

# External and Diagnostic Validity of the NTID Writing Test: An Investigation Using Direct Magnitude Estimation and Principal Components Analysis

Joseph H. Bochner

John A. Albertini

Vincent J. Samar

Dale Evan Metz

*Rochester Institute of Technology*

*The NTID Writing Test was developed to assess the writing ability of post-secondary deaf students enrolled at the National Technical Institute for the Deaf. The purpose of this study was to evaluate the NTID Writing Test's external and diagnostic validity using the technique of direct magnitude estimation. External validity herein refers to the relationship between ratings assigned to essays by experienced, professional judges (i.e., English teachers) and ratings assigned to the same essays by inexperienced non-professionals. Diagnostic validity refers to the degree to which judges are able to separately rate different dimensions of writing skill such as organization, content, language use, and vocabulary. The results of this study attest to the external validity of the NTID Writing Test, but they do not support the diagnostic validity of this and other writing tests of the same basic design. The implications of these findings are discussed with respect to the issue of rater bias, the metric properties of rating scales, and the perception of overall writing quality and of components of writing ability.*

In the last decade, schools, researchers, and testing agencies have relied increasingly on direct methods of assessing writing ability (Charney, 1984; Greenberg, Wiener, & Donovan 1986; Huot, 1990). Direct methods of writing assessment require that judges rate the quality of writing samples on a category scale (e.g., a holistic rating scale with values ranging from 1 to 6 points, a "poor" to "excellent" scale, or a 0- to 100-points scale). Direct approaches to writing assessment appear to have face validity. In contrast, indirect approaches seem to lack face validity because they typically are biased toward one or another aspect of writing ability (e.g., mechanics, grammar, or some aspects of style) while disre-

garding other important dimensions of writing (e.g., organization and content). Critics of indirect measures argue that a valid assessment of writing ability must include more than an evaluation of the examinee's mastery of conventions (Charney). Direct measures allow examinees to draw upon a range of linguistic and rhetorical knowledge, as well as upon knowledge of the conventions of the written language.

While judgments of writing samples appear to be valid tools for the assessment of the construct of global writing skill, the use of judgments raises additional validity questions. One issue pertains to the accuracy of experienced, professional judges' ratings of writing samples. In a review of the direct writing assessment literature over the last fifteen years, Huot (1990) emphasized the importance of judges' experience, asserting that "reader expectation shaped by personal and professional experience will always be a strong yet hard to define influence" (p. 255). Professional judges may produce biased evaluations based on their specialized experiences with particular populations of students. This may result in examinees being rank-ordered in writing ability in a way that is different from rankings assigned by inexperienced readers. The validity issue here is how well professional judges' ratings of writing ability generalize to other people's perceptions of writing ability. We will call this the *external validity* issue.

A second issue is the specificity of professional judges' ratings of particular properties of writing samples. Certain assessment protocols (i.e., analytic or multiple-trait scoring procedures) may require judges to separately rate the same writing sample on different subscales such as organization, content, and style. The purpose may be to ensure consistency of focus and emphasis among judges, or it may be to obtain diagnostic information about the examinees' strengths and weaknesses. According to Huot (1990), the majority of research over the last fifteen years indicates that readers tend to be more concerned about content and organization than sentence structure and mechanics. However, it is not clear that judges are actually capable of making discrete evaluations of component dimensions of writing skill. It may be difficult for judges to suppress their perception of one dimension when rating another. If so, then the diagnostic value of subscale ratings would be compromised. We will call this the *diagnostic validity* issue.

### NTID Writing Test

In this paper we examine both of these issues using the NTID Writing Test. Although this test was developed to assess the writing ability of post-secondary deaf students enrolled at the National Technical Institute for the Deaf (NTID), there is nothing in its design that restricts its use to

this population. It is appropriate for use with younger deaf students or with second-language learners. In fact, the design of the test is based to a large extent on work with adult second-language learners at the University of Hawaii (Jacobs, Zinkgraf, Wormuth, Hartfiel, & Hughey, 1981). Researchers and teachers have long noted similarities in the skill profiles and written English of deaf and second-language learners (Bochner, 1982; Bochner & Albertini, 1988; Goldberg & Boardman, 1974; Swisher, 1989).

Examinees are given thirty minutes to write an essay on an assigned topic. They are instructed to write as much as they can within the allotted time and are encouraged to use their best English. The rating scale used for the NTID Writing Test was adapted from the ESL Composition Profile developed by Jacobs and her colleagues (1981). Three independent judges then evaluate each essay in terms of four categories: organization, content, language use, and vocabulary. Each category is rated on a 0 to 25-point scale, and the subscores for each category are added together to form a composite score. The three judges' composite scores are averaged together to derive the writing test summary score. The ESL Composition Profile, in contrast, is comprised of the same four categories plus a fifth category, mechanics. Each of the five categories in the ESL Composition Profile has a different weight: content (30 points), organization (20 points), language use (25 points), vocabulary (20 points), and mechanics (5 points). A summary description of each subscale of the NTID Writing Test is included in Appendix A.

The decision to score four separate categories is, in part, intended to help maintain consistency among judges and is felt to be particularly useful in training new judges. Each of the professional judges used in this study has had considerable experience teaching English to deaf and hard-of-hearing students, and each participated in the development of the NTID Writing Test. The scoring criteria represent a consensus of their judgments. The training of new judges involves a somewhat formal, specialized, and time-consuming process whereby their judgments are brought into general agreement with the scoring criteria through the analysis of writing samples and discussion with experienced judges. New judges must demonstrate their proficiency by scoring a set of 25 pre-selected essays at a specific level of accuracy relative to the test developers' ratings for those essays.

Interrater reliabilities for the NTID Writing Test range from .61 to .87, with the average reliability for a single rater estimated at .75 with the use of Fisher's  $r$  to  $z$  transformation. Alpha reliabilities for the writing test summary scores, which represent the average of the scores assigned by three raters, range from .83 to .91 for all combinations of three raters. The rating scale, scoring procedure, and other details pertaining to the devel-

We examined both logarithmic and polynomial functions for their ability to fit the data better than a simple linear function. The simple linear function that best fit the data yielded a significant R-squared value of .868 [ $F(1,23) = 151.45, p < .0001$ ]. The logarithmic function that best fit the data yielded a significant R-squared value of .891 [ $F(1,23) = 187.251, p < .0001$ ]. The polynomial function that best fit the data was the second order polynomial shown in Figure 1. This function out-performed both the linear and logarithmic fits, yielding a significant R-squared value of .942 [ $F(2,22) = 179.37, p < .0001$ ]. A test for curvilinearity indicated that the polynomial function was significantly better than the linear function in fitting the data [ $F(1,22) = 28.20, p < .0001$ ].

### Diagnostic Validity

The results of the two PCAs are presented in Table 1. The 25-sample PCA produced only one factor that accounted for 94.5% of all variation among test scores. The factor scores on the four subscales are all quite high, indicating that each subscale measured essentially the same underlying dimension of writing skill as every other subscale. The failure of other factors to emerge from this analysis indicates that there was little or no additional systematic variation in writing quality beyond the single dimension of writing quality represented by Factor 1.

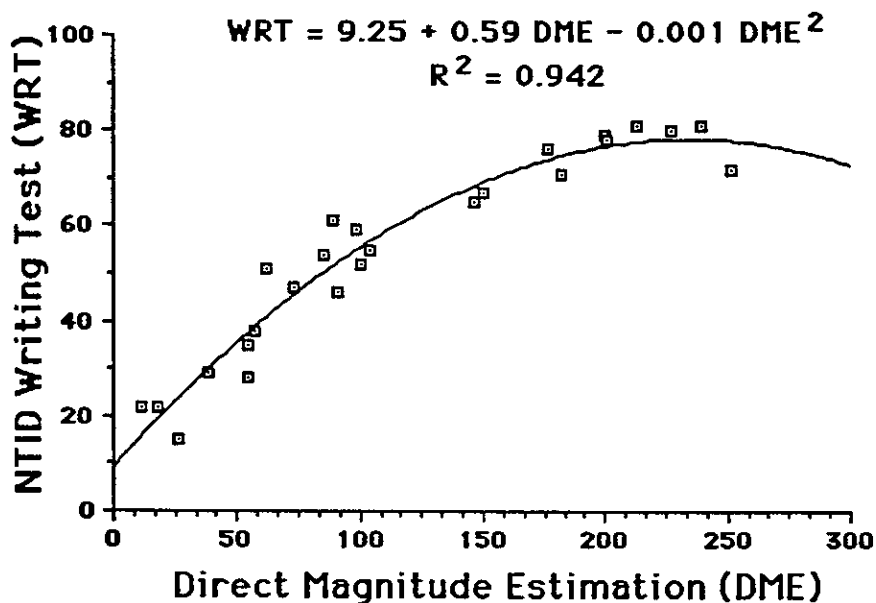


Figure 1. NTID Writing Test summary scores plotted as a function of direct magnitude estimation (DME) scores.

Table 1  
Factor Loadings from Principal Components Analysis (PCA)

	25-Sample PCA Factor 1*	654-Sample PCA Factor 1*
Organization	.967	.918
Content	.965	.931
Language Use	.975	.939
Vocabulary	.982	.943

\*Note: Factor 1, in the strict sense, is not defined precisely. It can be viewed as something like *overall writing quality*, but it is not exhausted by such a definition.

The large 654-sample PCA confirmed that the unidimensional factor structure of the 25-sample PCA is a general and robust property of the NTID Writing Test and not a peculiarity of the small experimental sample used in this study. This PCA accounted for 87% of all variation among test scores. Although this was slightly less than the amount of variance accounted for in the smaller 25-sample analysis, there were still no additional factors that emerged in the larger analysis. This indicates that little or no additional systematic variance in writing quality remained to be explained beyond that captured by the single emergent factor in this analysis.

### Discussion

#### External Validity

The results presented above suggest that experienced professional and inexperienced non-professional judges assign essentially the same relative order to examinees' writing samples. This is indicated by the simple increasing monotonic function that characterized the relationship between the non-professional judges' DME scores and the professional judges' NTID Writing Test summary scores. Since there was no evidence of substantial reordering of professionals' writing skill ratings by non-professionals, the qualities of writing skill valued by experienced English teachers seem to be in accord with those valued by the educated public. Therefore, it may be assumed that educated readers employ essentially the same underlying psychological construct of writing quality as do experienced English teachers. The NTID Writing Test, and presumably similar direct measures, appear to possess external validity, at least so far as the relative order of examinees' writing skills is concerned.

Nevertheless, the significant nonlinearity displayed in the function relating professional and non-professional judges' ratings raises a further issue regarding the distinction between professional and non-professional judges and regarding the metric properties of writing quality perceptions. The issue is whether the nonlinearity is due to a general effect of judges' experience or to a fundamental incompatibility between the underlying psychophysical function and the properties of rating scales. We will speculate on reasons for the nonlinearity and suggest an experiment to explore the issue further.

The first explanation is based upon the notion that judges' experience with the peculiarities of deaf students' written expression may inflate their judgments of writing quality for examinees in the mid-range of writing ability. Once examinees reach a basic skill level, experienced judges may overlook many characteristic violations of grammatical or stylistic conventions made by deaf writers. Presumably for experienced judges, these errors are well understood, expected, and decipherable. They are, therefore, not likely to be as detrimental to the comprehensibility and communicative effectiveness of written text for experienced judges as they might be for inexperienced judges. Under these circumstances, professional judges would tend to assign ratings to examinees with mid-range skills that are shifted somewhat toward the upper end of the rating scale. This would introduce a nonlinearity into the function that relates underlying writing ability to experienced judges' ratings.

This sort of experiential bias has been observed in studies of the intelligibility of deaf students' connected speech (Samar & Metz, 1988). Generally, a bias of this sort would not cause a fundamental reordering of examinees at different measured skill levels since the magnitude of the inflation is a direct function of underlying ability level. Rather, it would merely distort the psychophysical function through local, skill dependent shifts in the perceived magnitude of inter-examinee skill differences.

The notion that judges' experience may influence their evaluation of writing quality has been raised with regard to evaluating the quality of writing produced by students of English as a second language. For example, Stansfield and Ross (1988), in proposing a research agenda for the Test of Written English, have suggested that reader characteristics such as the following be investigated: age, profession, years of teaching experience, essay reading experience, field of teaching, foreign language background, experience living abroad, and frequency of interacting with foreigners. In addition, Carlson (1988), in a contrastive rhetoric study, has concluded that different communities of readers bring different approaches, labels, and definitions to an evaluation task. Similarly, the results of a study conducted by Park (1988) indicate that experiential

variables, such as linguistic (cultural) and academic background, can influence the evaluation of writing performance.

Another possible cause of nonlinearity relates to the metric properties of rating scales. In the present experiment, the DME ratings of non-professional judges may be taken as unbiased estimates of the psychophysical function mapping writing ability to the perception of writing quality. This is simply because the non-professionals are by definition inexperienced. Therefore, the nonlinearity between the professional and non-professional judgments might reflect the scale distortion effect of experience described above. However, it is also possible that the psychophysical function itself is not an equal interval function, but rather an equal ratio function. In this case, the nonlinearity in the function relating the two groups of judges might be due to the inappropriateness of an equal interval rating scale as a metric tool (Stevens, 1975). In this regard, Hamp-Lyons and Henning (1991) asked experienced professional judges to evaluate writing tests on category rating scales and reported a greater dispersion of scores at the upper end of the nine-point scoring range than at the lower end. These data imply that rating scales for writing ability are not composed of equal intervals.

In order to disentangle these two possible causes of the nonlinearity, the following experiment could be carried out. Professional and non-professional judges could be asked to evaluate writing samples using both DME and the NTID Writing Test category rating scale. This would entail substantial training of the non-professionals on the category scale scoring procedure. While the training would sensitize non-professional judges to the biases of experienced English teachers, the non-professionals would still be inexperienced with regard to the characteristic writing behavior of deaf students.

If the nonlinearity is due to professional experience with the population, we would expect to see it preserved in the plots of professional versus non-professional category scale ratings. We would also expect it to be preserved in the DME plots of these groups against each other, because the effects of experience should cause a similar inflation of the DME ratings by professional judges. However, if the nonlinearity is due to fundamental scale incompatibilities, then both the DME and the category scale plots of professional versus non-professional ratings would be linear, whereas the DME versus category scale plots for each group of judges separately would be nonlinear. Based on the finding of Hamp-Lyons and Henning (1991) mentioned above, we would expect the results of an experiment such as this to indicate that the width of intervals at the upper end of the rating scale is larger than the width of intervals in the low and middle regions of the scale. The simultaneous influence of an experiential

bias also may be demonstrated. However, determining the cause of the nonlinearity is clearly an issue for future research.

### *Diagnostic Validity*

The principal components analysis data presented above revealed a strong single-factor structure underlying the four sets of subscores. The absence of any clear multidimensional structure to the data set indicates that the subscores should not be used for the purpose of differential diagnosis. Although the NTID Writing Test has very high validity as a predictor of overall writing performance outside of the English classroom, its subscores provide no distinct information about specific dimensions of writing ability.

It is important to remember that these results do not preclude consideration of organization, content, language use, and vocabulary as bona fide components of the underlying writing quality construct. These components appear to be conceptually and intuitively distinct. Furthermore, they tend to rely upon different, often modularly distinct levels of language organization (Chomsky, 1986) such as syntax, lexical structure, and strategic and schematic conceptualization. Similar components of writing quality (i.e., content, organization, style and tone, surface features, and personal response of the reader) have been identified in a recent comparative study of achievement in written composition (Purves, 1992). Therefore, it is likely that these underlying dimensions varied meaningfully in the large-scale 654-sample factor analysis of this study. However, it appears from the factor results that judges are not capable of selectively rating the quality of an individual's control of these component subskills. Perhaps judges are not able to suppress their perceptual experience of one dimension when rating another (Carlson, 1988; Marsh & Ireland, 1984).

The question here is whether a judge can perceive and rate subskill categories separately. Some research pertaining to this question has been conducted in the area of speech intelligibility. A series of studies on the dimensional complexity of the speech intelligibility construct and its relationship to judges' ratings on both category and DME scales has revealed that psychometric ratings do not preserve the multidimensional structure of the underlying construct. By contrast, acoustic and articulatory measures of phonetic aspects of speech do reveal the expected multidimensional structure of the intelligibility construct and its significant multivariate relationship to global perceptual measures of intelligibility in connected speech (Metz, Samar, Schiavetti, Sitler, & Whitehead, 1985; Metz, Schiavetti, Samar, & Sitler, 1990; Samar & Metz, 1988, 1991). This demonstrated mismatch between objective acoustic and articulatory measures of speech communication and subjective psychometric mea-

asures of speech communication confirms that judges are sometimes incapable of separately attending to and assessing the quality of component dimensions known to exist in spoken utterances. The definitive experiment remains to be done with writing samples—comparing the factor structures of objective measures of language use and vocabulary, for example, to those of subjective ratings of these subskills.

Finally, since the factor analysis did not reveal a multifactor structure to the NTID Writing Test, the issue of the choice of appropriate subtest weights in the formula for computing overall summary scores does not arise. Any consistent choice of weights will result in the same rank ordering of students within the tolerance of measurement error.

### **Conclusion**

For the NTID Writing Test, the results of this study show that experienced professional judges' ratings of writing quality do in fact generalize to ratings produced by inexperienced non-professionals. These results, therefore, attest to the external validity of the test. However, the four subscales comprising the NTID Writing Test were found to be neither multidimensional nor multivariate. This finding indicates that subscores on the NTID Writing Test should not be used for the purpose of differential diagnosis and raises questions about the diagnostic validity of other tests of the same basic design.

The failure to demonstrate diagnostic validity implies that scores on the NTID Writing Test and on similar measures of writing skill should be treated as holistic ratings. This raises the issue of whether it might not be better for the NTID Writing Test to eliminate separate subskill ratings altogether and convert to a conventional holistic scoring rubric. The use of holistic scoring would be advisable for the NTID Writing Test if a sufficient level of reliability could be demonstrated. In addition, the use of DME ratings should be considered as an alternate means of scoring writing tests because DME ratings are more sensitive to differences in writing ability at the upper end of the rating scale.

### **References**

- Albertini, J., Bochner, J., Cuneo, C., Hunt, L., Nielsen, R., Seago, L., & Shannon, N. (1986). Development of a writing test for deaf college students. *Teaching English to Deaf and Second-Language Students*, 4(2), 5-11.
- Bochner, J. H. (1982). English in the deaf population. In D. G. Sims, G. G. Walter, & R. L. Whitehead (Eds.), *Deafness and communication: Assessment and training* (pp. 107-123). Baltimore: Williams and Wilkins.

- Bochner, J. H., & Albertini, J. A. (1988). Language varieties in the deaf population and their acquisition by children and adults. In M. Strong (Ed.), *Language learning and deafness* (pp. 3-48). New York: Cambridge University Press.
- Carlson, S. B. (1988). Cultural differences in writing and reasoning skills. In A. C. Purves (Ed.), *Writing across languages and cultures: Issues in contrastive rhetoric* (pp. 227-260). Newbury Park, CA: Sage.
- Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English*, 18, 65-81.
- Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. New York: Praeger.
- Goldberg, J. P., & Boardman, M. B. (1974). English language instruction for the hearing impaired: An adaptation of ESL Methodology. *TESOL Quarterly*, 8, 263-270.
- Greenberg, K. L., Wiener, H. S., & Donovan, R. A. (1986). *Writing assessment: Issues and strategies*. New York: Longman.
- Hamp-Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. *Language Learning*, 41, 337-373.
- Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research*, 60, 237-263.
- Jacobs, H., Zinkgraf, S., Wormuth, D., Hartfiel, V., & Hughey, J. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.
- Marsh, W., & Ireland, R. (1984). *Multidimensional evaluations of writing effectiveness*. (Report No. 840203). Sydney, Australia: University of Sydney (ERIC Document Reproduction Service No. ED 242 785).
- Metz, D. E., Samar, V. J., Schiavetti, N., Sitler, R. W., & Whitehead, R. L. (1985). Acoustic dimensions of hearing-impaired speakers' intelligibility. *Journal of Speech and Hearing Research*, 28, 345-355.
- Metz, D. E., Schiavetti, N., Samar, V. J., & Sitler, R. W. (1990). Acoustic dimensions of hearing-impaired speakers' intelligibility: Segmental and supra-segmental characteristics. *Journal of Speech and Hearing Research*, 33, 476-487.
- Park, Y. M. (1988). Academic and ethnic background as factors affecting writing performance. In A. C. Purves (Ed.), *Writing across languages and cultures: Issues in contrastive rhetoric* (pp. 261-272). Newbury Park, CA: Sage.
- Purves, A. C. (1992). Reflections on research and assessment in written composition. *Research in the Teaching of English*, 26, 108-122.
- Samar, V. J., & Metz, D. E. (1988). Criterion validity of speech intelligibility rating scale procedures for the hearing-impaired population. *Journal of Speech and Hearing Research*, 31, 307-316.
- Samar, V. J., & Metz, D. E. (1991). Scaling and transcription measures of intelligibility for populations with disordered speech: Where's the beef? *Journal of Speech and Hearing Research*, 34, 699-702.
- Stansfield, C. W., & Ross, J. (1988). A long-term research agenda for the Test of Written English. *Language Testing*, 5, 160-186.
- Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural, and social prospects*. New York: Wiley-Interscience.
- Swisher, M. V. (1989). The language-learning situation of deaf students. *TESOL Quarterly*, 23, 239-257.

## Appendix A

### Summary Description of NTID Writing Test Scoring Categories

*Organization* (25 points) includes such features as

- clear statement of topic placed appropriately
- intent is evident to reader
- plan of paper recognized by reader (i.e., paper is unified and coherent)
- appropriate transitions (e.g., markers and clear paragraphing)

*Content* (25 points) includes such features as

- paper addresses the assigned topic
- generalizations are supported by examples
- no extraneous material
- pertinence and noteworthiness of ideas

*Language Use* (25 points) includes such features as

- correct use of grammatical structures (sentence and discourse level) and punctuation
- correct use of complex structures
- intelligible spelling
- variety of style and expression
- clarity of reference

*Vocabulary* (25 points) includes such features as

- appropriate semantic use of vocabulary
- consistent register
- sophisticated choice of vocabulary
- appropriate use of figurative and idiomatic expressions

## Appendix B

### Instructions for DME Ratings

You are being asked to make judgments about the quality or "goodness" of student writing. You will read 25 short essays and judge how good each one is. The essays were written in response to the following topic:

You are in a new place with new people.  
 What do you like about NTID and the people here?  
 What don't you like about NTID and the people here?  
 Explain.

The essays span a wide range of quality. Some are very good, some are very poor, and some are in between. The attached samples illustrate this range of quality. Take a few minutes now to read the attached sample essays.

In judging writing quality, one essay will serve as the "standard" by which you will judge all other essays. The standard will have a quality (goodness) in the