On the Construct Validity of Multiple-Choice Items for Reading Comprehension

Huub van den Bergh
University of Utrecht

In this study 590 third-grade students took one of four reading comprehension tests with either multiple-choice items or open-ended items. Each also took 32 tests indicating 16 semantic Structure-of-Intellect (SI) abilities. Four conditions or groups were distinguished on the basis of the reading comprehension tests. The four 33 × 33 correlation matrices were analyzed simultaneously with a four-group LISREL model. The 16 intellectual abilities explained approximately 62% of the variance in true reading comprehension scores. None of the SI abilities proved to be differentially related to item type. Therefore, it was concluded that item type for reading comprehension is congeneric with respect to the SI abilities measured. *Index terms: construct validity, item format, free response, reading comprehension, Structure-of-Intellect model.*

Since the early 1920s the question of whether items in different formats tap the same mental functions has been an active area of research. Unfortunately, the results of different studies are equivocal. Some authors have claimed that there are no differences between tests with different item formats (Bracht & Hopkins, 1968; Carter & Crone, 1940; Cook, 1955; Weiss & Jackson, 1983), whereas others reached the opposite conclusion (Birenbaum & Tatsuoka, 1987; Coombs, Milholland, & Womer, 1956; Heim & Watts, 1967; Traub & Fisher, 1977; Vernon, 1962; Ward, Frederiksen, & Carlson, 1980).

A theoretical basis has been lacking in most studies. In most cases differences in observed scores due to item format are interpreted and explained in relation to the differences between recall and recognition. In this case the essence of recall is the generation or construction of an answer to an item. The respondent must formulate the answer in oral or written form, or must describe an idea if that is the answer desired. The essence of recognition, on the other hand, is that one or more alternatives are presented to the respondent; there is no requirement for overt generation of an answer. The respondent may select one of two strategies: a strategy in which an answer is generated or constructed, and a compare-and-delete strategy on the basis of the cues provided in the alternatives. Hence recognition is sometimes mediated by process characteristics of recall, that is, processes that do not depend on the presence of alternatives (see Brown, 1976; Gillund & Shiffrin, 1984). The recall/recognition paradigm encompasses differences in verbal abilities tested by means of open-ended and multiple-choice items (Doolittle & Cleary, 1987; Murphy, 1982; Ward, 1982; Ward et al., 1980), as well as the assessment of partial knowledge measured by means of multiple-choice items (Coombs et al., 1956).

The lack of a theoretical framework might be one of the reasons for the different research conclusions. For instance, some researchers interpret differences in means between two tests, or differences in observed variances, as indications for a difference in the intellectual abilities tested (see

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 14, No. 1, March 1990, pp. 1–12
© Copyright 1990 Applied Psychological Measurement Inc.
0146-6216/90/010001-12$1.85

Shohamy, 1984; van den Bergh, 1987). This seems a rather hasty conclusion, because the same process might be involved both in recognition and in recall.

There are other reasons for the differences in research results that may be of even greater importance. In none of the studies is an effect size specified; therefore the null hypothesis can always be rejected (see Cronbach, 1966). Nevertheless several conceptions are implicit with regard to the size of the effects. For instance, some authors test whether the correlation coefficients (corrected for test unreliability) between scores on tests differing in item format differ from unity (Hogan, 1981; Mellenbergh, 1971). Others interpret correlation coefficients of at least .80 as proof of the null hypothesis, which is that there are no differences in intellectual skills measured on tests with open-ended and multiple-choice items (Hurd, 1930; Paterson, 1926).

A second inadequacy concerns the research design. Most research is carried out according to a repeated-measurement design. Almost invariably, the test with open-ended items is administered first, followed by one or more tests in alternative formats. Then the product-moment correlation between the scores on the different tests is calculated. If a respondent must answer the same question, although in a different format, up to six times in a few weeks (Traub & Fisher, 1977), it is questionable whether the same traits are measured on the first and on the last administration.

The administration procedures can be improved by randomizing the tests administered on each occasion (Hogan, 1981). In this type of study the main interest lies in the shared variance of the scores on tests differing in item format. Therefore, a repeated-measurement design in which all tests are administered at all occasions seems the most appropriate solution (van den Bergh, Eiting, & Otter, 1988). In that case the amount of variance shared by the scores on two tests different in item format can be compared to the amount of variance shared by the scores on tests with the same format.

A third problem seems to be the questionable use of psychometrics. Correlation coefficients corrected for attenuation on the basis of the often-used coefficient α may be an overestimate of the true correlation between two variates (Lord & Novick, 1968, p. 138). Also, in many studies the criterion validity of tests differing in item format is investigated by testing differences in the correlation coefficients of scores on both types of test with a criterion (Culpepper & Ramsdale, 1983; Hopkins, George, & Williams, 1985; Hopkins & Stanley, 1981; Weiss & Jackson, 1983). Hence it is assumed that both test scores have an equal regression on the true score; it is assumed that both types of test are at least tau-equivalent.

Strictly speaking, this assumption of tau-equivalence is not necessary; the research question translated into a testable form only concerns the congenericity (Jöreskog, 1971) of different test types. That is, the tests measure the same trait(s) except for errors in measurement. The scores on tests with multiple-choice items have a higher mean (Benson, 1981; Benson & Crocker, 1979; Kinney & Eurich, 1938; Samson, 1983; Shohamy, 1984) and lower observed and true-score variances than scores on tests with open-ended items (Carter & Crone, 1940; Croft, 1982; Frary, 1985; van den Bergh, 1987). Therefore, tests differing in format cannot be considered tau-equivalent a priori; in all cases a special test is needed to see whether this assumption is met.

This study was not only concerned with whether there were differences in intellectual abilities tested due to item format, but was also concerned with the nature of any differences observed. Hence it is insufficient to take into account only theoretical notions of differences in item format. Intellectual abilities that are differentially involved in answering open-ended and multiple-choice items must be defined and taken into account as well.

One of the most concise models for intellectual abilities is the Structure-of-Intellect (SI) model (Guilford, 1971, 1979; Guilford & Hoepfner, 1971). In the SI model, abilities are defined on three dimensions. The *operations* dimension concerns the intellectual activities to be performed. Five operation types are distinguished: cognition (C), divergent production (D), evaluation (E), memory (M), and convergent production (N). The *contents* dimension characterizes basic kinds or areas of information: behavioral (B), figural (F), semantic (M), and symbolic (S) information. The *products* dimension concerns the organization of the information to be processed. A distinction is made between classes (C), implications (I), relations (R), systems (S), transformations (T), and units (U). The intersection of the three dimensions is indicated by a trigram; for example, CMT denotes cognition of semantic transformations, or the ability to see changes in meaning or interpretation.

The SI abilities involved in carrying out an assignment are, of course, dependent on the subject matter. In this study, where the subject area was reading comprehension, the primary interest was in abilities having to do with meaning or understanding. Therefore this study was restricted to the semantic part of the SI model (see Meuffels, 1982; Spache, 1963).

The difference between recognition and recall can be translated into semantic SI abilities. Recall, which refers to the retrieval of items from memory storage in order to meet certain objectives, can be assigned to the SI abilities of divergent and convergent production (Guilford, 1971). *Divergent production* refers to an intellectual ability that pertains primarily to information retrieval. *Convergent production* is the prevailing function when input information is sufficient to determine a unique answer, which also must be retrieved from memory.

In recognition two strategies may be involved. Convergent and divergent abilities may be essential if a respondent answers multiple-choice questions in the same way as open-ended questions. On the other hand, the presentation of alternatives may give rise to differences in cognition and evaluation abilities. Cognition abilities refer to immediate awareness, immediate discovery, or rediscovery of information, whereas evaluation abilities concern comparison according to a set criterion and making a decision about criterion satisfaction. Both seem important: The respondent may ''know'' the answer to a multiple-choice item as soon as the alternatives are seen, or may actually compare the different alternatives presented. Besides, the cognition abilities (e.g., CMS, CMT, and CMU) are crucial for reading comprehension (Guilford, 1979; Hoeks, 1985; Meuffels, 1982); for instance word knowledge, or CMU, is one of the best (single) predictors of reading comprehension (Guilford, 1979; Hoeks, 1985; Mezynsky, 1983; Sternberg & Powell, 1983).

This study explored the question of whether items for reading comprehension are congeneric irrespective of the format of the items, that is, whether they are congeneric with respect to the abilities tested. This general question can be formulated more precisely: It was hypothesized that open-ended questions for reading comprehension appeal to divergent- and convergent-production abilities, whereas cognition and evaluation abilities are prevalent in multiple-choice items. These differences will be observed in differences in the regression weights of reading comprehension scores on the SI ability scores.

Another question investigated was: What SI abilities are involved in answering items in traditional reading comprehension tests? It might be argued that apart from the already mentioned SI abilities (cognition, evaluation, convergent-production, and divergent-production), SI memory abilities might also be of importance for the answering of reading comprehension items.

The role of memory abilities is emphasized in current theories of text comprehension. In these models it is asserted that (experienced) readers identify and relate text topics as they read (Just & Carpenter, 1980; Kintsch & van Dijk, 1978; Lorch, Lorch, & Matthews, 1985). Hence the answering of reading comprehension items might require the activation and revision of the text representation (Lorch, Lorch, & Mogan, 1987). Therefore, an efficient method of encoding information seems crucial. The import of topic structure might then be reflected in the memory for semantic systems (MMS) and the evaluation of semantic systems (EMS); see Ackerman (1984a, 1984b). In short, it was hypothesized that memory abilities (especially MMS) are crucial for the answering of reading comprehension items, but are not differentially involved in answering open-ended and multiple-choice items.

Method

Examinees

The examinees were 590 third-graders from 12 different Dutch high schools for lower vocational education or lower secondary education. Each student was randomly assigned to one of the four conditions. Conditions were defined on the basis of the reading comprehension test administered (see Table 1). Administration time for all reading comprehension tests was 100 minutes.

Dependent Variables

Two reading comprehension tests were constructed on the basis of two central Dutch language exams for lower vocational education. All items were open-ended; some were essay questions and others were short-answer questions. The tests were constructed so that the distribution of items, according to the categories of a classification scheme for reading comprehension items (Davis, 1968), was equal.

The two tests were pretested in another group of 438 third-graders from both types of secondary education. The alternatives of the four-choice items constructed were based on incorrect answers provided by the students. This resulted in four tests with 25 items each: two tests with the original open-ended items and two tests with the same items in a multiple-choice format.

In the final study each student was randomly assigned to one of the four reading comprehension tests. The answers of the students to the open-ended items were all rated once by an experienced rater, whose stability proved to be high as indicated by the correlation between total reading comprehension scores on the first rating and on a second rating three weeks later ($r = .91$, $N = 50$). The answers of 100 students were also rated by a second rater. The agreement between both raters was satisfactory ($r = .86$, $N = 100$).

Because this study used a randomized group design, every respondent took only one reading comprehension test. Therefore, every respondent answered the reading comprehension items only once in either a multiple-choice or an open-ended format. Hence any dependency of the answers due to memory effects from an earlier to a later administration was impossible. The SI ability tests were administered to all respondents.

Table 1 presents summary data for the four reading comprehension tests. In addition to the reliability estimate α of a reading comprehension test, the split-half reliability estimate is reported. The split-half estimate was obtained by ordering the items on the basis of both their difficulty (p) values and item-test correlations. The two items nearest on both dimensions were assigned randomly to one of the two subtests (Gulliksen, 1950). Then the reliability estimate was obtained by application of the Spearman-Brown formula for parallel tests to the correlation between the subtests.

As could be expected, the means of both multiple-choice tests were higher than the means of the same tests in open-ended item format. None of the tests was very easy, nor do they appear to have been extremely demanding. The standard deviations and the homogeneity of both multiple-choice tests are lower than the respective figures for the

Table 1
Number of Examinees (N), Mean,
Standard Deviation (SD), Coefficient α,
and Split-Half Reliability Estimates (r_{xx})
for Two Reading Comprehension Tests

Item Format	N	Mean	SD	α	r_{xx}
Test A					
Open-ended	167	15.11	4.06	.84	.84
Multiple-choice	129	16.26	3.16	.63	.64
Test B					
Open-ended	149	14.02	4.08	.83	.83
Multiple-choice	145	17.37	3.57	.63	.63

open-ended tests. The small differences between the two reliability estimates are also worth noting; coefficient α is only a fraction lower than the theoretically more sound split-half reliability.

SI Ability Tests

In this study 16 of the 30 semantic abilities were measured by means of two tests each (see Table 2). For every ability a choice was made of two of the three tests designed to measure a specific SI ability (van den Bergh, 1989). The main criterion was the proportion of shared variance; the two tests with the highest communality on the SI ability (from a previous analysis; see van den Bergh, 1989) were selected (see Table 2).

Almost every student took all 32 tests; only about 10 percent of the students had missing scores on one or two tests. The correlations between the achievement scores on the SI tests supported the SI model with oblique ability factors; other models based on theories of Cattell (1963), Spearman (1927), Guttman (1965), and Thurstone (Thurstone & Thurstone, 1941) did not fit the data as well. The adjusted goodness-of-fit index and the root-mean-square residual for this model were estimated at .94 and .07, respectively (van den Bergh, 1989).

As can be seen in Table 2, the majority of the tests shared most of their variance with the other tests for the same ability. There are some exceptions, however; the communality of the test "Time order" is low, for instance. The estimated correlations between factors were low, ranging from −.121 (between EMU and MMU) to .950 (between NMS and NMT), with a mean of .224.

Analyses

Four correlation matrices were calculated. For each group in which the respondents took the same reading comprehension test, the correlations among the achievement scores on the SI tests and the correlation with the reading comprehension score were calculated.

The four correlation matrices were analyzed simultaneously with LISREL (Jöreskog & Sörbom, 1986). Because of the criticism on model fitting, which in essence comes down to chance capitalization, a cross-validation was performed on the variables and the models to be fitted were specified in advance.

Because every SI ability was measured with two tests, the variable set could be split into two parts. In each variable set each SI ability was measured by one test. If the results converged, there was additional evidence of the significance of the model accepted. Only in the final analysis could all variables be taken into account, because in that case the SI factors were better defined, resulting in smaller standard errors of the regression weights. On the other hand, specification of the models to be fitted minimized chance capitalization, because none of

Table 2
Estimated Communalities for the SI Tests

SI Test	SI Ability	Communality
Family relations	CMS	.632
Transitivities	CMS	.578
Different meanings	CMT	.626
Word picture translations	CMT	.544
Accuracy of word meanings	CMU	.646
Rote word knowledge	CMU	.669
Constrained associations	DMR	.762
Number associations	DMR	.412
Different customs	DMU	.579
Different problems	DMU	.619
Inferences	EMI	.507
Syllogism	EMI	.604
Word-matrix evaluation	EMS	.630
Title fabrication	EMS	.584
General remarks	EMU	.812
Sensible sentences	EMU	.823
Name and animal	MMI	.845
Name and profession	MMI	.863
Memorizing facts	MMS	.616
Memorizing orders	MMS	.904
Memorizing double meanings	MMT	.388
Memorizing homophones	MMT	1.000
Memorizing pictures	MMU	.716
Memorizing words	MMU	.741
Missing links	NMI	.440
Order associations	NMI	.878
Contradistinctions	NMR	.661
Vocabulary completion	NMR	.869
Time order	NMS	.368
Sentence order	NMS	.563
Object synonyms	NMT	.663
Unusual objects	NMT	.679

the model specifications was altered in order to obtain a better fit.

Models

Two measurement models and one structural model were specified for the analyses. Both measurement models define the relation between the observed and the latent scores, whereas the structural model specifies the relation between the latent variables. The two measurement models are:

$$\mathbf{y}^{(g)} = \boldsymbol{\Lambda}_y^{(g)}\boldsymbol{\eta}^{(g)} + \boldsymbol{\varepsilon}^{(g)} \qquad (g = 1, 2, 3, 4) \tag{1}$$

$$\mathbf{x}^{(g)} = \boldsymbol{\Lambda}_x^{(g)}\boldsymbol{\xi}^{(g)} + \boldsymbol{\delta}^{(g)} \qquad (g = 1, 2, 3, 4) \tag{2}$$

In Equation 1 $\mathbf{y}^{(g)}$ refers to the observed reading comprehension score in the gth subpopulation; $\boldsymbol{\Lambda}_y^{(g)}$ refers to the regression of the observed reading comprehension score on the latent score $\boldsymbol{\eta}^{(g)}$, with unity as variance because correlational analyses are performed. The last term in Equation 1, $\boldsymbol{\varepsilon}^{(g)}$, is the residual of the reading comprehension score, with variance $\boldsymbol{\Theta}_\varepsilon^{(g)}$. Because in every subpopulation only one reading comprehension test is administered, $\mathbf{y}^{(g)}$ refers to one dependent variable (which is therefore restricted; see below).

In Equation 2, $\mathbf{x}^{(g)}$ denotes a column vector of 16 independent variables (in the final analysis, 32 variables) measured in the gth subpopulation. $\boldsymbol{\Lambda}_x^{(g)}$ refers to a matrix of regression weights of 16 variables on 16 factors (in the final analysis, 32 independent variables on 16 factors) on the latent vector $\boldsymbol{\xi}^{(g)}$. Finally, $\boldsymbol{\delta}^{(g)}$ refers to the residuals of the SI tests, with variance $\boldsymbol{\Theta}_\delta^{(g)}$.

It is assumed that there is neither covariance between $\boldsymbol{\eta}^{(g)}$ and $\boldsymbol{\varepsilon}^{(g)}$ nor between $\boldsymbol{\xi}^{(g)}$ and $\boldsymbol{\delta}^{(g)}$. Furthermore it is assumed that there is no covariance among or between the residuals $\boldsymbol{\varepsilon}^{(g)}$ and $\boldsymbol{\delta}^{(g)}$. Therefore the matrix $\boldsymbol{\Theta}_\delta^{(g)}$ is diagonal with the residual variances of the SI ability scores, and $\boldsymbol{\Theta}_\varepsilon^{(g)}$ is a matrix with only one element. The correlation between the SI factors can be estimated in the matrix $\boldsymbol{\Phi}^{(g)}$, which is a symmetric matrix with unities on the diagonal.

In the structural model the regression of the latent reading comprehension ability score on the latent SI abilities is defined as

$$\boldsymbol{\eta}^{(g)} = \boldsymbol{\Gamma}^{(g)}\boldsymbol{\xi}^{(g)} + \boldsymbol{\zeta}^{(g)} \qquad (g = 1, 2, 3, 4) \tag{3}$$

In Equation 3 $\boldsymbol{\Gamma}^{(g)}$ is a row vector with regression weights of the latent reading comprehension score ($\boldsymbol{\eta}^{(g)}$) on the SI ability factors ($\boldsymbol{\xi}^{(g)}$) in the gth subpopulation, and $\boldsymbol{\zeta}^{(g)}$ refers to the unexplained variance of the latent reading comprehension score in the gth subpopulation.

Equation 3 permits simultaneous analysis of the relationships between the latent variables in the four subsamples. Hypotheses can be tested by varying the restrictions over groups on the gamma parameters ($\gamma_j^{(g)}$), while holding other restrictions over groups and analyses invariant.

The invariant restrictions over groups and analyses concern the second measurement model (see Equation 2). Hence restrictions are placed on $\boldsymbol{\Lambda}_x^{(g)}$ and $\boldsymbol{\Theta}_\delta^{(g)}$ and on the correlation between the SI ability factors specified in the matrix $\boldsymbol{\Phi}^{(g)}$. These restrictions are

$$\boldsymbol{\Lambda}_x^{(1)} = \boldsymbol{\Lambda}_x^{(2)} = \boldsymbol{\Lambda}_x^{(3)} = \boldsymbol{\Lambda}_x^{(4)} \tag{4}$$

$$\boldsymbol{\Theta}_\delta^{(1)} = \boldsymbol{\Theta}_\delta^{(2)} = \boldsymbol{\Theta}_\delta^{(3)} = \boldsymbol{\Theta}_\delta^{(4)} \tag{5}$$

$$\boldsymbol{\Phi}^{(1)} = \boldsymbol{\Phi}^{(2)} = \boldsymbol{\Phi}^{(3)} = \boldsymbol{\Phi}^{(4)} \tag{6}$$

With respect to the first measurement model (Equation 1), a set of invariant restrictions over analyses is placed on both the $\boldsymbol{\Lambda}_y^{(g)}$ and $\boldsymbol{\Theta}_\varepsilon^{(g)}$ parameters. In all analyses the $\Lambda_y^{(g)}$ parameters are fixed at the square root of the split-half reliabilities for the four reading comprehension tests (see Table 1), whereas $\boldsymbol{\Theta}_\varepsilon^{(g)}$ is fixed at 1 minus these reliability estimates. For instance, $\Lambda_y^{(1)}$ and $\Theta_\varepsilon^{(1)}$ were fixed at .916 $[(.84)^{1/2}]$ and .16, respectively.

In the first model to be tested, it was assumed that there are no differences in regression of the reading comprehension factor on the SI factors due to the reading comprehension test. It was also assumed that there are no differences between the tests due to the format of the items. Hence the following restriction was imposed:

$$\gamma_j^{(1)} = \gamma_j^{(2)} = \gamma_j^{(3)} = \gamma_j^{(4)} \qquad (j = 1, 2, \ldots, 16) \tag{7}$$

$$\mathrm{Var}(\zeta^{(1)}) = \mathrm{Var}(\zeta^{(2)}) = \mathrm{Var}(\zeta^{(3)}) = \mathrm{Var}(\zeta^{(4)}) \tag{8}$$

where the subscript j refers to the SI vector involved. This model is called the *no-difference model.*

In the second model, the *format model*, the restrictions are loosened a bit; differences in regression weights due to the reading comprehension test are allowed, but again no differences between formats are permitted. This can be written as

$$\gamma_j^{(1)} = \gamma_j^{(2)}, \quad \gamma_j^{(3)} = \gamma_j^{(4)} \quad (j = 1, 2, \ldots, 16) \tag{9}$$

$$\mathrm{Var}(\zeta^{(1)}) = \mathrm{Var}(\zeta^{(2)}), \quad \mathrm{Var}(\zeta^{(3)}) = \mathrm{Var}(\zeta^{(4)}) \tag{10}$$

The difference in fit between the first and second models tested is in fact the effect size referred to above.

In the third model to be tested, no restrictions are placed on the regression of the reading comprehension factor on the cognition, evaluation, convergent-production, and divergent-production ability factors. All of these regression weights, γ parameters, and (variances of the) disturbance terms (ζ) must be estimated. Hence differences in regression weights as well as differences in residual variances are allowed. Note that in this model the regressions of the memory abilities are constrained over item format. This model is called the *difference model.*

Model Fit

Several procedures may be used to assess the fit of a model and to obtain parameter estimates. The maximum likelihood procedure is usually a very efficient method. If the scores have a multivariate normal distribution, the parameter estimates are precise (in large samples) and the fit of a model can be assessed using a chi-square-distributed statistic. However, if the test scores have no multivariate normal distribution, an estimation procedure without assumptions on the distribution of the test scores may be preferable (Jöreskog & Sörbom, 1986, p. 28). Because the SI ability test scores have no multivariate normal distribution (van den Bergh, 1989), the unweighted least-squares estimation method is preferred. In that case the fit of the models must be assessed using a goodness-of-fit index, indicating the total amount of variance and covariance accounted for by the model and the root-mean-square residual (Jöreskog & Sörbom, 1986, p. 40). Note, however, that both statistics have an unknown sampling distribution (Jöreskog & Sörbom, 1986).

With respect to the question of when to reject the null hypothesis, substantial differences in abilities measured due to item format must appear across tests. The abilities measured by two tests with the same item format can then be divided into two kinds: abilities related to the latent trait measured (reading ability) and abilities related to the item format. Of course, two tests with the same format for the same latent trait will not correlate perfectly because of differences in texts (Biemiller, 1977; Klein-Braley, 1983; Neville & Pugh, 1977) or differences in items. Therefore, effect size was related to the differences in abilities between two reading comprehension tests. That is, the results were interpreted as an effect due to item format only if the differences in regression weights of the reading comprehension test scores with a different format on SI ability test scores were larger than the differences in regression weights between two different reading comprehension tests with the same format.

The difference in fit between the format model and difference model is the test of the first hypothesis that there are no differences due to the format of the items. The null hypothesis is rejected only if the difference in fit between the format model and the difference model is larger than the difference in fit between the no-difference model and the format model.

Results

Table 3 shows the fit of all three models for both variable sets. It can be seen in Table 3 that the fit of the no-difference model for both variable sets is clearly worse than that of the format model. Hence it must be concluded that there are differences in regression weights of the reading comprehension factor on the SI factors due to the reading comprehension test involved. The difference in fit between the format model and the difference model, based on either the goodness-of-fit index or the root-mean-square residual, is less than the difference in fit between the no-difference model and the format model. This holds again for both variable sets. According to the criterion specified above, it must be concluded that there is no sub-

Table 3
Goodness-of-Fit Index (GFI) and
Root-Mean-Square Residuals (RMR) of Three Models
for Two Variable Sets for the Relation Between
16 SI Abilities and Reading Comprehension

Model		Variable Set I		Variable Set II	
		GFI	RMR	GFI	RMR
I	No-Difference Model	.834	.095	.901	.089
II	Format Model	.925	.081	.948	.077
III	Difference Model	.946	.077	.955	.071

stantial difference in regression weights of $\eta^{(g)}$ on $\xi^{(g)}$ due to the format of the items. Hence the answering of open-ended and multiple-choice items is congeneric with respect to the SI abilities measured.

In Table 4 the parameter estimates are presented as estimated under the restrictions of the format model (see Equations 9 and 10). The goodness of fit of this model, with all 32 SI variables included, did not differ very much from the goodness-of-fit estimates of the format model presented in Table 3 (goodness-of-fit index = .938; root-mean-square residual = .071).

As can be seen in Table 4, a relatively large part of the variance in the true reading comprehension scores is explained by the 16 SI abilities. The point estimates $(1 - \zeta)$ varied from 62% for Test A to 66% for Test B.

The parameter estimates in Table 4 imply that reading comprehension has a moderate regression on most SI abilities. The high regression weights of the memory factors (MMI, MMS, MMT, and MMU) and EMS coincide with current theories of reading comprehension. The low regression weights of reading ability on CMU indicate that variance in word knowledge did not play an important role in answering the items of both tests.

The negative weights of the two divergent-production abilities (DMR and DMU) indicate that students who had a high score on these tests had a low achievement score on the reading comprehension tests. These negative weights also can be interpreted as suggesting an interaction effect between divergent-production/evaluation abilities and reading comprehension. In order to inspect the data with respect to this hypothesis, the scores for DMR and DMU, and the sum of the scores for the evaluation ability tests (EMI, EMS, and EMU), were dichotomized at their medians. This resulted in two 2 × 2 × 2 cross-tables (see Table 5).

The interaction between the (2 ×) four cell-frequencies in Table 5 for the high-divergent-production ability group was tested by estimating this effect in a loglinear model (Fienberg, 1980) by

$$\hat{\mu}_{ijk} = \sum \beta_{ijk} \log x_{ijk} \qquad (i, j, k = 1, 2) \qquad (11)$$

where $\hat{\mu}_{ijk}$ indicates the interaction effect between reading and evaluation ability for the high-divergent-production ability group, and β_{ijk} is a contrast

Table 4
Parameter Estimates Under the
Restrictions of the Format Model

SI Ability	Test A	Test B
CMS	.376	.251
CMT	.314	.698
CMU	.117	.204
DMR	−.507	−.284
DMU	−.255	−.209
EMI	.407	.356
EMS	.471	.546
EMU	.110	.235
MMI	.465	.114
MMS	.744	1.078
MMT	.633	.716
MMU	.688	.686
NMI	.474	.620
NMR	.306	.268
NMS	.496	.728
NMT	.394	.689
Var(ζ)	.382	.349

Table 5
Cross-Tablulation of the Number of Respondents Per Cell for Divergent Production Abilities (DMR and DMU), Evaluation Abilities (EMI + EMS + EMU) and Reading Comprehension (L = Low; H = High)

	DMR				DMU			
	L		H		L		H	
EMI + EMS + EMU	L	H	L	H	L	H	L	H
Low	102	50	66	60	98	52	70	58
High	77	75	33	119	68	81	42	113

assigned to the cell frequencies of cell x_{ijk} (subject to the restriction $\sum\beta_{ijk} = 0$). If the contrasts are selected according to the specified hypothesis (i.e., 0 for the four low-divergent-production ability cells and 1, −1, −1, 1 for the high-production ability cells), μ_{ijk} can be estimated. The large-sample variance of the contrast was estimated by

$$\sigma^2_{\hat{\mu}} = \sum[(\beta_{ijk})^2(x_{ijk})^{-1}] \qquad (i, j, k = 1, 2) \qquad (12)$$

The standardized effect was then estimated by dividing the estimated effects by their standard deviations. Table 6 presents the results of this analysis for both divergent-production abilities.

The data in Table 6 imply that both effects were significant. Students with high divergent-production ability scores and high evaluation ability scores evaluated their answers on the reading comprehension task. Students who had high divergent-production abilities but low evaluation abilities read less well. This supports the importance of evaluation abilities for answering reading comprehension items.

Discussion

It was not possible to demonstrate a substantial difference in intellectual abilities measured with either open-ended items or multiple-choice items for reading comprehension. Therefore, open-ended and multiple-choice items for reading comprehension are evidently congeneric with respect to the SI abilities measured. Because a relatively large proportion of the variance in true reading comprehension scores was accounted for by the SI ability tests, it is tempting to conclude that if semantic abilities are involved differentially in answering reading comprehension items with a different format, they only play a minor role.

The lack of a substantial difference due to item format does not mean there are no differences at all between items in different formats. It only means that the differences in semantic abilities tested due to the format of the items seem small relative to the differences in abilities tested with (largely) comparable reading comprehension tests. Furthermore, the lack of a substantial difference due to item format only concerns semantic abilities, and no generalization to non-semantic abilities is warranted.

The results of this study do not suggest that the answering processes for different item formats are identical. Respondents may obtain the same total reading comprehension score using the same intellectual skills, although the processes involved differ, without one being more efficient than the other (Sternberg, 1980).

Table 6
Results of the Analysis of an Interaction Effect ($\hat{\mu}_{ijk}$) Between Divergent Production Ability, Evaluation Ability, and Reading Comprehension for the High-Divergent-Production Ability Group

Ability	$\hat{\mu}_{ijk}$	Var($\hat{\mu}_{ijk}$)	Standardized Effect
DMR	1.378	.071	5.171
DMU	1.178	.064	4.656

The results imply that answering multiple-choice questions is not solely a matter of comparing and deleting alternatives. Convergent- and divergent-production abilities appear to be important as well. Nor did the answering of open-ended items prove to be solely a question of production abilities, as indicated by the interaction between divergent-production abilities, evaluation abilities, and reading comprehension. Good readers seem to evaluate their answers on both divergent-production and reading comprehension items, whereas students with high divergent-production abilities and low evaluation abilities seem to go astray because of their lack of evaluation of their answer with reference to either the text or the item they were supposed to answer. This conclusion poses a problem for most research on the basis of the SI model, because it is usually assumed that intellectual tasks such as reading can be described as a linear combination of single SI abilities; interaction effects are usually neglected.

The results suggest, therefore, that students seem to construct their answers to multiple-choice items to the same degree as when they answer open-ended reading comprehension items. Of course, these results are not generalizable to a situation in which a compare-and-delete strategy for answering multiple-choice items is explicitly taught, as may be the case in the last weeks or months before important examinations (Gipps, Steadman, Blackstone, & Stiener, 1983; Goslin, 1967).

Memory abilities proved to be correlated with the answering of reading comprehension items. This is in concordance with models of text comprehension, in which the activation and revision of the text representation forms an essential part. The importance of memory for semantic systems, together with the importance of evaluation of semantic systems, both indicating the activation and revision of text representations, can be interpreted as support for these models.

The difference in abilities tested between the two reading comprehension tests observed in Table 4 has two possible explanations. The first concerns the different structures of the texts for which the items were constructed. In the second text (Test B) more transformations had to be recognized and produced, whereas the first text (Test A) was more explicit. On the other hand, the differences in abilities measured might also be due to small differences in items, although this did not appear in the classification of the items.

References

Ackerman, B. P. (1984a). The effects of storage and processing complexity on comprehension repair in children and adults. *Journal of Experimental Child Psychology, 37,* 303–343.

Ackerman, B. P. (1984b). Storage and processing constraints on integrating storage information in children and adults. *Journal of Experimental Child Psychology, 38,* 64–92.

Benson, J. (1981). A redefinition of content validity. *Educational and Psychological Measurement, 41,* 793–802.

Benson, J., & Crocker, L. (1979). The effects of item format and reading ability on objective test performance. *Educational and Psychological Measurement, 39,* 381–387.

Biemiller, A. (1977). A relationship between oral reading rates for letters, words and simple texts in the development of reading achievement. *Reading Research Quarterly, 12,* 259–266.

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats—it does make a difference for diagnostic purposes. *Applied Psychological Measurement, 11,* 385–395.

Bracht, G. H., & Hopkins, K. D. (1968). *Comparative validities of essay and objective tests* (Research Paper No. 20). Boulder CO: Laboratory of Educational Research, University of Colorado.

Brown, J. (1976). An analysis of recognition and recall and of problems in their comparison. In J. Brown (Ed.), *Recall and recognition.* London: Wiley.

Carter, H. D., & Crone, A. P. (1940). The reliability of new-type or objective tests in a normal classroom situation. *Journal of Applied Psychology, 24,* 353–368.

Cattell, R. B. (1963). *Abilities—their structure, growth and action.* Boston: Houghton Mifflin.

Cook, D. L. (1955). An investigation of three aspects of free-response and choice-type tests at college level. *Dissertation Abstracts,* 1351.

Coombs, C. H., Milholland, J. E., & Womer, F. B. (1956). The assessment of partial knowledge. *Educational and Psychological Measurement, 16,* 13–37.

Croft, A. C. (1982). Do spelling tests measure the ability to spell? *Educational and Psychological Measurement, 42,* 715–723.

Cronbach, L. J. (1966). New light in test strategy from decision strategy. In A. Anastasi (Ed.), *Testing prob-*

lems in perspective. Washington DC: American Council on Education.

Culpepper, M. M., & Ramsdale, R. (1983). A comparison of multiple-choice and essay tests of writing skill. *Research in the Teaching of English, 41,* 295–298.

Davis, F. B. (1968). Research in comprehension in reading. *Reading Research Quarterly, 3,* 499–535.

Doolittle, A. E., & Cleary, T. A. (1987). Gender-based differential item performance in mathematics achievement items. *Journal of Educational Measurement, 24,* 157–166.

Fienberg, S. E. (1980). *The analysis of cross-classified categorical data*. Cambridge: MIT Press.

Frary, R. B. (1985). Multiple-choice versus free-response: A simulation study. *Journal of Educational Measurement, 22,* 21–31.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 9,* 2–67.

Gipps, C., Steadman, S., Blackstone, T., & Stiener, B. (1983). *Testing children: Standardized testing in local educational authorities and schools*. London: Heinemann.

Goslin, D. A. (1967). *Teachers and testing*. New York: Russell Sage Foundation.

Guilford, J. P. (1971). *The nature of human intelligence*. London: McGraw-Hill.

Guilford, J. P. (1979). *Cognitive psychology with a frame of reference*. San Diego: EDITS.

Guilford, J. P., & Hoepfner, R. (1971). *The analysis of intelligence*. New York: McGraw-Hill.

Gulliksen, M. (1950). *Theory of mental tests*. New York: Wiley.

Guttman, L. (1965). A faceted definition of intelligence. In K. Eiferman (Ed.), *Studies in psychology, Scripta Hierosolymitana 14* (pp. 166–181). Jerusalem: Hebrew University.

Heim, A. W., & Watts, R. B. (1967). An experiment on multiple-choice versus open-ended answering in a vocabulary test. *British Journal of Educational Psychology, 37,* 339–346.

Hoeks, J. (1985). *Vaardigheden in begrijpend lezen* [Abilities in reading comprehension]. Unpublished doctoral dissertation, University of Amsterdam.

Hogan, T. P. (1981). *Relationships between free-response and choice type tests of achievement: A review of literature*. Washington DC: National Institute of Education.

Hopkins, K. D., George, C. A., & Williams, D. D. (1985). The concurrent validity of standardized achievement tests by content area using teachers' ratings as criteria. *Journal of Educational Measurement, 22,* 177–182.

Hopkins, K. D., & Stanley, J. C. (1981). *Educational and psychological measurement and evaluation* (6th ed.). Englewood Cliffs NJ: Prentice-Hall.

Hurd, A. W. (1930). Comparison of short-answers and multiple-choice type tests covering identical subject content. *Journal of Educational Research, 26,* 28–30.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36,* 109–132.

Jöreskog, K. G., & Sörbom, D. (1986). *LISREL: Analysis of linear structural relationships by maximum likelihood, instrumental variables and least squares methods*. Uppsala, Sweden: University of Uppsala.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87,* 329–354.

Kinney, C. B., & Eurich, A. C. (1938). A summary of investigations comparing different types of tests. *School and Society, 36,* 540–544.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of discourse comprehension of simple prose. *Psychological Review, 85,* 363–394.

Klein-Braley, C. (1983). A cloze is a cloze is a question. In J. W. Oller (Ed.), *Issues in language testing research*. Rowley MA: Newbury House.

Lorch, R. F., Lorch, E. P., & Matthews, P. D. (1985). On-line processing of the topic structure of a text. *Journal of Memory and Language, 24,* 350–362.

Lorch, R. F., Lorch, E. P., & Mogan, A. (1987). Task effects and individual differences in on-line processing of the topic structure of a text. *Discourse Processes, 10,* 63–80.

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

Mellenbergh, G. J. (1971). *Studies in studietoetsen* [Studies in educational testing]. Unpublished doctoral dissertation, University of Amsterdam.

Meuffels, B. (1982). *Studies over taalvaarrdigheid* [Studies in language abilities]. Unpublished doctoral dissertation, University of Amsterdam.

Mezynsky, K. (1983). Issues concerning the acquisition of knowledge: Effects of vocabulary training on reading comprehension. *Review of Educational Research, 53,* 253–279.

Murphy, R. J. (1982). Sex differences in objective test performance. *British Journal of Educational Psychology, 52,* 213–219.

Neville, M. H., & Pugh, A. K. (1977). Context in reading and listening: Variations in approach to cloze tasks. *Reading Research Quarterly, 11,* 12–31.

Paterson, D. G. (1926). Do old and new type examinations measure different mental functions? *School and Society, 24,* 246–248.

Samson, D. M. M. (1983). Rasch and reading. In J. van Weeren (Ed.), *Practice and problems in language testing 5*. Arnhem, The Netherlands: CITO.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing, 1,* 147–169.

Spache, G. D. (1963). *Toward better reading.* Champaign IL: Gerrard.

Spearman, C. (1927). *The abilities of man.* London: McMillan.

Sternberg, R. J. (1980). A proposed resolution of curious conflicts in the literature of linear syllogisms. In R. S. Nickerson (Ed.), *Attention and performance* (vol. VII). Hillsdale NJ: Erlbaum.

Sternberg, R. J., & Powell, J. S. (1983). Comprehending verbal comprehension. *American Psychologist, 38,* 878–893.

Thurstone, L. L., & Thurstone, T. G. (1941). Factorial studies of intelligence. *Psychometric Monographs,* No. 2.

Traub, R. E., & Fisher, C. W. (1977). On the equivalence of constructed-response and multiple-choice tests. *Applied Psychological Measurement, 1,* 355–369.

van den Bergh, H. (1987). Open en meerkeuzevragen tekstbegrip: Een onderzoek naar de relatie tussen prestaties op en meerkeuzevragen [Open-ended and multiple-choice items: A study of the relationships between achievements on open-ended and multiple-choice items for reading comprehension]. *Tijdschrift voor Onderwijsonderzoek, 12,* 304–312.

van den Bergh, H. (1989). De correlationele structuur van enkele schriftelijke taalvaardigheden [The correlational structure of several language abilities]. *Tijdschrift voor Taalbeheersing, 11,* 38–59.

van den Bergh, H., Eiting, M., & Otter, M. (1988). Differentiële effecten van vraagvorm bij aardrijkskunde en natuurkunde examens [Differential effects of item format in geography and science exams]. *Tijdschrift voor Onderwijsonderzoek, 13,* 270–284.

Vernon, P. E. (1962). The determinants of reading comprehension. *Educational and Psychological Measurement, 22,* 269–286.

Ward, W. C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Applied Psychological Measurement, 6,* 1–11.

Ward, W. C., Frederiksen, N., & Carlson, S. B. (1980). Construct validity of free-response and machine-scorable forms of a test. *Journal of Educational Measurement, 17,* 11–29.

Weiss, D., & Jackson, R. (1983). *The validity of descriptive tests of language skills: Relationships to direct measures of writing skill ability and grades in introductory college English courses* (Report LB-83-4; ETS-RR-83-27). New York: College Entrance Examination Board.

Acknowledgments

This research was funded by the Institute of Educational Research (SVO) in The Netherlands (Project 4075). The author wishes to thank Mindert Eiting, Jan Elshout, Gideon Mellenbergh, and Bert Schouten for their advice and comments.

Author's Address

Send requests for reprints or further information to Huub van den Bergh, University of Utrecht, Faculteit der Letteren, Trans 10, 3512 JK Utrecht, The Netherlands.