DOCUMENT RESUME

ED 414 291 TM 027 720

AUTHOR	Sireci, Stephen G.; Foster, David F.; Robin, Frederic; Olsen, James
TITLE	Comparing Dual-Language Versions of an International Computerized-Adaptive Certification Exam.
PUB DATE	1997-03-26
NOTE	37p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, March 24-27, 1997).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Adaptive Testing; Adults; *Certification; Comparative Analysis; *Computer Assisted Testing; Cultural Differences; English; German; International Education; *Item Bias; *Languages; Licensing Examinations (Professions); Test Bias; *Test Format; Test Validity

ABSTRACT

Evaluating the comparability of a test administered in different languages is a difficult, if not impossible, task. Comparisons are problematic because observed differences in test performance between groups who take different language versions of a test could be due to a difference in difficulty between the tests, to cultural differences in test taking behavior, or to a difference in proficiency between the language groups. The international certification testing programs conducted by Novell, Inc. are exceptional examples of the complex psychometric demands inherent in multiple language assessment programs. Novell's international certification program includes tests administered in 12 languages. Many of these tests are computerized adaptive (CAT), complicating comparisons across tests and languages. This paper reports the results of a study comparing English and German language versions of a high-stakes Novell CAT certification exam. The two versions of the test were compared by analyses including separate and concurrent item response theory calibrations. Results with 1,668 English-language candidates and 922 German-language candidates indicate that the English and German CATs are highly similar, and that the tests appear to be unidimensional in both the English and German versions. It is also concluded that the German candidate sample was more proficient than the English sample, and that 2 of 15 items functioned differentially across the languages. The source of the differential item functioning was identified post hoc using bilingual subject matter experts. The comparability of the passing scores, and other critical validity issues are discussed. (Contains 4 tables, 11 figures, and 20 references.) (Author/SLD)

Running Head: COMPARING DUAL-LANGUAGE CATs

Comparing Dual-Language Versions of an International Computerized-Adaptive Certification Exam

Stephen G. Sireci
University of Massachusetts, Amherst

David F. Foster.
Novell, Inc.

Frédéric Robin
University of Massachusetts, Amherst

James Olsen
Alpine Media

Paper presented at the annual meeting of the National Council on Measurement in Education as part of the symposium "Large-Scale Test Adaptation Projects: Designs, Results, and Suggestions for Improving Practice," March 26, 1997, Chicago, Illinois.

Comparing Dual-Language Versions of an International Computerized-Adaptive Certification Exam

Abstract

Evaluating the comparability of a test administered in different languages is a difficult, if not impossible, task. Such comparisons are problematic because observed differences in test performance between groups who take different language versions of a test could be due to a difference in difficulty between the tests, cultural differences in test taking behavior, or to a difference in proficiency between the language groups. The international certification testing programs conducted by Novell are exceptional examples of the complex psychometric demands inherent in multiple-language assessment programs. Novell's international certification program includes tests administered in twelve different languages. Many of these tests are computerized adaptive, complicating comparisons across tests and languages. This paper reports the results of a study comparing English- and German-language versions of a high-stakes computerized adaptive Novell certification exam. The two versions of the test were compared by a series of analyses including separate and concurrent IRT calibrations. The results indicated that the English and German computerized adaptive tests are highly similar, and that the tests appear to be unidimensional in both the English and German language versions. It was also concluded that the German candidate sample was more proficient than the English sample, and that two out of fifteen items functioned differentially across the two languages. The source of the differential item functioning was identified post hoc using bilingual subject matter experts. The comparability of the passing scores, and other critical validity issues are discussed.

Introduction

Novell Inc. develops some of the most popular computer networking software in the world. This software is critical for managing today's business environments that rely on sophisticated computer technology. Because computer networks must be operated, supervised, and managed by professionals who are competent with respect to the newest network software technology, there is an increasing need to certify individuals for competence in several areas of network operations. To meet this demand, Novell developed an extensive international certification program involving over 1,100 testing centers, 400 of which are outside of North America. For many individuals who work with computer networks, Novell certification makes an enormous difference in regard to salary, promotion, and marketability. Novell's certification program is high-stakes, and plays an important role in ensuring the successful operation of many organizations.

There are several admirable features of Novell's certification program that enable it to survive in an international, constantly changing marketplace, while simultaneously maintaining psychometric standards of test quality. These features are described in detail by Foster, Olsen, Ford, and Sireci (1997). True to form, all of Novell's certification exams are computerized, and the majority are computerized adaptive tests (CAT). These tests contain a mix of multiple-choice and authentic assessment items. The authentic assessment items involve completing tasks on the network that are commonly performed by network administrators. Thus, Novell's tests are conspicuously job-relevant and content-valid.

To meet the needs of its international customers, Novell currently administers its certification tests in 12 different languages. Foreign language testing accounts for about 21% of all Novell certification testing; typically, over 4,000 foreign language exams are administered per month. The availability of the tests in foreign languages allows networking skills to be measured in a manner that is not confounded by English language proficiency. Given the high demand for these foreign language tests, and the short time available for producing tests linked to current software, it is not feasible to develop unique item banks in all languages. Instead, Novell pilots and calibrates all test items in English and translates them into the other relevant languages. The item translation process is comprehensive, and the end result is translated CAT item pools that are parallel to the original English-language item pools. This practice has resulted in the successful implementation of Novell's certification CATs throughout the world.

A unique feature of Novell's foreign language tests is that, by "clicking" on an icon, a candidate taking a foreign-language version of a test can access the English-language version of any test item. This feature is helpful for those candidates who are somewhat proficient in English or who may be more familiar with the English versions of some computer terms.

Although Novell's certification program is a modern-day success story, little research has been done to evaluate the comparability of their certification tests across languages. It is widely known that when a test is translated from one language to another unintended differences in the difficulty or clarity of the test can be introduced (Geisinger, 1994; Hambleton, 1993; Sireci, 1997b). To what degree may such unintended differences be affecting Novell's tests? Are the different language versions of Novell's tests comparable across languages? Is the level of proficiency required to pass the test the same across languages? Initial attempts to answer these questions are provided in this study. Specifically, we compare the comparability of the English- and German-language versions of one of Novell's most popular exams. To conduct this evaluation, we use a variety of strategies ranging from response time analyses to analyses of differential item functioning.

Method

Instrument

The CATs evaluated here are drawn from a pool of approximately 100 test items. Before becoming operational, all items in the pool were screened and calibrated using the three-parameter logistic item response theory (IRT) model (Lord & Novick, 1968; Hambleton, Swaminathan, & Rogers, 1991). Following standard procedures at Novell, the tryout statistics and IRT item parameters were calibrated from comprehensive field tests of candidates who take the test in English. Candidates have 30 minutes to complete the exam, but the vast majority finish in under 20 minutes. Candidates must answer a minimum of 15 test items; the maximum number of items administered is 25. All items are scored dichotomously. The CAT algorithm for these tests includes a content balancing feature to ensure that the candidates are tested on all of the relevant features of the software. All items in the pool were originally developed in English and were subsequently translated into German using local bilingual translators in Germany. Bilingual test specialists at Novell supervised and evaluated the translations in accordance with the recent guidelines for adapting tests proposed by the International Test Commission (Hambleton, 1994). About 20% of the items are performance-based "simulation" tasks involving little verbiage.

Subjects

The data for this study come from 1,668 candidates who took the English language version of the test between September and December 1996, and 922 candidates who took the test in German during this same period. For the item response time analyses, item exposure comparison, and content area exposure comparison, the data from all candidates were used. However, due to small sample sizes in many cells of the examinee-by-item matrix (a common feature of CAT administrations) a subset of candidates was used for the IRT analyses. The extraction of this subset is described below. The passing rates on the exam for those candidates comprising the final subsets were 76.0% and 73.5% for the English and German samples, respectively. These passing rates are higher than those observed for both language groups in 1996, which were 60% and 61% for the English and German candidates, respectively.

Data Analyses

CAT Functioning

To evaluate the functioning of the CAT in its English and German versions, the proportions of candidates in each language who took each item were compared (item exposure comparison). The average numbers of items from each content area that were administered in each language were also compared. In addition, the distribution of scaled scores for candidates who passed, failed, and received an incomplete score (i.e., ran out of time) were compared across the English and German groups. The average number of items taken by the U.S. and German candidates was also compared. These analyses provided a general inspection of the functioning of the CAT across the two languages.

Response time analyses

The amount of time it took to answer the English- and German-language version of each item was compared (response time comparisons). These comparisons looked at overall response times, as well as the response times for those candidates who answered the item correctly, and those who answered it incorrectly. In addition to the average response time, the distributions of response time were compared across language groups.

Dimensionality analyses

The development of Novell's certification tests assumes that a unidimensional IRT model is sufficient for scaling candidates' responses to the test questions. In addition, only one score is calculated for candidates. Thus, it is assumed that these tests are unidimensional. To evaluate this assumption, the dimensionality of both the English- and German-language CATs were evaluated using multidimensional scaling (MDS). MDS is an appropriate method for evaluating the dimensionality of test items because it does not assume a linear relationship among test items (similar to the IRT model) and is relatively easy to use and interpret (Chen & Davison, 1996; De Ayala & Hertzog, 1991; Sireci, 1997a). To apply MDS to these data, the association between two test items was calculated using the pseudo paired comparison (PC) statistic proposed by Chen and Davison (1996). The PC statistic is derived from the two joint probabilities that can occur when a person passes one item in a pair and fails the other item (i.e., passing item *j* and failing item *k* or vice versa). Thus, these probabilities are conditional on passing one, and only one, of the two items in the pair. The PC statistic can be directly related to the item difficulty parameters (b-parameters) of the one-parameter (Rasch) IRT model:

$$\pi_{jk} = \frac{\exp(b_k - b_j)}{1 + \exp(b_k - b_j)} \qquad (1)$$

where π_{jk} is the PC (association) statistic between items *j* and *k*, and b_j and b_k are the difficulty parameters from the Rasch model (Chen & Davison, 1996, p. 312). The sample estimate of π_{jk} is computed by dividing the number of examinees who passed item *j* and failed item *k* by the number of examinees who passed only one of the two items. This sample index is not symmetric across *j* and *k*, and so each is reduced by .50 to form the symmetric dissimilarity index used as the input for the MDS analysis.

Due to the nature of CAT administrations, all candidates did not respond to all items in the pool. In fact, there were several cells in the candidate-by-item matrix that had extremely small sample sizes. Therefore, the dimensionality analyses were carried out using a subset of 16 items to which at least 300 German and English candidates responded. These 16 items spanned the most popular content areas. However, they did not span the entire difficulty range. The majority of the items were of easy-to-moderate difficulty, which explains why they were the most frequently administered. However, because the passing score is set within this range, the predominance of easier items was not considered a major threat to the internal validity of the study. Dissimilarity indexes were computed for all 120 item pairings involving the 16 studied items. Separate indexes were computed for the English and German data sets.

Chen and Davison (1996) recommend fitting one-dimensional and two-dimensional MDS models to the matrix of item dissimilarities and comparing the results. This comparison is qualitative; however, it is linked directly to theories describing the behavior of unidimensional and multidimensional test data. If a set of test items are unidimensional, the unidimensional MDS solution should demonstrate good fit to the data, and a "U" or "C"-shaped pattern should appear when two dimensions are fit to the data (a well known characteristic of overfitting unidimensional data using MDS). Both one and two-dimensional MDS solutions were fit to the data for both the English and German data. The fit and interpretability of these solutions was compared across the two language groups.

IRT Analyses

Comparing theta estimates based on English and German parameters

Due to the sample size requirements necessary for IRT analysis, and due to the sparse candidate-by-item matrix, the same 16 item-subset used in the dimensionality analyses was used in the IRT analyses. In addition, only candidates who responded to at least eight of these items were included. Thus, the data sets used in the IRT analyses comprised item responses from 965 English (U.S.), and 505 German candidates. The MULTILOG calibration software (Thissen, 1991) using the marginal maximum likelihood method for computing item and person parameters (Bock & Aitkin, 1981) was used for all analyses. Using the three-parameter logistic IRT model (3PL), the relationship between candidates facility with Novell's software (i.e., theta) and the three parameters used to describe the item is given by the equation:

$$P_i(\theta) = c_i + (1-c_i)\,\frac{e^{D_{a_i}(\theta-b_i)}}{1+e^{D_{a_i}(\theta-b_i)}}\ (i = 1,2,\ldots n). \qquad (2)$$

where $P_i(\theta)$ is the probability that a randomly selected candidate with facility θ will answer item *i* correctly, b_i is the difficulty parameter for the item (i.e., location of the item on the θ scale), a_i is the discrimination parameter of the item (i.e., proportional to the slope of $P_i(\theta)$ at the point $\theta = b_i$), c_i is the "guessing" parameter, which represents the probability that candidates with extremely low ability can answer item i correctly (i.e., the lower asymptote of the curve), and n is the number of items in the pool (Hambleton, et al., 1991). The two-parameter logistic model (2PL) sets c_i equal to zero, which assumes that guessing is not an important variable for explaining candidates' responses to the items. The 2PL is particularly applicable to the items calibrated in this study because the multiple-choice items typically involved a large number of response options (usually six or seven) and the simulation items required constructed responses.

To evaluate whether the use of item parameter estimates calculated from English candidates was appropriate for estimating scores (thetas) for the German candidates, a series of IRT analyses was conducted. The first set of analyses compared the fit of the two- and three-parameter logistic IRT models to the data. Subsequently, three sets of item parameter estimates were calibrated. The first analysis estimated the item parameters using the data from the German candidates. The second analysis estimated the item parameters using the entire sample of U.S. candidates. The third analysis estimated the item parameters using a random sample of U.S. candidates that was equal to the sample size of the German candidates (to evaluate the sample size effect). These three sets of item parameters were then used to estimate separate IRT scaled scores (thetas) for the Germans. The similarity of these separate theta estimates, and their impact on passing the test, were compared.

Evaluating differential item functioning

IRT analyses were also used to search for differential functioning between translated versions of an item. If the English and German version of an item can be modeled using the same item parameters, then the item may be considered to function similarly across languages. To evaluate the functioning of the items across the two languages, the IRT likelihood ratio (LR) procedure for detecting differential item functioning (DIF) was used (Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988, 1993; Wainer, Sireci, & Thissen, 1991). An important feature of the marginal maximum likelihood method for fitting an IRT model is that a likelihood index is provided for each model fit to the data. This index can be used to compare the fit of different hierarchical models applied to the same data. The likelihood ratio (LR) chi-square statistic (G^2) compares the difference in fit to the data between two hierarchical models as

$$G^2(d.f.) = 2 \log \left[\frac{\textit{likelihood of more general model}}{\textit{likelihood of more constrained model}}\right] \qquad (3).$$

The G^2 statistic is distributed as chi square, with degrees of freedom equal to the difference between the number of parameters estimated in each model. A more general (less constrained) model will always fit the data better than a more constrained model (i.e., the more parameters fit to the data, the better the data are modeled). The G^2 statistic tests whether the improvement in fit is statistically significant. Incorporating constraints into a general model yields parsimony, and so constrained models are preferred whenever the improvement in fit of a more general model is not statistically significant (Thissen, Steinberg, & Gerrard, 1986; Thissen, et al., 1988, 1993).

Chi-square LR tests were used to determine if the English and German versions of the items were equivalent. To conduct these "translation DIF" tests, an "ALL DIF" model was initially fit to the data. This model calibrated separate parameters for all the English items and their German counterparts. Thus, it was the most general model fit to the data. The mean and standard deviation of the theta scale is arbitrary in an IRT model and so these values must be set for at least one of the groups. To define the scale, the mean theta for the Germans was set to zero and the standard deviation of both groups was set to one. The chi-square LR statistic for this model was compared to those values obtained from fitting constrained versions of this general model. The first constrained model fit to the data was the "NO DIF" model that calibrated a common set of parameters for the English items and their German counterparts. This highly constrained model represented item equivalence across languages. Intermediate models between these two extremes were also fit to the data to evaluate the differential functioning of specific items. The LR approach has been widely applied to the investigation of DIF across subgroups who take a test in a single language (Thissen, et al., 1988, 1993; Thissen, et al., 1986; Wainer, 1995; Wainer, et al., 1991), and has recently been applied to the translation DIF problem (Sireci & Berberoglu, 1997).

Collateral Item Information

A problem in evaluating DIF across translated versions of an item is that differences in proficiency between the separate language groups, and differences in item difficulty due to translation or other factors, may be confounded. If the location or dispersion of the proficiency distributions for the two groups are different, items that function differentially across languages may appear statistically equivalent, or items that are really equivalent may appear to function differently. This problem can occur if the differences between the proficiency distributions of the two language groups are not accounted for in the IRT model and a systematic difference in difficulty is present in the majority of translated items (Sireci, 1997b; Sireci & Swaminathan, 1996). To provide an estimate of the "true" difference between the means of the proficiency distributions for the English and German groups, collateral information on the items was used to identify items which were most likely to be equivalent across the two languages. First, a group of expert English-German bilingual translators, proficient with Novell's software, were asked to evaluate the 16 items studied and rank the "best eight" in terms of equivalence across English and German. Second, those items that were simulation items involving little text were identified. Third, items that exhibited similar response time distributions for both language groups were identified. It was hypothesized that these three criteria could be used to identify a subset of items appropriate for anchoring the responses for both groups onto a common scale.

Results

Item Exposure and Content Area Comparisons

The item exposure proportions were calculated by dividing the number of times an item was administered by the total number of candidates in each language group. For each item, a separate proportion was calculated for the U.S. and German candidates. Figure 1 displays a scatter plot of these proportions for each group. Although there is some variability across languages, it is evident from the figure that the items were administered in roughly the same proportions in each language.

[Insert Figure 1 About Here]

Differences in number of items administered from each content area were evaluated by calculating the average number of items from each content area that were administered to a typical candidate. These content area averages were computed by dividing the number of items administered from each content area by the total number of candidates in each language group. These averages are plotted concurrently for the English and German groups in Figure 2. The averages fall along a straight line, indicating that the content balancing feature of the CAT appears to be functioning similarly in both languages.

The average number of items taken by candidates in each group was also calculated. The average candidate who took the English version responded to 15.5 questions, the average candidate who took the German version responded to 15.7 questions. Thus, the number of items administered on average is consistent across the different language versions of the CAT.

[Insert Figure 2 About Here]

Item Response Time Analyses

The average (median) time it took for candidates to answer each item was calculated separately for each language group. The majority of items were answered, on average, in less than two minutes. However, a few of these items took an average of five minutes or more. These items were the simulation items, one of which took about a minute longer for the average U.S. candidate to answer. Due to this spread of response times, the square root of the average response times were compared across the two language groups. These values are plotted concurrently in Figure 3. With only a few exceptions, on average, the German candidates take longer to answer the items. The average response time for the U.S. candidates was 74.3 seconds per item; the average response time for the German candidates was 80.1 seconds per item. This finding is not surprising for two reasons. First, German is a "longer" language morphemically. For most of the items, the German versions required more text than the English versions. Thus, the German items probably take longer to read. Second, some of the German candidates may be taking advantage of the option to access the English version of the item. Although accessing the

English version of an item is almost instantaneous, reading both versions of an item would increase the response time. Given that the exam has a very liberal time limit, the fact that the Germans may require more time to answer the questions is not likely to negatively affect their probability of passing.

[Insert Figure 3 About Here]

For the 16 items used in the dimensionality and IRT analysis, further examination of the German and English response times were conducted. For each of these items, the distribution of response times was compared across the two groups. Three response time distributions were computed for each group: response time for all candidates, response time for candidates who answered the item correctly, and the response time for candidates who answered the item incorrectly. Given that, overall, the Germans took longer to respond to each item, the distributions were considered to be similar if there was a shift in the response time distributions, but the shape was consistent for each group. Using this process, the labels "very similar," "similar," and "dissimilar" were given to the translated item pairs to describe the similarity of the response time patterns across languages. Figures 4 and 5 provide an example of items that were considered to have very similar and dissimilar response time distributions, respectively. In Figure 4, the shape of the response time distributions are roughly the same across all three comparisons. In Figure 5, the German response times appear bimodal, especially for the "incorrect" group, whereas the U.S. data are unimodal. Although a more empirical criterion or larger sample sizes would have been helpful, the visual inspections did reveal similarities and differences in the item response times across groups. These inspections were considered useful for determining how similarly the items functioned across the two groups.

[Insert Figure 4 About Here]

[Insert Figure 5 About Here]

Comparison of Scaled Score Distributions

The test data analyzed are placed on a score scale ranging from 200 to 800. The distributions of scaled scores were compared for those who passed, failed, and received an incomplete (also a failing score) on the test across the two languages. These distributions, presented in Figure 6, were extremely similar across the two language groups. An interesting observation noted in Figure 6 is that there is a substantial drop in frequency over the 560-to-590 scaled score range. This drop probably reflects the "mastery search" feature of the CAT algorithm, which is designed to identify the highly proficient candidates early on, so that their tests can be ended more quickly than candidates who are closer to the passing score.

[Insert Figure 6 About Here]

Dimensionality Analyses

The one-dimensional MDS models demonstrated good fit to the data for both the English and German candidates. The one-dimensional solution accounted for 96% of the variance in the item dissimilarities for the English items, and 92% of the variance in the item dissimilarities for the German items. The STRESS (badness of fit measure) and R-squared (proportion of variance accounted for) indexes for the one- through three-dimensional MDS solutions for each data set are presented in Table 1. Relatively little improvement in fit occurs when adding one or two dimensions to the model. The one-dimensional solution scaled the items similarly in both languages. The Pearson correlation between the scale values from the English and German one-dimensional solutions was .92. These scale values were highly related to the item p-values (proportion correct) for each language group. The correlation between the item coordinates and p-values was .94 for the English items and .93 for the German items.

[Insert Table 1 About Here]

The case for unidimensionality is supported further when interpreting the results of the two-dimensional solution. Figure 7 presents the configuration of the English items in the two-dimensional MDS space. The classic (inverted) u-shaped pattern is observed, suggesting that over-fitting occurred in two dimensions. The configuration of German items in the two-dimensional MDS space is portrayed in Figure 8. A similar scaling pattern is observed. In comparing the two solutions, it appears that the items are scaled similarly along the first dimension in both solutions. Item number five has a notably larger coordinate on the second dimension for the German data. This item could account for the small difference in fit (4%) of the one-dimensional solution between the English and German data. In sum, the results of the MDS analyses suggest that the test data are unidimensional for both language groups, and that the items scale similarly in both languages. However, the solutions provided preliminary evidence that item number five functions differentially across the two languages, possibly contributing to multidimensionality in German.

[Insert Figure 7 About Here]

[Insert Figure 8 About Here]

IRT Results

Preliminary analyses

A 3PL IRT model was fit to the data separately for the English and German candidates. Item number sixteen exhibited near-zero discrimination parameters in both groups and so it was eliminated from all further analyses. Subsequently, a 3PL and a 2PL model were fit to both data sets of the remaining fifteen items, again, separately for each language group. A comparison of twice the negative of the log likelihoods for the models in each group indicated that the two parameter model represented a more parsimonious fit (i.e., the 3PL did not exhibit improved fit; $G^2_{(15)}$ English=8, $G^2_{(15)}$ German=7; both probabilities are around .90). Therefore, the 2PL was used in all subsequent analyses.

Comparing theta estimates from English- and German-calibrated items

Thetas (IRT-scaled scores) were estimated for the German candidates based on the item parameters calibrated from their responses to the items. Next, thetas were estimated for these same German candidates using the item parameters calibrated from the U.S. candidates who responded to the English versions of the items. Thus, each German candidate had two theta estimates, one based on the parameters derived from the U.S. candidate pool (similar to the standard practice at Novell), the other based on the item parameters estimated directly from their responses. These two theta estimates were compared to evaluate the impact of the assumption that the U.S.-derived item parameters were appropriate for use in German. The correlation between these two theta estimates was .91, indicating a strong relationship between the proficiency estimates produced using either set of item parameters (83% shared variance). Figure 9 presents the scatter plot of these two different theta estimates for the German candidates. To determine if the observed theta differences were due to calibration sample size differences, a random sample of 505 English candidates (equal to the sample size for the German candidates) was used to calibrate new item parameters. These item parameters were then used to calculate new thetas for the Germans. The correlation between the thetas based on the complete English sample and the reduced English sample was .99, which indicates that the difference in theta estimates calculated from the U.S. or German responses is probably not due to sample size.

[Insert Figure 9 About Here]

Although the correlation between the theta estimates using the English or German item parameters was high, some important differences were observed. The difference between the means of the two separate theta distributions was .07. The mean theta estimate for Germans was reduced by .07 when the parameter estimates were derived from the German response data. This reduction was a consequence of the fact that nine of the fifteen b-parameters were lower (i.e., easier) when calibrated using the German sample. This finding could be due to either the items becoming easier when translated into German, or to the German candidates being of higher ability than the English candidates.

The effect of using the English or German parameter estimates on passing the test was also evaluated. The point on each theta scale that corresponded to the passing score used on the operational CAT was determined. Using this criterion, 76.0% of the German candidates would pass if the English item parameters were used, and 69.3% of the German candidates would pass if the German parameters were used. As stated above, 73.5% of these 505 German candidates actually passed, based on their complete tests. This value is close to the value obtained using the

English parameter estimates in this study[1]. The lower passing rate (69.3%) that would be obtained if the items were calibrated separately in German (and not equated to the English item parameters) can be due to either the items being easier in German, or to the German candidates being more proficient on average.

The item parameters estimated from the U.S. and German samples are not on the same scale. Because translated items cannot be considered equivalent, we cannot interpret the differences between these two sets of parameters at face value. However, inspection of the differences in these parameters across languages is informative. Table 2 presents the item parameters estimated for the items in each language. Also indicated in Table 2 are the collateral information for each item: the ranking of "best translated" items for those eight items so identified by the expert translators, the subjective classification of response time similarity, and the indication of the simulation items. To make comparison of the parameter estimates easier, scatterplots of the English and German difficulty and discrimination parameters are presented in Figures 10 and 11, respectively. The most striking observation from these figures and Table 2 is the very different parameters estimated for item number five. This item was very difficult and poorly discriminating in German, and of moderate difficulty and discrimination in English. Interestingly, this item was one of the three simulation items, and was also the item that appeared aberrant in the MDS analysis. Another notable observation is the wide variation of the discrimination parameters across the two groups. This finding is probably due to the relatively small samples of candidates and items. After item number five, item number fourteen was the next item that displayed relatively larger differences across the two groups. This item was less discriminating and more difficult in German than in English.

[Insert Figure 10 About Here]

[Insert Figure 11 About Here]

[Insert Table 2 About Here]

Search for translation DIF

As a first step in disentangling the potential "translation difference" effect from the "group proficiency effect," the "ALL DIF" and "NO DIF" models were fit to the data. To complete this analysis, the data for the English (n=965) and German (n=505) candidates were calibrated concurrently. The design of the analysis treated the data set as comprising 30 test items, with each language group having missing data on 15 of these 30 items. In the ALL DIF model, separate parameters were estimated for the English version of an item and its German

[1] It should be remembered that for the purposes of this study, the theta estimates for the 505 German candidates were not computed from their entire set of item responses, but rather from their responses to only those of the 16 items selected for analysis. On average, candidates responded to about 8 or 9 of these 16 items, and so the estimates are less precise than those that would come from the operational CAT.

counterpart. In the NO DIF model, the parameters for an English item and its German counterpart were constrained to be equal. To define the scale, the mean theta for the Germans was set to zero and the standard deviation of both groups was set to one. Thus, the ALL DIF model involved the estimation of 61 parameters (15 a-and b-parameters for the 15 English items, 15 a-and b-parameters for the 15 German items, and the mean of the theta distribution for the English group). The NO DIF model involved the estimation of 31 item parameters (a common set of 15 a- and b-parameters for the 15 English and German items, and the mean theta for the English group).

The LR test comparing the ALL DIF and NO DIF models indicated that the NO DIF model can be rejected outright (see Table 3). Allowing the parameters to be estimated separately for each group results in statistically significant improvement in fit to the data ($G^2_{(30)}$=140, p<.001). The next set of analyses aimed toward locating the source of the DIF.

The results of the separate item parameter calibrations suggested that items five and fourteen were the most different across the two samples. Thus, the next model fit to the data was the "ITEM 5 DIF" model. This model estimated separate parameters for the English and German versions of item number five, and equivalent parameters for the remaining fourteen English-German item pairs. The ALL DIF model also exhibited statistically significant improvement in fit when compared to this model ($G^2_{(28)}$=49, p<.01). The next model fit to the data, "ITEMS 5 & 14 DIF," estimated separate item parameters only for the English and German versions of these two items. The ALL DIF model did not fit the data better than the ITEMS 5 & 14 DIF model ($G^2_{(26)}$=29, p=.32), indicating that calibrating equivalent item parameters for the English and German versions of the other thirteen items fit the data as well as calibrating separate parameters for the different language versions of all items. The parsimonious fit of this model suggests that only items five and fourteen function differentially across the English and German languages. The other thirteen items appear to be functioning similarly.

[Insert Table 3 About Here]

The estimated mean proficiency for the English group varied (of course) across the different models described above. To test whether the mean proficiencies of the two groups were equal, additional models were fit to the data. The results of the various models fit to the data are summarized in Tables 3 and 4. These table include the mean theta parameters in each model estimated for the English group. Table 3 summarizes the results focusing on locating DIF; the results in Table 4 summarize those models used to evaluate group differences. In the ALL DIF model, the English mean is .66 standard deviations higher than the German mean (see Table 3). This finding makes sense because if the items were all different across the two languages, on average, the Germans took more easy items. However, we know this model is not correct. After accounting for the fact that some of the items (ITEMS 5 & 14 DIF model) or all of the items (NO DIF model) are equivalent, the English mean is estimated to be at least one-third of a standard deviation lower than the German mean. This finding also makes sense because if some of the items are equivalent across the two groups, the higher b-values obtained for the separately-calibrated English items are incorrect.

Acceptance of the "ITEMS 5 & 14 DIF" model involves concluding that the Germans are more proficient, on average, than the U.S. group (by .36 standard deviation units). This hypothesis was tested by fitting a new "ITEMS 5 & 14 DIF" model to the data that constrained the means of the two groups to be equal. The original "ITEMS 5 & 14 DIF" model, that allowed the English and German means to vary, exhibited statistically significant improvement in fit to the data than the model constraining the means to be equal ($G^2_{(1)}$=17, p<.001). Thus, it appears that the German sample is more proficient than the English sample (see Table 4).

One final model was fit to the data as a final check on whether the observed difference between the English and German groups was valid. It is unlikely, but possible, that the thirteen items identified as functioning equivalently across the two languages, really *do* function differentially across languages, but look equivalent because the IRT theta metric is biased. Theoretically, this could happen if the translation of these thirteen items into German made them all systematically easier and the English group mean was higher than the German group mean by about the same amount. The only way to evaluate this type of rival hypothesis is to defend the cross-lingual equivalence of a set of items in a manner independent of the calibration model (Sireci, 1997b). To do this, the two remaining simulation items (item numbers 10 and 15) and the two items that were identified as representing the two "best" translations (item numbers 3 and 8) , were considered equivalent and were used to comprise a common item set (anchor) across the two language groups. The equivalence of these four items was supported by the observation that the response time distributions for the items were rated "similar" or "very similar." This final model, the "ANCHOR" model was fit to the data both with and without constraining the means to be equal. The ANCHOR model that estimated separate means for the two groups fit the data statistically significantly better than the model that constrained the means to be equal ($G^2_{(1)}$=30, p<.001; see Table 4). In this model, the English group mean was .61 standard deviation units lower than the German group. However, this model is less parsimonious than the "ITEM 5 & 14 DIF" model which fit the data well using fewer parameters. Therefore, the .36 difference noted in the former model is a better estimate of the average difference in proficiency between the English and German samples.

[Insert Table 4 About Here]

The important finding is that in both cases, the English group mean was estimated to be lower than the German group mean, supporting the conclusion regarding DIF for only items five and fourteen. It should be noted that the validity of any anchor model is dependent on whether the items selected for the anchor are appropriate. Unfortunately, the validity of an anchor can never be absolutely determined. Nevertheless, this analysis provided additional evidence that the conclusions reached above regarding item and group equivalence are correct.

Discussion

Through a comprehensive series of analyses, the results of this study indicate that the English and German versions of the Novell certification examination studied are remarkably similar. The types of items administered, and the number of items administered were consistent across the two language groups. In addition, the dimensionality of the test was consistent (unidimensional) across the two languages.

When evaluating the passing rates based on the parameter estimates calibrated from the English and German samples, a difference of about 6.7% was observed. More Germans passed our "smaller" version of the CAT when the parameter estimates were calibrated from the U.S. candidates. Given that it was later concluded that the German candidates were of higher proficiency, the higher passing rate observed using the U.S.-derived item parameters is sensible. Therefore, the practice of using item parameter estimates calibrated from U.S. samples seems appropriate.

The results also underscore the importance of monitoring the functioning of the items across all languages. If DIF is observed for some items, the U.S.-based parameters may not be appropriate. However, on the other hand, just because an item functions differentially across languages does not necessarily signify a translation problem. For example, if the German sample was unaware of a particular feature of the software being tested, but this feature was commonly known by the English sample, the item may function differentially across languages, but such differential functioning would be expected and does not threaten the validity of the test scores. In any case, "translation DIF" should be studied. If the DIF observed for an item cannot be explained, the use of a common set of parameters across languages for the item may not be appropriate.

Post hoc qualitative analyses of the two items that displayed DIF in this study produced some interesting conclusions. For item five, the item that was much more difficult and much less discriminating in German, an external analysis was able to explain the DIF. Bilingual test specialists at Novell reviewed hard copies of the item in English and German. At first, they were unable to figure out why the item functioned differentially across the two languages. Fortunately, they decided to compare the English and German versions of the item as displayed on the computer screen. They immediately discovered a formatting error. One of the correct answers was located in the wrong portion of the screen in the German version of the item, thus making it harder for the Germans to determine the correct answer. Novell corrected the formatting of the item, and within 24 hours, the corrected item was downloaded worldwide. The Novell translation team was unable to explain the DIF observed for item 14. In fact, this item was previously considered the "fourth best" translated item. However, a bilingual translator in Germany asserted that the wording of the item in English was confusing, and that the confusing wording was retained in the German translation. As he put it "*Von hinten durch die Brust ins Auge*," which is a German saying describing a "very complicated word order, which is very often used in passive voice" (Volker Enkrodt, personal communication April 4, 1997). This explanation is consistent with the observed relatively poor discrimination of the item in both English and

German. Thus, it was somewhat comforting that the DIF observed empirically was explainable by post hoc inspection of the items.

The fact that only two of fifteen studied items functioned differentially across languages, in general, supports the validity of Novell's translation process. This relatively minor amount of DIF may pose less of a problem in a CAT environment than in a linear test administration. If DIF contributes to a candidate answering an item incorrectly, the item will probably not lower the estimate of error around the candidates' score, and subsequent items should correct the estimate. Evaluating DIF across languages will reduce the probability that a candidate would be administered a DIF item, and the probability that a candidate would be administered two DIF items (in the same direction) would be even lower. Nevertheless, if DIF items are presented to a candidate in a CAT, the best case scenario is that the test administration is extended; the worst case scenario is that a candidate's final score is raised or lowered inappropriately. Thus, it is important to evaluate items for DIF due to translation.

A likely reason that the majority of test items studied here functioned similarly in the two languages is that many of the items involved less verbiage than that found on many other tests, such as academic or personality tests. For example, the simulation items required candidates to perform tasks on the actual operating system. The test is highly content-valid because the objectives measured are defined clearly and concretely (Foster et al., 1997). In addition, candidates in all countries have experience operating the software before taking the test. Because machine language does not have to be "translated," the software performs equivalently across languages. Given that rigorous translation procedures were followed, the construct measured in the two languages is considered to be equivalent, and that verbal skills are not a major factor for success on the exam, it is not surprising that the different language versions of this exam function equivalently.

A notable feature of this study is the multiple methods that were used to evaluate the translation equivalence problem. First, the functioning of the entire CAT was compared across languages. Second, the effect of assumptions underlying the CAT algorithm (that using English item parameters for computing German thetas) was explicitly tested. Third, when evaluating the functioning of the items across languages, known threats to the internal validity of cross-lingual DIF studies were explicitly modeled and evaluated. Each IRT model fit to the data was motivated by a specific hypothesis or rival hypothesis. The use of collateral information to anchor the two scales allowed the anchor to be constructed independent of the calibration model. Although there are advantages and limitations of using subjective judgment to form an anchor, the fact that both the empirical and judgmental "common" items led to the same conclusion supported the validity of the final conclusions.

The results also demonstrated that some types of collateral item information are useful. Although one of the simulation items displayed DIF, the DIF was explainable due to a severe formatting error. Thus, items that are performance-based and involve little text appear useful for anchoring score scales across language groups. However, as the results of this study illustrate, their functioning in each language needs to be evaluated. The use of an independent team of expert translators to identify the "best translated" items was equivocal. One of the seven items they identified as a relatively good translation exhibited DIF. This finding supports the general conclusion that translated items cannot be considered equivalent without empirical verification. With respect to the item response time comparisons, it was very difficult to qualitatively interpret the distributions across languages. No problems were suspected when looking at the response time distributions for those items that were later identified as DIF items. The two items that appeared differently in terms of response time distributions (items 6 and 11, see Table 2) did not exhibit DIF. Thus, although expert judgment, less verbal item types, and analyses of response time distributions are helpful for identifying similarity of item functioning across languages, sensible caution must be exercised. Rather than relying on purely statistical or collateral criteria, a comprehensive series of analyses, such as that presented in this study, is needed.

An interesting observation noted in this study is that the item that exhibited the largest DIF in the IRT analyses also appeared to be aberrant in the MDS analyses (item 5). Future research should be done to investigate the utility of MDS for evaluating DIF.

A notable limitation of this study is that thetas were computed for the English and German candidates using about half of the items they actually answered. Furthermore, we evaluated only a small subset of items from the item pool. These limitations were necessitated by the data available. If possible, future research should investigate the item and test equivalence problems using designs that require different language test takers to respond to the larger set of common, translated items.

References

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. Psychometrika, 46, 443-459.

Chen, T., & Davison, M. L. (1996). A multidimensional scaling, paired comparisons approach to assessing unidimensionality in the Rasch model. In G. Engelhard, & M. Wilson (Eds.). Objective measurement: Theory into Practice (Volume 3), Norwood, NJ: Ablex.

De Ayala, R. J., & Hertzog, M. A. (1991). The assessment of unidimensionality for use in item response theory. Multivariate Behavioral Research, 26, 765-792.

Foster, D., Olsen, J. B., Ford, J., & Sireci, S. G. (1997, March). Administering computerized certification exams in multiple languages: Lessons learned from the international marketplace. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Geisinger, K. F. (1994). Cross-cultural normative assessment: translation and adaptation issues influencing the normative interpretation of assessment instruments. Psychological Assessment, 6, 304-312.

Hambleton, R. K. (1993). Translating Achievement tests for use in cross-national studies. European Journal of Psychological Assessment, 9, 57-68.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: a progress report. European Journal of Psychological Assessment, 10, 229-244.

Hambleton, R. K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. European Journal of Psychological Assessment, 11, 147-157.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.

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

Sireci, S. G., (1997a). Dimensionality issues related to the National Assessment of Educational Progress. Commissioned paper by the National Academy of Sciences/National Research Council's Committee on the Evaluation of National and State Assessments of Educational Progress, Washington, DC: National Academy of Sciences.

Sireci, S. G. (1997b). Problems and issues in linking tests across languages. Educational Measurement: Issues and Practice, 16,12-19.

Sireci, S. G., & Berberoglu, G. (1997, March). Evaluating translation DIF using bilinguals. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Sireci, S. G., & Swaminathan, H. (1996, October). Evaluating translation equivalence: So what's the big DIF? Paper presented at the annual meeting of the Northeastern Educational Research Association, Ellenville, NY.

Thissen, D. (1991). Multilog: Multiple categorical item analysis and test scoring using item response theory, version 6 [computer program]. Mooresville, IN: Scientific Software.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group mean differences: The concept of item bias. Psychological Bulletin, 99, 118-128.

Thissen, D, Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D, Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. Applied Measurement in Education, 8, 157-186.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. Journal of Educational Measurement, 28, 197-219.

Table 1

Fit Measures for MDS Solutions

	English Items		German Items	
Solution	STRESS[a]	R-squared[b]	STRESS[a]	R-squared[b]
1-Dimension	.12	.96	.17	.92
2-Dimensions	.08	.98	.13	.93
3-Dimensions	.06	.98	.10	.94

[a] This index represents the departure of the data from the model, and so smaller values indicate better fit.

[b] This index represents the percentage of variance among the item dissimilarities accounted for by the model.

Table 2

Separately Calibrated English and German Item Parameter Estimates and Collateral Information

Item No.	English b-param.	German b-param.	Difference (Eng. b – Germ. b.)	English a-param.	German a-param.	Expert Rank[a]	Response Time Comparison	Simulation ?
12	-4.16	-2.58	-1.58	0.77	1.34	3	Very Similar	
13	-4.04	-3.14	-0.9	0.53	0.89	5	Very Similar	
8	-3.78	-3.02	-0.76	0.93	1.27	2	Similar	
7	-3.25	-1.68	-1.57	0.5	0.95		Similar	
3	-2.36	-2.11	-0.25	0.64	0.86	1	Similar	
1	-1.48	-2.27	0.78	0.91	0.58	7	Very Similar	
2	-1.18	-1.3	0.12	0.76	0.7		Similar	
10	-0.94	-1.51	0.57	0.85	0.97	8	Very Similar	Yes
6	-0.92	-1.47	0.55	0.92	0.72		Dissimilar	
14	-0.79	0.31	-0.31	0.53	0.33	4	Very Similar	
4	-0.73	-0.75	0.02	0.67	1.44		Very Similar	
15	-0.53	-1.46	0.93	1.82	1.56		Very Similar	Yes
11	-0.3	-0.91	0.61	0.68	0.97		Dissimilar	
9	-0.23	-0.59	0.36	0.55	0.8		Very Similar	
5	-0.09	3.75	-3.84	1.2	0.28		Similar	Yes

[a] The item rated most equivalent across languages is ranked "1," etc. The item ranked "6th best" was the item deleted initially due to poor discrimination in both language groups.

Table 3

Summary of Search for Differential Item Functioning

Model	English θ	German θ	No. of Params.	G^2	*df*	p
ALL DIF	0.66	0	61			
NO DIF	-0.34	0	31	140	30	<.001
ITEM 5 DIF	-0.43	0	33	49	28	<.010
ITEMS 5 & 14 DIF	-0.36	0	35	28	26	0.32

Table 4

Summary of Search for Group μ Differences

Model	English θ	German θ	No. of Params.	*G^2*	*df*	p
NO DIF	-0.34	0	31			
NO DIF $\mu_E = \mu_G$	0	0	30	16	1	<.001
ITEMS 5 & 14 DIF	-0.36	0	33			
ITEMS 5 & 14 DIF $\mu_E = \mu_G$	0	0	32	18	1	<.001
ANCHOR	-0.61	0	53			
ANCHOR $\mu_E = \mu_G$	0	0	52	30	1	<.001

Figure 1

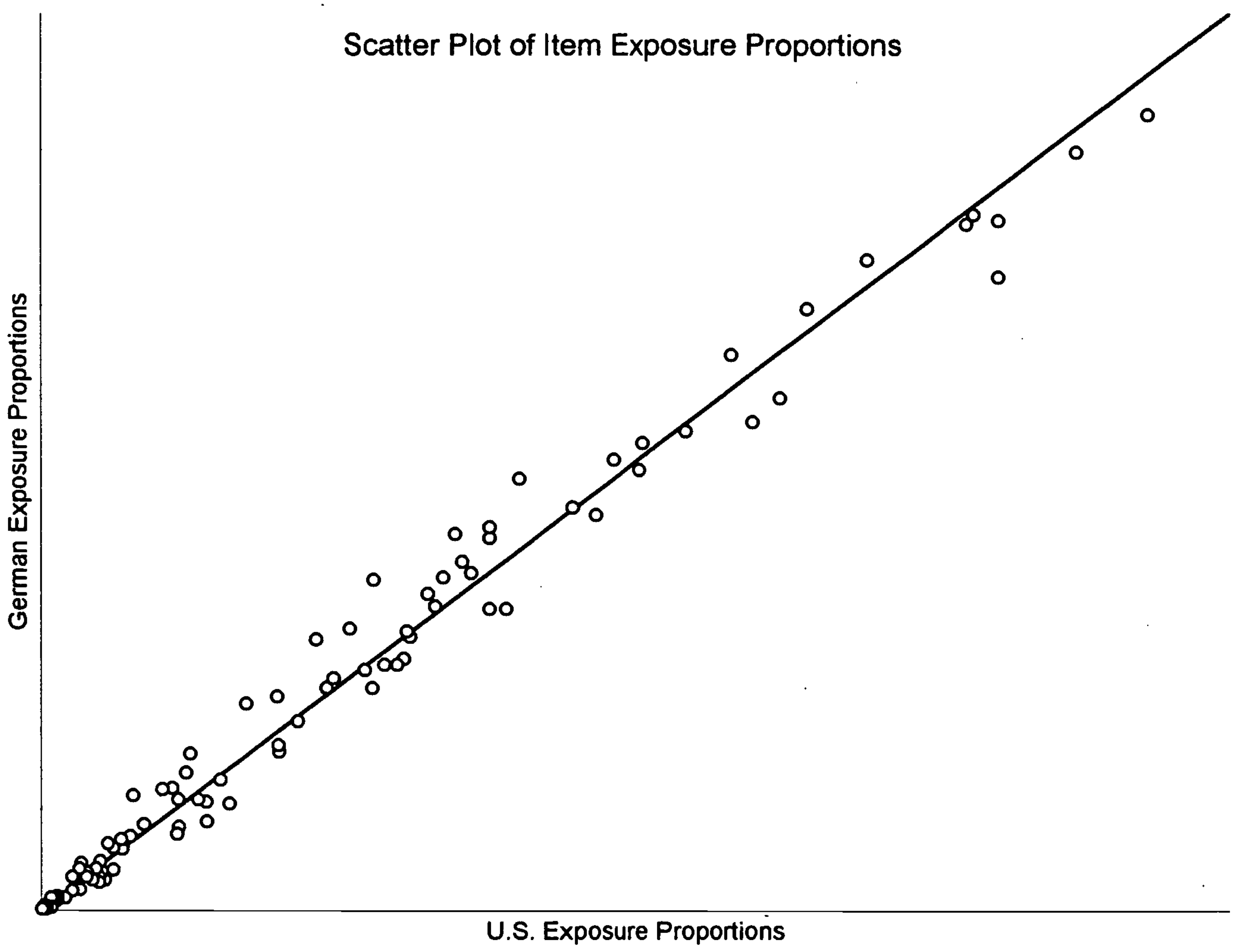

Scatter Plot of Item Exposure Proportions
German Exposure Proportions
U.S. Exposure Proportions

Figure 2

Scatter Plot of Content Area Averages

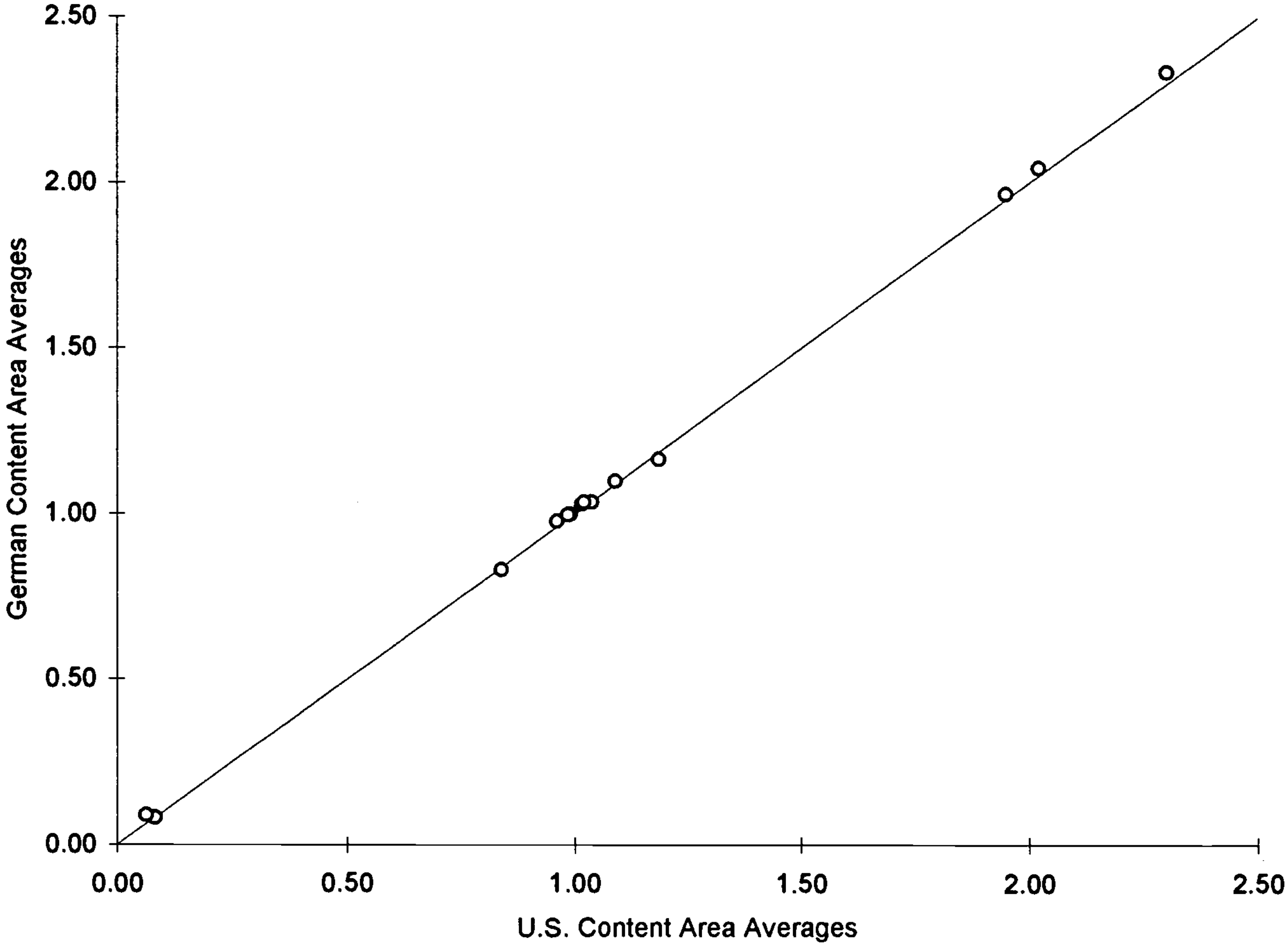


Figure 3

Scatter Plot of Square Root of Average Item Response Times

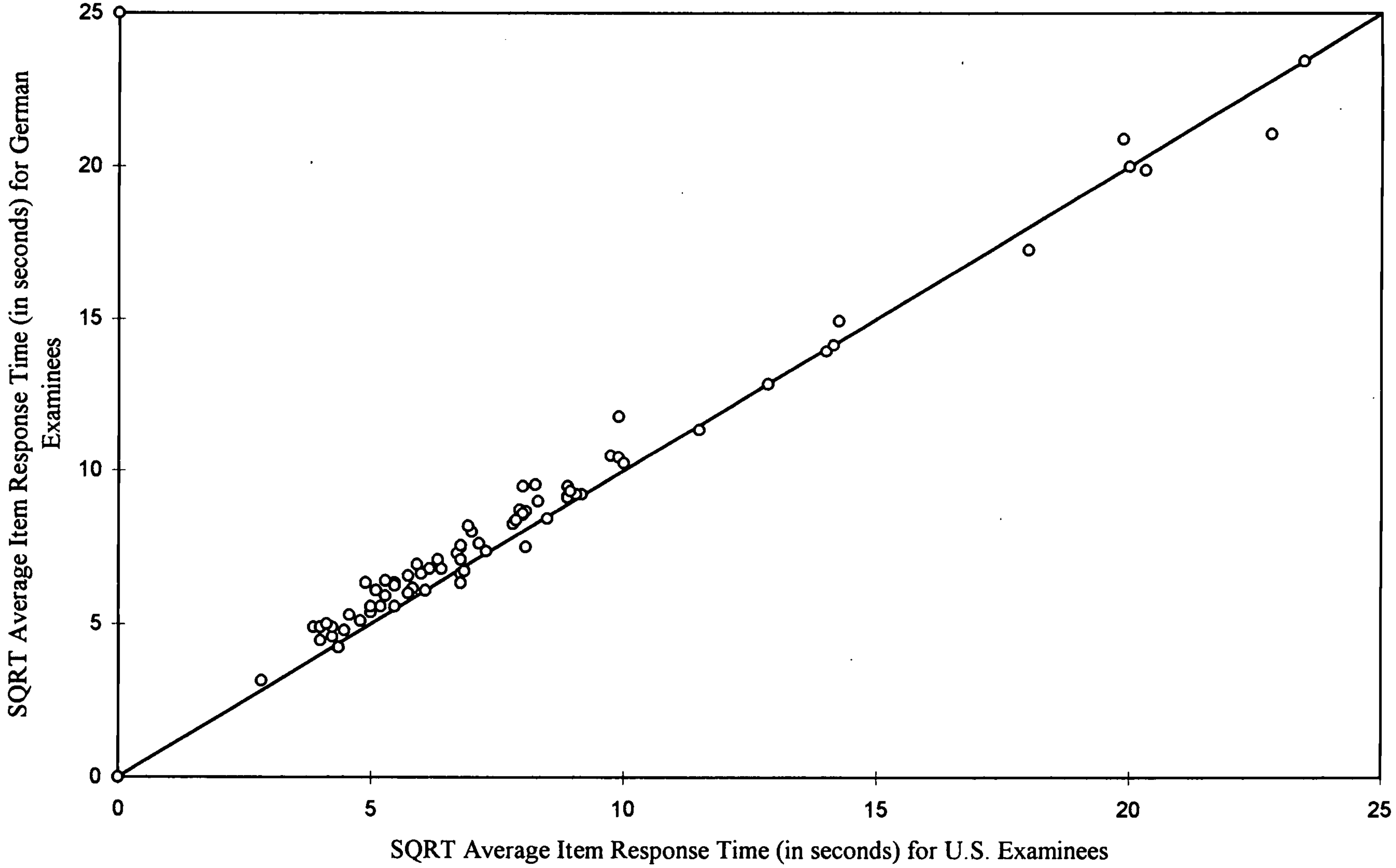


Figure 4
Response Time Frequency Distribution For "Very Similar" Item

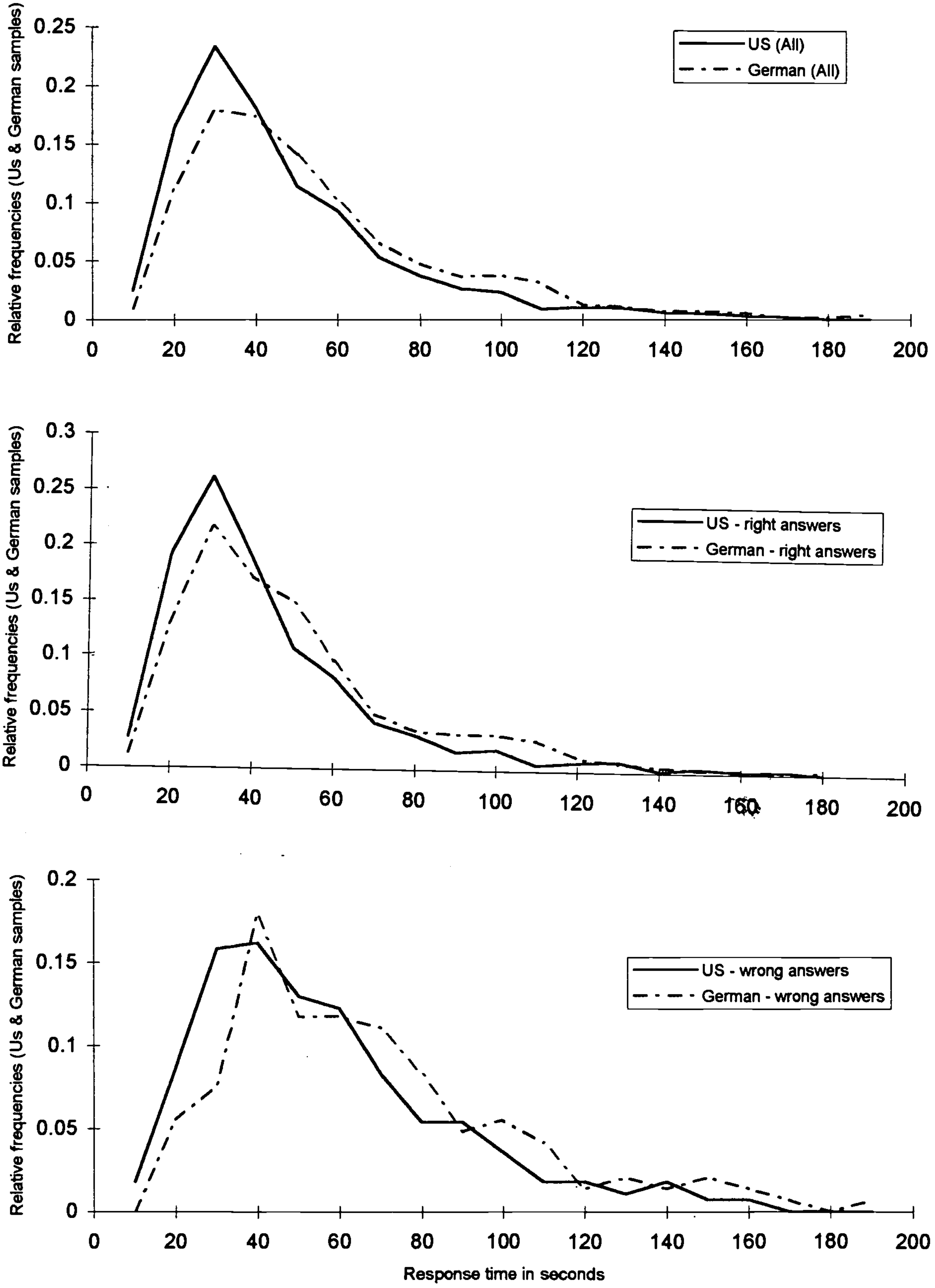


Figure 5:
Response Time Frequency Distribution For "Dissimilar" Item

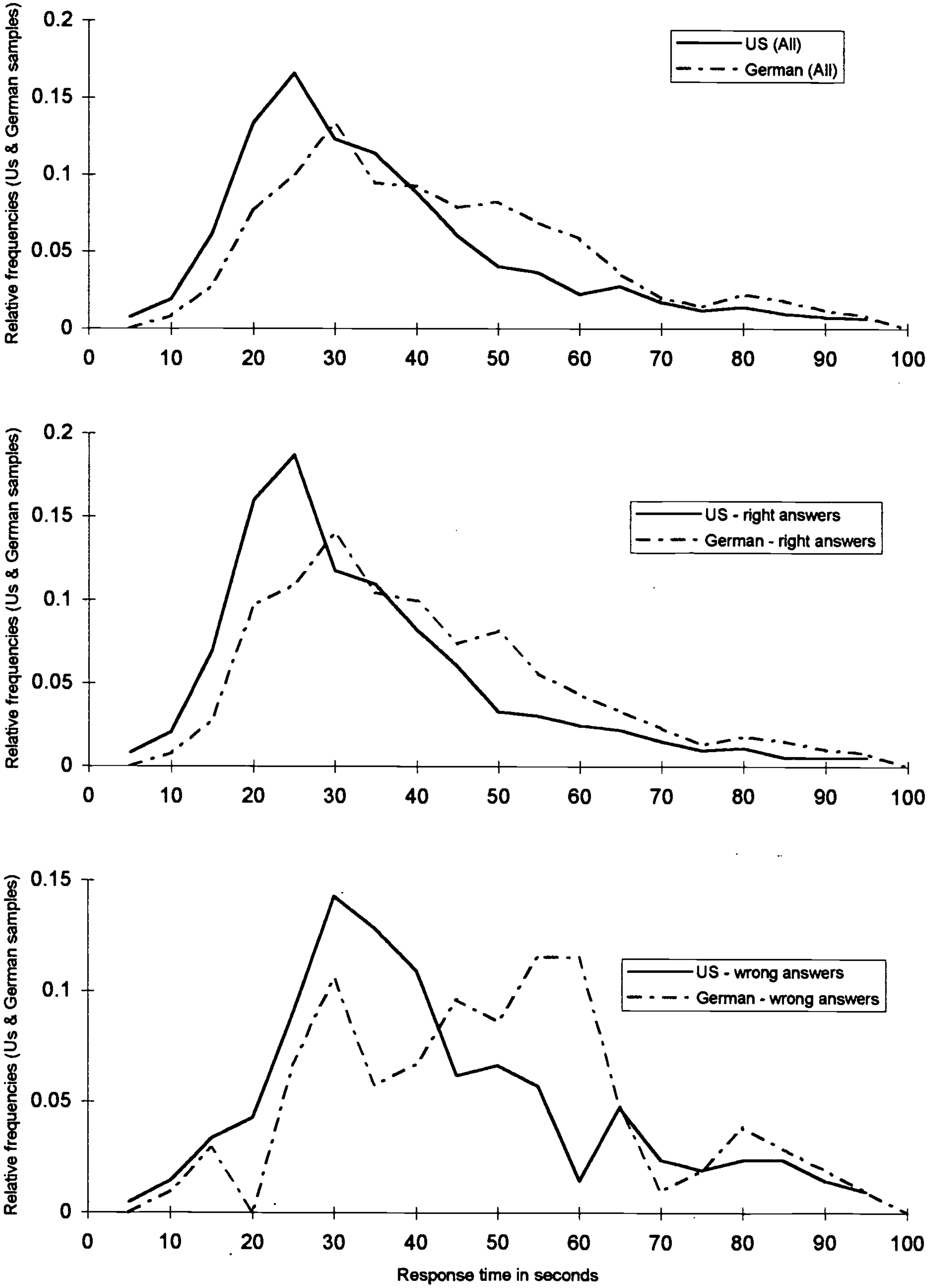


Figure 6

U.S. and German Scaled Score Distributions

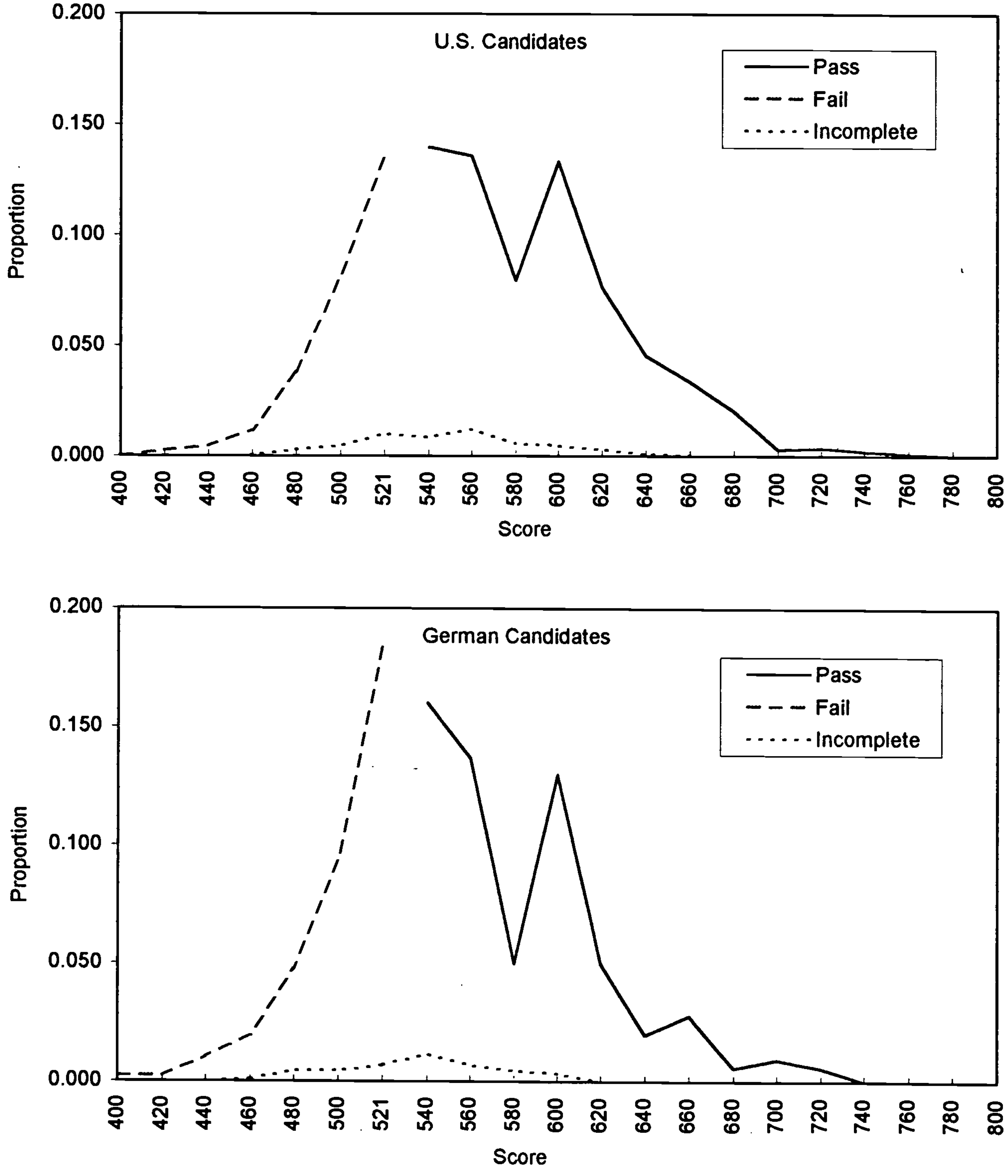


Figure 7

2-Dimensional MDS Solution for English Items

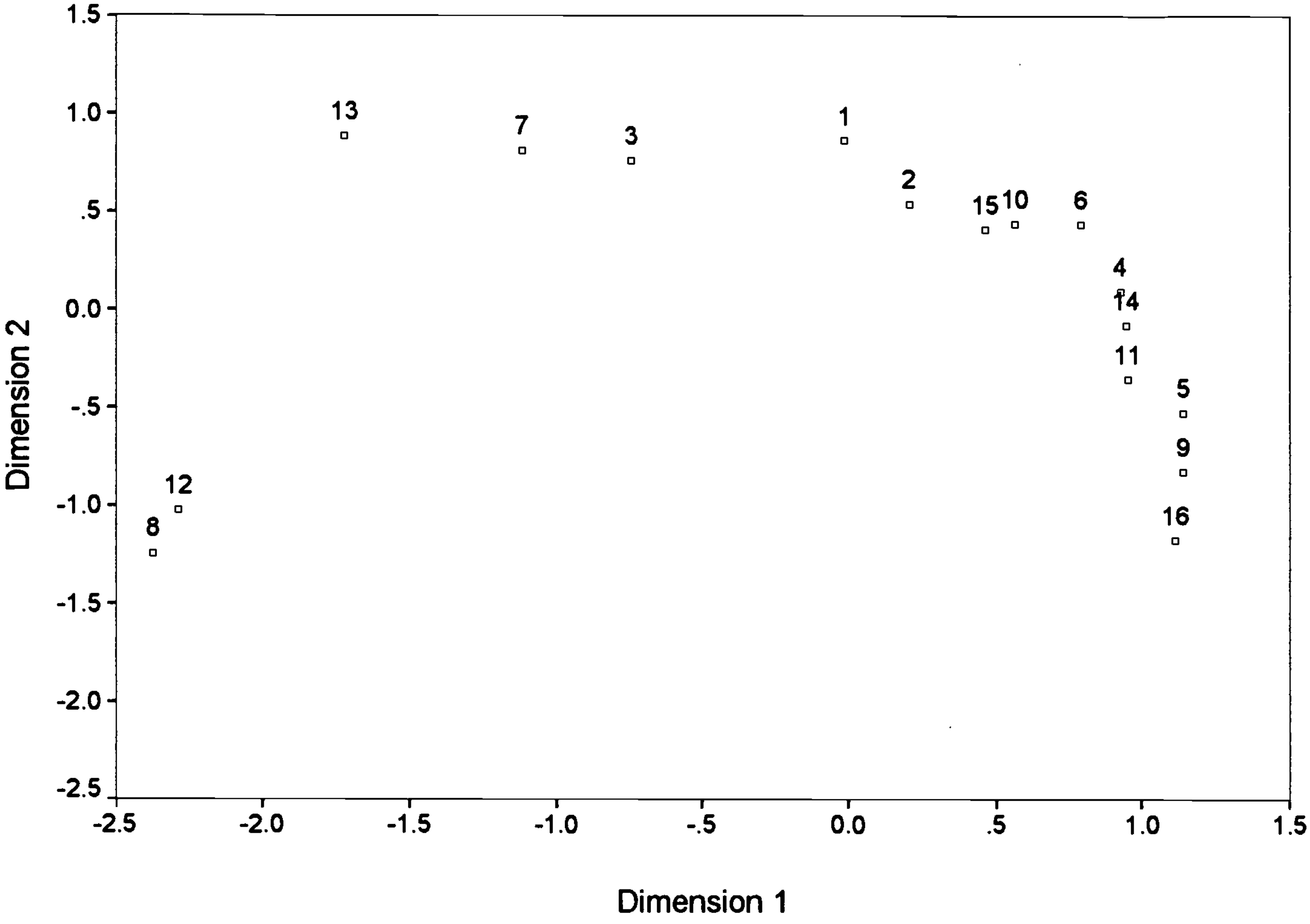


Figure 8

2-Dimensional MDS Solution for German Items

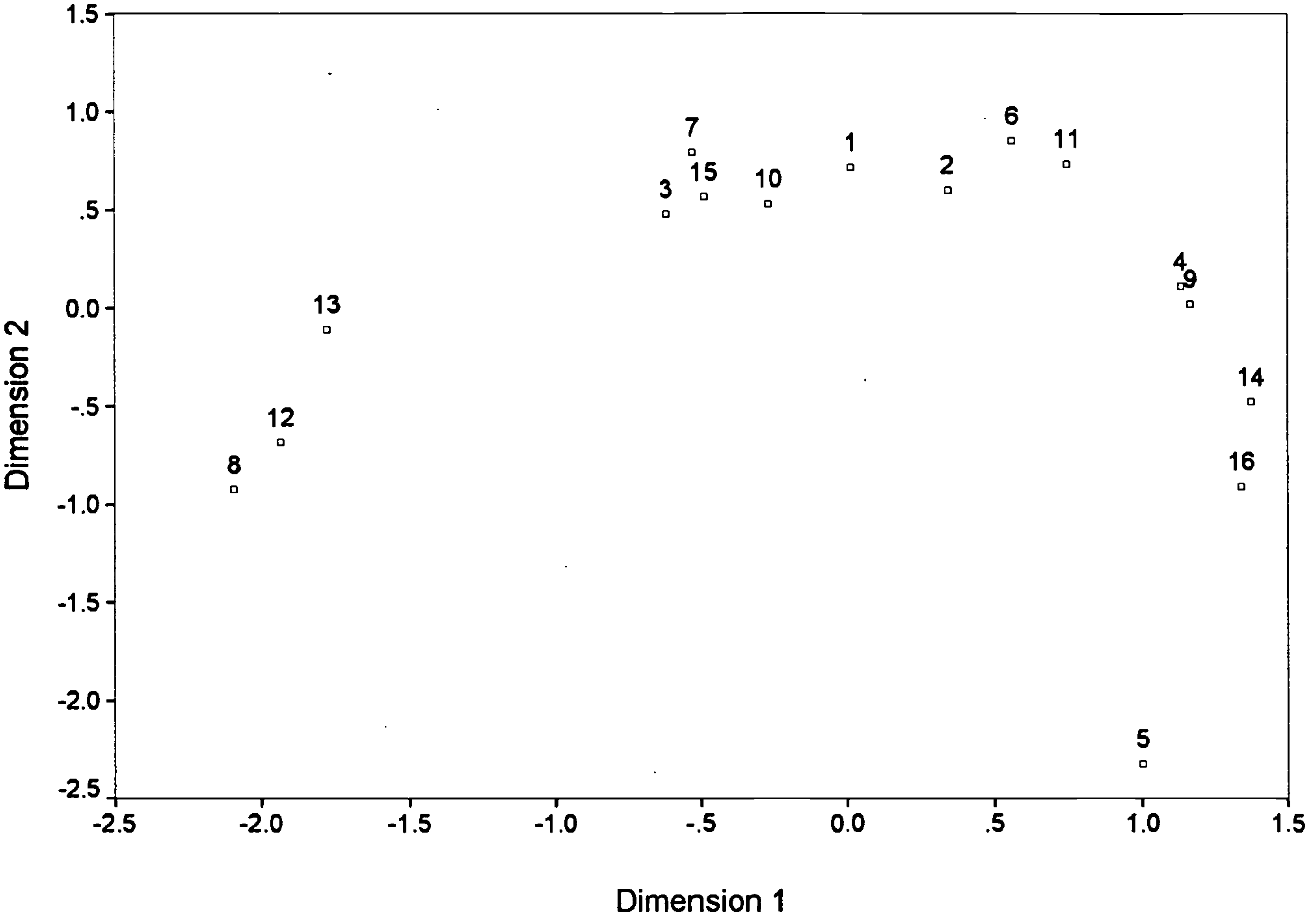


Figure 9

Scatter Plot of German Thetas

Derived From U.S. and German Data

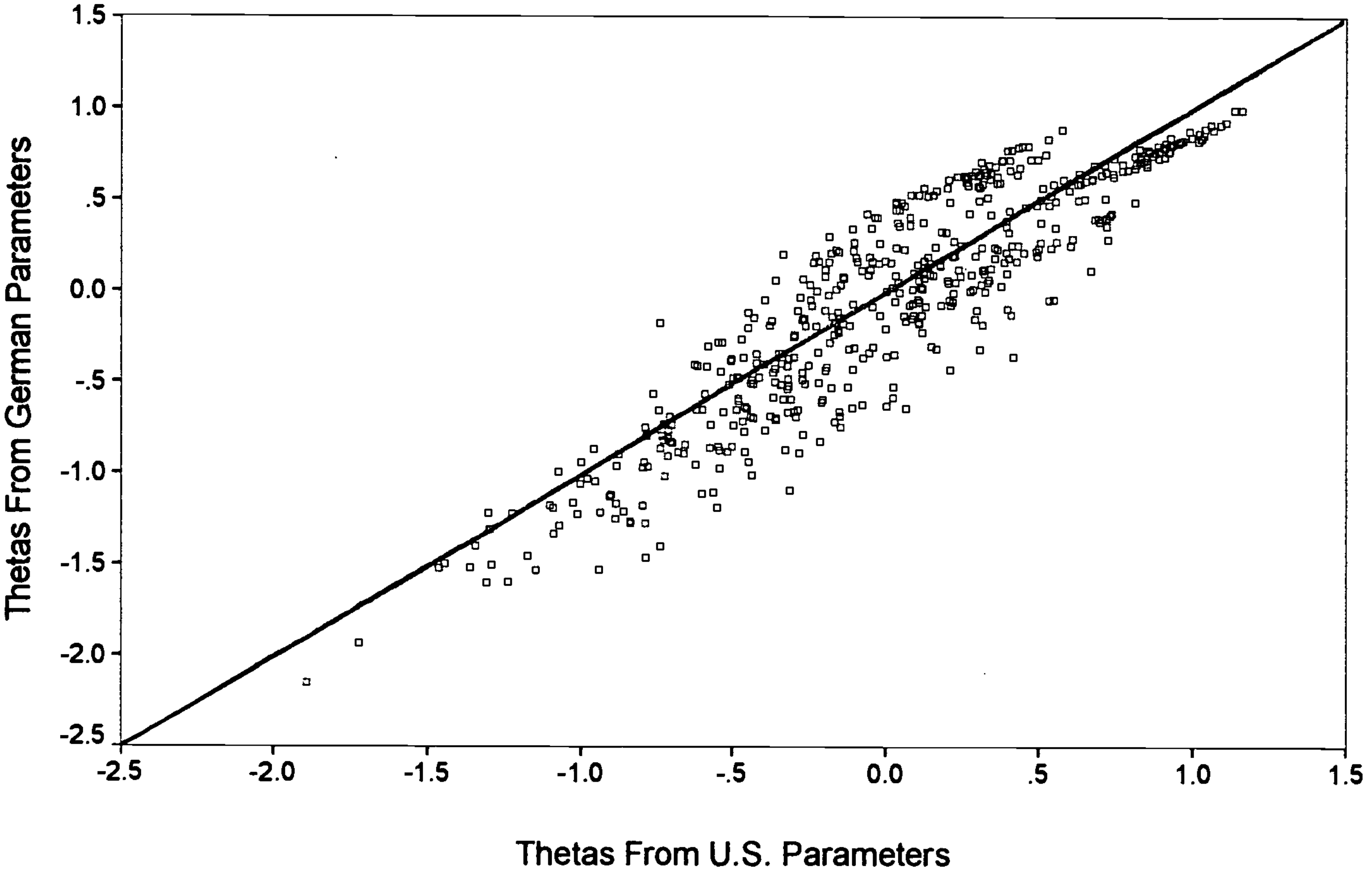


Figure 10

English b's Plotted Against German b's

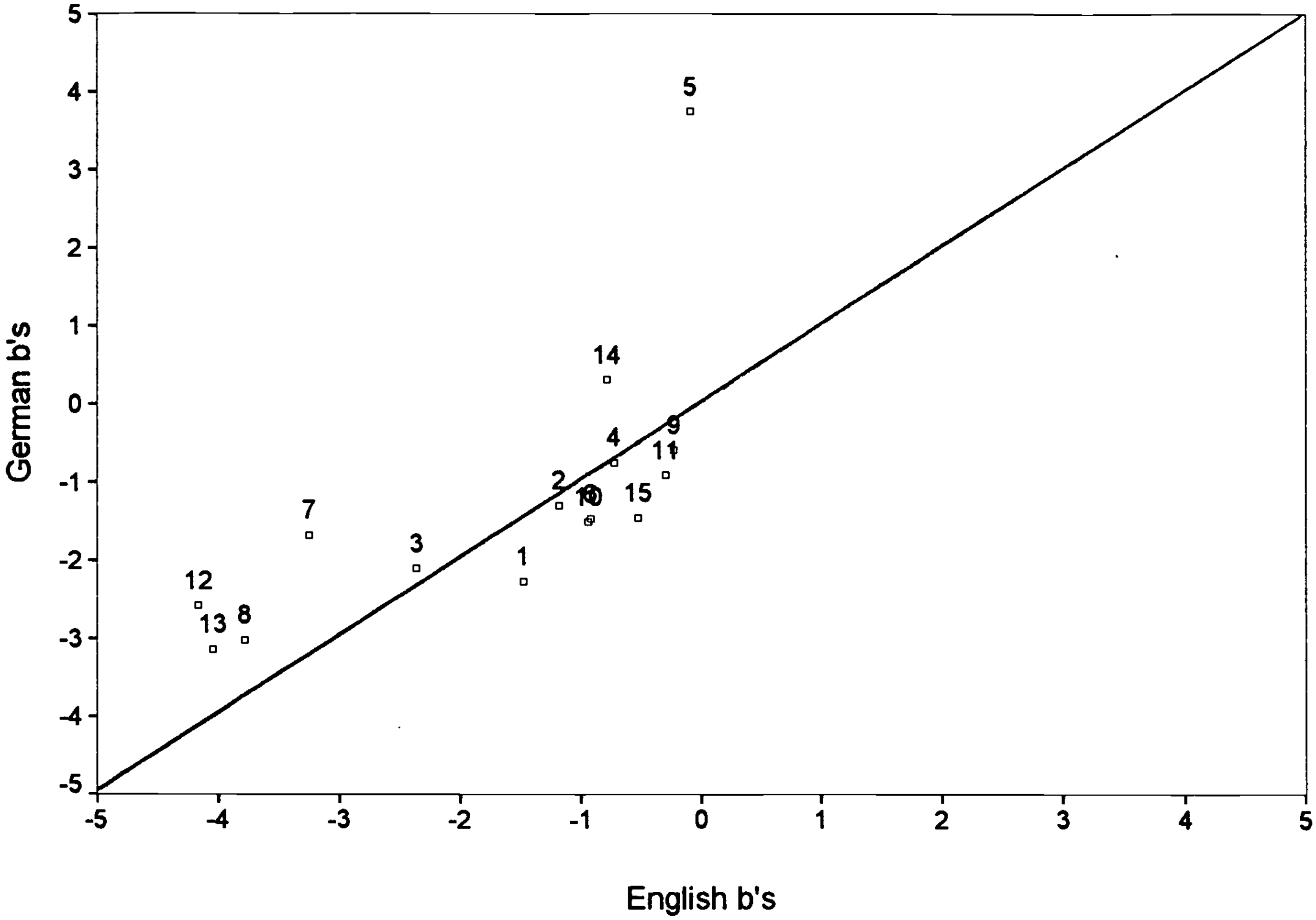
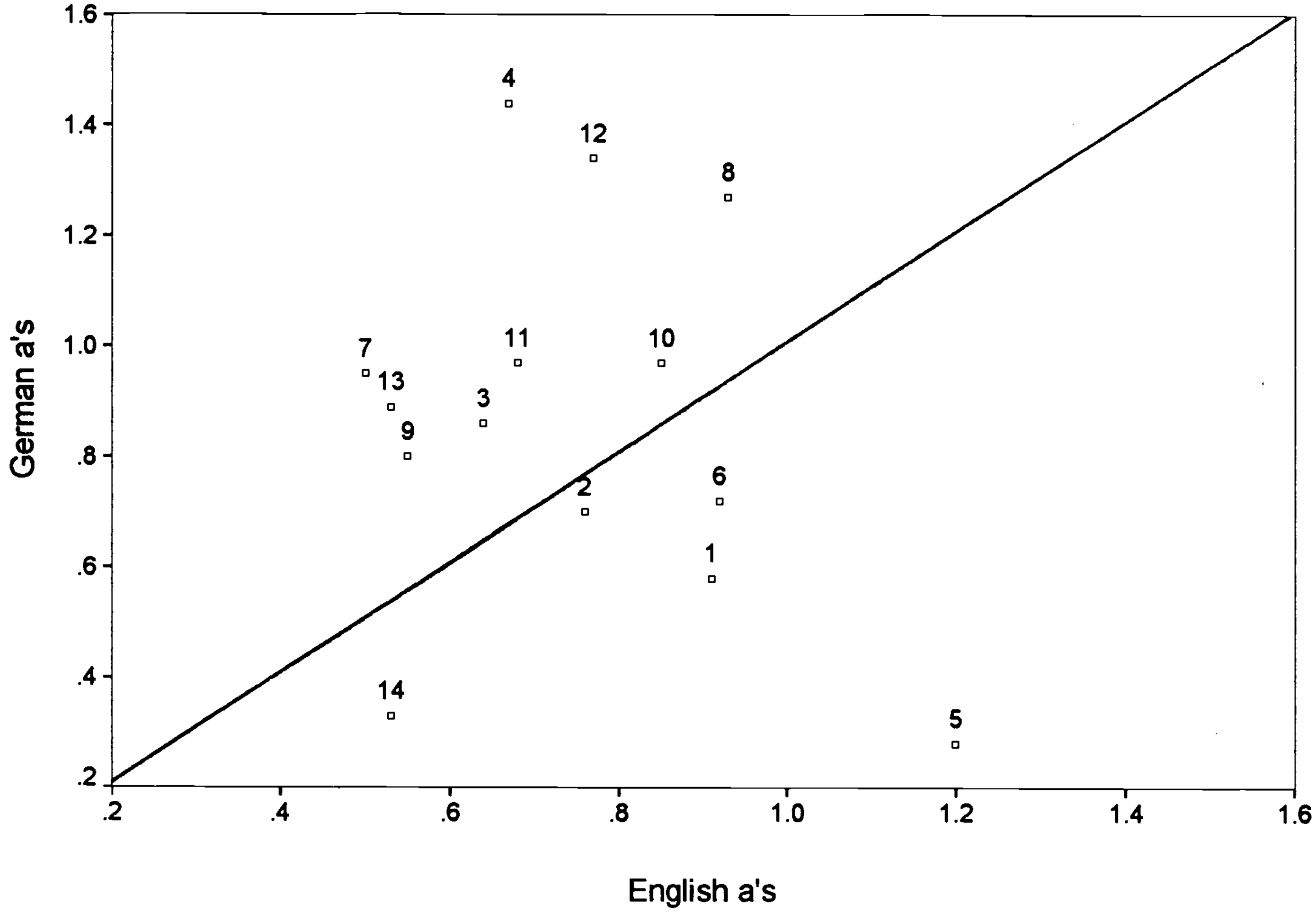


Figure 11

English a's Plotted Against German a's



NCME 1997

TM027720

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Comparing Dual-Language Versions of an International Computerized-Adaptive Certification Exam

Author(s): Stephen G. Sireci, David F. Foster, Frédéric Robin, & James Olsen

Corporate Source: University of Massachusetts

Publication Date: NCME Paper 1997

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

☑ Sample sticker to be affixed to document

Check here
Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Sample
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document □

or here
Permitting reproduction in other than paper copy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY
Sample
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature:	Position: Assistant Professor
Printed Name: Stephen G. Sireci	Organization: UMASS
Address: School of Education 156 Hills South University of Massachusetts Amherst, MA 01003-4140	Telephone Number: (413) 545-0564 Date: 6/2/97

OVER

CUA

THE CATHOLIC UNIVERSITY OF AMERICA
Department of Education, O'Boyle Hall
Washington, DC 20064
202 319-5120

February 24, 1997

Dear NCME Presenter,

Congratulations on being a presenter at NCME[1]. The ERIC Clearinghouse on Assessment and Evaluation invites you to contribute to the ERIC database by providing us with a written copy of your presentation.

We are gathering all the papers from the NCME Conference. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality. You can track our process of your paper at http://ericae2.educ.cua.edu.

Please sign the Reproduction Release Form on the back of this letter and include it with **two copies** of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can drop off the copies of your paper and Reproduction Release Form at the **ERIC booth (523)** or mail to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to:

NCME 1997/ERIC Acquisitions
O'Boyle Hall, Room 210
The Catholic University of America
Washington, DC 20064

Sincerely,

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE

[1] If you are an NCME chair or discussant, please save this form for future use.