?
   a. At a monument
   b. On a beach
   c. In a city
   d. On a bus

2. What will Ramon and Félicité visit next?
   a. Mayan ruins
   b. An animal preserve
   c. A sun god temple
   d. An island off the Yucatan peninsula

Appendix B

Segment 2: A visitor

Situation: For your next entry, your teacher would like you to write some questions for a student from a French/German/Spanish-speaking area who will be coming to visit your class soon. The class has an opportunity for a question-and-answer session with the visitor and your teacher wants the class to be well prepared with questions.

Warm-up: Think about what you want to ask the French/German/Spanish-speaking visitor, then respond below in French/German/Spanish or in English. You may want to ask about the climate, interesting places to visit, about what young people do for work and entertainment (i.e., music, food, going out, etc.).

Things you want to know:

Task: In your journal, write at least five questions for the French/German/Spanish-speaking student who is coming to your class. You might want to include questions about: (1) the climate; (2) what young people do for work and entertainment; (3) interesting places to visit, etc.

Write at least five questions for the visiting student in French/German/Spanish.

Appendix C
Assessment Team: Rating Criteria for Entrance Proficiency Writing Test

- NH **Novice-High**	**Intermediate-Low**	**Intermediate-Mid** + IM
Novice High level examinees complete the tasks below the intended level. A NH performance will not sustain some of the necessary features of the Intermediate-Low profile. It will lack the quality and the quantity of the IL performance.	Examinees at the Intermediate-Low level adequately complete the writing tasks. Communication is successful, although there are errors.	An Intermediate-Mid level performance is above the required level
COMPREHENSIBILITY FREQUENTLY INCOMPREHENSIBLE • errors in spelling, grammar, punctuation, vocabulary, and frequent lapses into non-target language interfere with comprehensibility for a sympathetic reader	COMPREHENSIBILITY GENERALLY COMPREHENSIBLE • errors in spelling, grammar, punctuation, vocabulary, and occasional lapses into non-target language do not interfere with comprehensibility for a sympathetic reader	COMPREHENSIBILITY ALMOST ALWAYS COMPREHENSIBLE • errors in spelling, grammar, punctuation, vocabulary, and rare lapses into non-target language do not interfere with comprehensibility for a sympathetic reader
TASK FULFILMENT TASK DEMANDS UNFULFILLED • response is not appropriate to the task as it is specified in bold on the test • amount of writing is insufficient to meet task requirements	TASK FULFILLMENT TASK DEMANDS ADEQUATELY FULFILLED • response to task, as specified in bold on the test, is appropriate • number of sentences/clauses is sufficient to meet task requirements	TASK FULFILLMENT TASK DEMANDS SURPASSED • response elaborates beyond task requirements as specified in bold on the test • amount of writing is greater than needed to meet task requirements

- NH Novice-High		Intermediate-Low		Intermediate-Mid + IM	
VOCABULARY DOES NOT DEMONSTRATE ADEQUATE SCOPE FOR TASK • does not show ability with target language in relation to the topic • inability to go beyond memorized phrases • word choices are repetitive		VOCABULARY DEMONSTRATES ADEQUATE SCOPE FOR FOR TASK • sufficiently broad vocabulary to attempt to create with target language in relation to the topic • ability to go beyond memorized phrases • word choices are minimally repetitive		VOCABULARY DEMONSTRATES GREATER SCOPE THAN REQUIRED FOR TASK • breadth of vocabulary goes beyond task requirements; writer creates with target language • goes beyond memorized requirements • word choices are minimally repetitive	
BELOW SENTENCE-LEVEL PRODUCTION • does not sustain basic sentence-level production beyond a few memorized patterns; resorts to list and fragments • can use present tense of some common verbs; sometimes uses infinitives for conjugated verbs		SENTENCE-LEVEL PRODUCTION • overall, sample shows ability to write basic sentence-level discourse (S-V-C) • can express present and future time (with compound future or adverbs of time); generally cannot use past tenses. • sentence patterns are minimally repetitive		SENTENCE + LEVEL PRODUCTION • shows ability to go beyond basic sentence structure, with attempts to link sentences • can use: 1) present tense of most regular and some common irregular verbs; 2) the compound future; 3) sporadically the *passé composé* of high-frequency verbs • emerging ability to vary stylistic features	

0 ——— Score ——— 1

UNRATABLE SAMPLE
NOT ENOUGH TO EVALUATE; 5 SIMPLE SENTENCES OR LESS