Research Report

Rasch Rating Scale Modeling of Data From the Standardized Letter of Recommendation

Sooyeon Kim

Patrick C. Kyllonen

Research & Development

October 2006
RR-06-33

Rasch Rating Scale Modeling of Data From the Standardized Letter of Recommendation

Sooyeon Kim and Patrick C. Kyllonen
ETS, Princeton, NJ

October 2006

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

The Standardized Letter of Recommendation (SLR), a 28-item form, was created by ETS to supplement the qualitative rating of graduate school applicants' nonacademic qualities with a quantitative approach. The purpose of this study was to evaluate the following psychometric properties of the SLR using the Rasch rating-scale model: dimensionality, reliability, item quality, and rating category effectiveness. Principal component and factor analyses were also conducted to examine the dimensionality of the SLR. Results revealed (a) two secondary factors underlay the data, along with a strong higher order factor, (b) item and person separation reliabilities were high, (c) noncognitive items tended to elicit higher endorsements than did cognitive items, and (d) a 5-point Likert scale functioned effectively. The psychometric properties of the SLR support the use of a composite score when reporting SLR scores and the utility of the SLR in higher education and in admissions.

Key words: Standardized Letter of Recommendation, rating-scale model, noncognitive constructs, higher education, factor analysis, dimensionality

Acknowledgments

We would like to thank Kevin Meara, Fredric Robin, Kevin Larkin, Anna Kubiak, and Daniel Eignor for many diverse and helpful comments on earlier drafts of this paper. The authors also gratefully acknowledge the editorial assistance of Kim Fryer.

Introduction

Many in the graduate education community believe that noncognitive variables (i.e., nonacademic, socioaffective, affective-motivational, and personality variables) should play a role in graduate admissions. Faculty members report that nonacademic personal qualities such as motivation, integrity, and the ability to work with others are important for success in higher education (Walpole, Burton, Kanyi, & Jackenthal, 2001; Willingham, 1985). Many conceptual and empirical investigations of noncognitive factors have been conducted (Briel et al., 2000; Hough, 2001; Kuncel, Hezlett, & Ones, 2001; Schmidt, 1994). Nevertheless, interest in the role of noncognitive factors in higher education has not led to large-scale development of new assessments that might be used for guidance or admission purposes (Kyllonen, Walters, & Kaufman, 2005). No standardized assessment is available to measure such qualities, primarily because many policymakers and scientists do not believe that qualities based on subjective human ratings can be measured validly or reliably.

Graduate admissions staff have often mentioned a need for noncognitive indicators to augment the cognitive measures on the Graduate Record Examinations® (GRE®; Briel et al., 2000). A wide range of noncognitive factors (e.g., attitudes, learning skills, motivation, teamwork, affective competencies) affects students' educational outcomes. Accordingly, noncognitive characteristics play multiple roles in admissions, attrition, grade point average, and time to degree completion (Kyllonen, 2005).

Standardized Letter of Recommendation

How do faculty members judge whether students possess the personal qualities that higher education demands without a standardized format? In practice, faculty members use letters of recommendation, along with students' personal statements (Walpole et al., 2001). Reliance on letters of recommendation, however, can be problematic. These letters often reflect their authors' writing skills as much as they reflect candidates' strengths. Letters often contain circumspect, nuanced language, or code words, which can be misunderstood. Letters also may not address a particular personal quality about which an admissions committee desires information, and this omission can be misinterpreted as intentional when it may be inadvertent.

Letters of recommendation are typically treated as qualitative data on which an interpretive analysis is performed. Faculty members believe that letters contain reliable and valid information on candidates (Walpole et al., 2001); however, it is difficult to evaluate the accuracy

of those beliefs with qualitative data and analyses. It is also difficult, if not impossible, to determine the psychometric qualities of conventional letters of recommendation. Large-scale studies on the actual validity of letters of recommendation are needed to make their use more objective and systematic. To achieve this goal, the Standardized Letter of Recommendation (SLR) has been proposed as a quantitative method to supplement the qualitative approach to rating applicants' nonacademic qualities (Walters, Kyllonen, & Plante, 2004).

The SLR, a 28-item form, supplements the qualitative rating of graduate school applicants' nonacademic qualities with a quantitative approach. The SLR is designed to capture in quantitative form essentially the same judgments about candidates as do conventional letters. The SLR is being used in selecting ETS summer interns and fellowship recipients and also is being pilot tested in various university settings (Kyllonen, 2005). The seven scales that comprise the SLR—*knowledge and skill*, *creativity*, *communication skills*, *teamwork*, *motivation*, *self-organization*, and *professionalism and maturity*—emerged from a comprehensive review of the literature (Briel et al., 2000; Walpole et al., 2001) and from numerous faculty interviews and focus groups (Walters et al., 2004). The 28 items, with four items for each of the seven scales, were developed from those sources and consultation with expert personality psychologists. This quantitative form allows ratings to be evaluated with respect to their psychometric properties, such as dimensionality (i.e., on how many underlying qualities do raters judge candidates) and reliability, in an objective manner. Research to develop the SLR was described by Walters et al. (2004) in detail.

Using factor analytic procedures, Kyllonen and Kim (2005) investigated the psychometric properties of the SLR and identified the structure of cognitive/noncognitive qualities based on data from 430 faculty members, each of whom rated a single graduate school applicant. This study showed a lack of ceiling effect in ratings and very high reliabilities for both the overall SLR and its subscales, supporting its psychometric properties. A strong common factor accounted for 80% of the common variance among the 28 items, and both the scree test and eigenvalue > 1 criteria suggested a two-factor solution. Although these findings provided an indication of the functional unidimensionality of the SLR scale, two-, three-, four-, and five-factor solutions were also computed to confirm the theoretical structure of the SLR (i.e., seven scales) using the same empirical data set. Based on various factor pattern matrices, the major conclusions of the previous study were: (a) faculty members not only rated students along one

general dimension, but also distinguished among various qualities, and (b) faculty members differentiated between cognitive and noncognitive qualities and they reliably differentiated teamwork, professionalism, motivation, and communication skills within the noncognitive realm. However, those conclusions can be considered to be somewhat premature due to some methodological limitations (e.g., use of factor analysis model only, use of a single data set). We revisit the dimensionality issue of the SLR scale in the present study, using a number of different approaches, the most important of which involved application of the Rasch model.

The Rasch Measurement Model

Compared to classical test theory (CTT), item response theory (IRT) based measurement models rely on different analytical approaches to evaluate an instrument's psychometric properties (Embretson & Reise, 2000). The simplest of the IRT models is the Rasch model. A series of studies have been conducted to investigate the differences between Rasch and exploratory factor analysis (Chang, 1996; Green, 1996; Smith, 1996; Wright, 1996). Some of the problems that factor analytic procedures pose under CTT can be averted with the use of the Rasch model under IRT as shown in those studies.

The Rasch model uses the raw score to estimate trait ability and places the estimated trait ability on the same metric (i.e., logit scale) with item difficulty estimates. The overlap between trait ability and item difficulty distributions on the logit scale can then be examined to determine whether the instrument is appropriate for the given sample. If the measurement instrument works properly (i.e., the IRT model fits the data), the estimation of item parameters does not depend on the specific sample used and unbiased estimates of item parameters may be obtained from unrepresentative samples (Embretson & Reise, 2000, p. 23). As with any IRT model, the Rasch model assumes responses on an ordinal level, thus avoiding the discussion of Likert scales as "quasimetric" scales or the usual problems associated with nonnormality of data. In addition, with small sample sizes the Rasch model provides more stable parameter estimates when compared to two- or three-parameter logistic (2PL or 3PL) IRT models. Primarily for this reason, the Rasch model was used in the present study to investigate the psychometric properties of the SLR scale that was rated using the conventional Likert scale.

The Rasch rating-scale model (RSM; Andrich, 1978) is appropriate for polytomous data from the Likert response format employed in this study. The RSM describes the probability that person *n* will be observed in a specific rating-scale category *x* on a particular item *i*. The equation

for this probability contains three parameters: raters' perception of their ratees (β_n), the item's endorsability (δ_i), and a set of threshold parameters (τ_k). For the RSM, the distance between each threshold parameter is assumed to be constant across all items:

$$P\left(X_{ni}=x\right)=\frac{\exp\sum_{k=0}^{x}\left[\beta_n-\left(\delta_i+\tau_k\right)\right]}{\sum_{x=0}^{m}\exp\sum_{k=0}^{x}\left[\beta_n-\left(\delta_i+\tau_k\right)\right]},\ x=0,1,...,m \qquad (1)$$

where $P\left(X_{ni}=x\right)$ is the probability that person *n* is observed in the rating-scale category *x* on item *i*, which has *m* + 1 rating-scale categories, and

$$\left[\beta_n-\left(\delta_i+\tau_0\right)\right]=0.$$

Parameters for this model are estimated using joint maximum likelihood estimation as implemented in WINSTEPS (Linacre & Wright, 2000). The respondent measure refers to the raters' tendency to endorse items as descriptions of their perceptions of their ratees. The item calibration refers to the difficulty of endorsing a particular item, and the threshold calibration refers to the difficulty of assigning a rating of *k* versus *k*–1 on the rating scale in question. The Rasch model and associated fit statistics can be used to identify items that define a single linear dimension, subject to the constraint that the model is appropriate for that set of items.

Purposes

The present study is an extension of our previous work (Kyllonen & Kim, 2005) for the SLR using the Rasch measurement model. More specifically, the purpose of the present study is to confirm the psychometric properties (e.g., reliability, dimensionality) of the SLR using the RSM. Within the Rasch model, we used category frequency, average measures, thresholds, item-fit indices, and separation and reliability indices to examine the psychometric properties of the SLR and its scales. In addition, a Rasch dimensionality analysis is conducted to determine (a) the degree to which the SLR exhibits sufficient internal consistency to support an assumption of unidimensionality and (b) whether relations among items within a scale are consistent with theory-based expectations.[1] Both conventional principal component analysis and factor analysis under CTT are also conducted to confirm the results from the Rasch dimensional analysis. In

sum, we are proposing to supplement the qualitative rating of graduate school applicants' nonacademic qualities with quantitative approaches based on factor analysis under CTT and on the Rasch measurement under IRT.

Methods

Participants

In the present study, we used the same data set as in the previous studies (Kyllonen & Kim, 2005; Walters et al., 2004). The participants are 430 faculty members (67% male and 33% female) from a variety of American university departments (e.g., biology, political science, and psychology) who receive students' GRE score reports. Faculty participants were screened for their willingness to participate and on their levels of involvement in graduate admissions (e.g., writing or reading letters of recommendation). They were each asked to rate a student for whom they most recently wrote a letter of recommendation by completing the standardized letter of recommendation (SLR). The faculty members were mostly academic advisors, committee members, department chairs, instructors, research supervisors, or mentors of the targeted students. About 87% of the faculty members reported that they had known the targeted students for 1 year or longer. In a subsequent telephone interview, they were asked for their opinions of the SLR; the results of those interviews are described elsewhere (Walters et al., 2004).

Measure: Standardized Letter of Recommendation

The SLR's scales and items were developed into a Web-based instrument that was refined through focus groups and usability studies (Walters et al., 2004). The SLR is composed of seven theory-based dimensions, each of which includes four items: (a) knowledge and skill, (b) creativity, (c) communication skills, (d) teamwork, (e) motivation, (f) self-organization, and (g) professionalism and maturity. A 5-point response set was used to rate the 28 items: 1 (below average), 2 (average), 3 (above average), 4 (outstanding), and 5 (truly exceptional). A sixth point (6) existed for do not know. "Do not know" responses were treated as missing data; accordingly, the numbers of faculty members from whom data were actually available varied across items from 384 to 428. Of the 430 raters, 38% (N = 165) had at least one item missing. Overall, 4.7% of the 12,040 possible item ratings (430 x 28) were missing from the data set.

Cronbach's alphas for the seven scales ranged from .84 to .89 and intercorrelations among them were high and fairly homogeneous (.54 to .81), indicating the existence of a general

factor across dimensions. No ceiling effects emerged, and the items' distributions did not depart from univariate normality. Detailed descriptive statistics for the SLR are presented elsewhere (Kyllonen & Kim, 2005).

Missing Data

Most statistical procedures exclude from analysis cases for which any values are missing; including only complete cases, however, can result in the loss of a significant amount of information. Accordingly, we used multiple imputation, which has been shown to yield accurate replacement values (Smits, Mellenbergh, & Vorst, 2002), to deal with missing values. According to Yuan's (2000) multiple imputation efficiency criterion, we generated in the previous study (Kyllonen & Kim, 2005) five complete data sets based on a Markov chain Monte Carlo (MCMC) method built into the SAS PROC MI procedure. We selected one of the five data sets at random for the present study, because results derived from these data sets in the previous study were very similar (e.g., factor loading estimates were identical to two decimal places for all five data sets).

Procedure

As mentioned previously, the first objective of the present study was to determine the psychometric qualities of the items, scales, and response format using the Rasch measurement model. To do this, we examined item fit and reliability to ensure the overall quality of the SLR and its scales, along with the adequacy of the response set's range. Although the Likert-format response set has been used very extensively, we confirmed the optimal range of the response set on the SLR.

In the next step, we conducted a principal component analysis of the residuals after fitting the RSM, as implemented in WINSTEPS. This analysis involves computing the residuals, that is, the observed responses minus their expected values under the Rasch model. These residuals are subjected to a principal component analysis. If unidimensionality holds, then all recovered components should be at the noise level. Through application of this procedure, the amount of variance that each extracted principal component accounts for can be examined to determine the SLR's dimensionality. The major justification for using an IRT model to analyze item-level response data from the SLR is the assumption that the instrument is unidimensional, at least operationally. Otherwise, there would be little reason to choose a model that makes unidimensionality a requirement for measurement.

In addition, a principal component analysis (PCA) under CTT, which assumes no a priori number of factors, was conducted using a Pearson product moment correlation matrix. The patterns of loadings and the item content breakdown may confirm theory-based expectations for interitem relationships or may identify unintended relationships due to weakness in the item content or response formats. Furthermore, exploratory factor analyses (EFA) were conducted in the present study as certain evidence indicated that the SLR was multidimensional. In this case, we analyzed the SLR data using *principal factor analysis* (not the maximum likelihood method used in the previous study; Kyllonen & Kim, 2005) as an estimation method to allow a consistent comparison with the Rasch PCAs. The findings based on different models or assumptions were compared. It is worth noting that although we used the same data set as in the previous study, all the analyses conducted in the present study are new.

Results

Item Statistics

Table 1 presents Rasch-based item statistics for each of the 28 items ($N = 427$; data from 3 respondents whose ratings reached either the maximum [140] or minimum [28] score were excluded from the item calibration procedure). The lower the item difficulty value, the higher the endorsement (i.e., they are likely easy items). The items rated highest (i.e., lowest difficulty) were *demonstrates honesty and sincerity* (–.95), *is dependable* (–.90), and *maintains high ethical standards* (–.78); those rated lowest (i.e., highest difficulty) were *is among the most creative person I know* (1.01), *produces novel ideas* (.78), *writes with precision and style* (.71), and *makes decisions easily* (.65). In general, noncognitive items tended to elicit higher endorsements than did cognitive items.

Item fit measure. In Rasch measurement, infit and outfit have been used as quantitative measures of the discrepancy between a statistical model and the observed data set (Gustafson, 1980) based on signal-to-noise ratio theory. An acceptable fit range of 0.6 to 1.4 is used for rating-scale data (Bond & Fox, 2001, p.179). All items except *is rarely hostile or distrustful* (1.53) were within the range of reasonable fit; *has appropriate skills to perform effectively* (.66) and *has an unusually high level of energy* (1.34) attained marginal fit. The infit values of 1.53 and 1.34 indicate that 53% and 34% more variation, respectively, emerged in the observed data than the Rasch model predicted; this occurred when responses display haphazard tendencies (Bond & Fox, 2001, p.177).

Table 1

Item Statistics From Rating-Scale Analysis (N = 427)

Form/Item	Difficulty	Error	Infit	Outfit	*r*
Knowledge & skill (α = .89)					
1. Has sufficient knowledge of the field	.01	.07	.69	.70	.75
2. Has appropriate skills to perform effectively	–.11	.07	.66	.67	.76
3. Has a broad perspective on the field	.57	.07	.90	.93	.70
4. Is among the brightest person I know	.45	.07	.92	.93	.72
Creativity (α = .88)					
5. Produces novel ideas	.78	.07	1.06	1.06	.68
6. Generates multiple solutions to problems	.63	.07	1.02	1.02	.71
7. Is intensely curious about the field	–.44	.08	1.11	1.10	.72
8. Is among the most creative persons I know	1.01	.07	.83	.84	.73
Communication skills (α = .86)					
9. Demonstrates clear and critical thinking	.04	.07	.79	.79	.77
10. Speaks in a clear, organized, and logical manner	.19	.07	1.06	1.05	.70
11. Writes with precision and style	.71	.07	1.23	1.23	.66
12. Listens well and responds appropriately	–.19	.07	.81	.79	.77
Teamwork (α = .84)					
13. Shares ideas easily	.08	.07	.87	.88	.73
14. Supports the efforts of others	–.09	.07	1.19	1.18	.67
15. Makes decisions easily	.65	.07	.88	.89	.73
16. Is rarely hostile or distrustful	–.72	.08	1.53	1.54	.61
Motivation (α = .84)					
17. Maintains high standards of performance	–.40	.07	.78	.78	.78
18. Can overcome challenges and setbacks	–.18	.07	.95	.95	.73
19. Sets realistic goals	.38	.07	.76	.76	.77
20. Has an unusually high levels of energy	–.19	.07	1.34	1.33	.65

(Table continues)

Table 1 (continued)

Form/Item	Difficulty	Error	Infit	Outfit	*r*
Self–organization (α = .85)					
21. Organizes work and time effectively	.03	.07	1.09	1.09	.72
22. Shares ideas and findings with others	.20	.07	1.06	1.06	.69
23. Makes good decisions	.19	.07	.80	.79	.76
24. Can work independently of others	–.76	.08	1.00	.98	.73
Professionalism & maturity (α = .88)					
25. Maintains high ethical standards	–.78	.08	1.19	1.16	.71
26. Demonstrates honesty and sincerity	–.95	.08	1.16	1.11	.69
27. Is dependable	–.90	.08	1.08	1.05	.72
28. Regulates own emotions appropriately	–.22	.07	1.20	1.18	.67

Note. r = correlation between item and total score.

Reliability. In the Rasch model, reliability is estimated for both persons and items. The person separation reliability (Wright & Masters, 1982), which estimates how well the instrument differentiates persons on the measured variable, was .96. The person separation index for estimating the spread of persons on the measured variable was 4.6, expressed in standard error units. This value indicates good separation among persons. Reliability and separation for items, estimated in the same manner as for persons, were .98 and 6.7, respectively, indicating excellent psychometric qualities for the SLR.

Rating-Scale Diagnostics

In practice, most Likert scales tend to be unequally spaced instead of conforming to the assumption of equal spacing between points in the response set. Rasch analysis lets the responses of the persons using the rating scale determine the spacing actually in effect for them during their ratings. Rasch measurement diagnostics were used to evaluate how well the five categories that make up the response set functioned to create an interpretable measure. For each category, we examined the shape of the distribution and the number of endorsements the response received. As presented in Table 2, the distribution of the observed frequencies was negatively skewed, with no more than 2% of the total endorsements falling in the first category (below average). Low-frequency categories can be problematic because they do not include enough observations

to allow an estimation of stable threshold values. However, because the average endorsements increase monotonically across the rating scale, collapsing categories was viewed as unnecessary.

Table 2

Summary of Rating-Scale Diagnostics

Category	Observed count (%)	Expected score measure	Average measure	Threshold	Infit MNSQ	Outfit MNSQ
1. Below average	213 (2%)	–4.51	–1.96	—	1.51	1.49
2. Average	1,613 (13%)	–2.20	–.70	–3.35	.96	.96
3. Above average	3,960 (33%)	.01	.51	–.98	.90	.90
4. Outstanding	4,451 (37%)	2.20	1.74	1.01	.89	.89
5. Truly exceptional	1,719 (14%)	4.49	3.03	3.32	1.10	1.09

Note. Category, observed count, and percentage indicate the numbers of raters who chose a particular response category, summed for each category across all 28 items. Average measure is the mean of measures in a category predicted by a model.

The other pertinent scale characteristics, threshold and infit statistics, support the same conclusion. Because the threshold distance that defined a distinct proposition on the variable was larger than 1.4 and less than 5 logits (Linacre, 1999), the SLR response set clearly distinguished between category options (see Figure 1). The raters needed 2.31 logits to go from below average at –4.51 to average at –2.20 and 2.29 logits to go from outstanding at 2.20 to truly exceptional at 4.49. For those raters, to move from Category 1 to 2 is as easy as to move from Category 4 to 5, indicating strong agreement. For this reason, the use of factor analysis would appear to be appropriate in this study as the Likert score could be defined as having approximately equally spaced intervals. In addition, none of infit (outfit) mean square measures was greater than 2 (Linacre, 1999), indicating that no noise was introduced into the measurement process. Based on all of this information, we concluded that the SLR's response set functioned well.

Dimensionality

The 28 rating-scale items in Table 1 were designed to define seven aspects of applicant characteristics that would be captured with a single measure of the SLR for each of the 430 raters. The measurement question is: Do the 430 raters use the 28 items in a way that allows a single measure to be constructed?

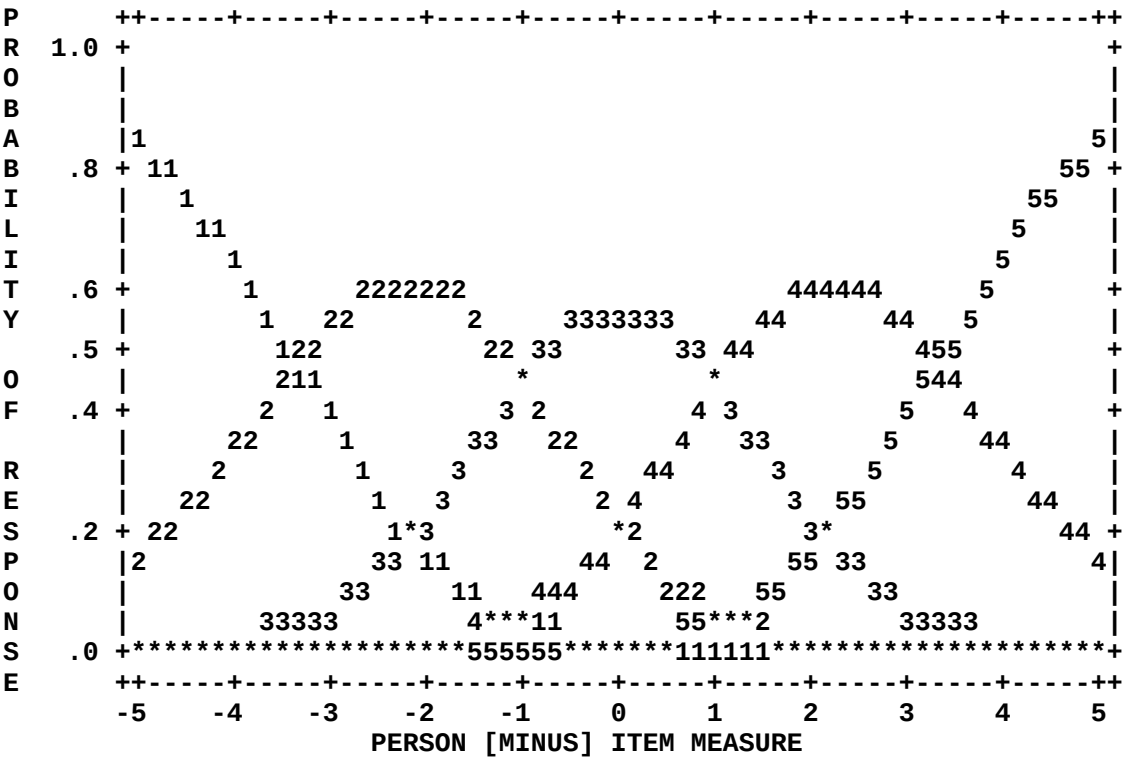

Figure 1. **Probability curves for a well-functioning five-category rating scale.**

Note. The probability curves show the probability of endorsing a given rating-scale category for every agreeability-endorsement difference estimate. The threshold estimates in Table 2 correspond to the intersection of rating-scale categories, indicating the point at which there is an equal probability of choosing either of two adjacent category options.

Rasch item fit. Item fit measures such as mean square (MNSQ) infit and outfit statistics, which were previously discussed, can also be used to determine how well each item contributes to defining one common construct as evidence of scale unidimensionality. As mentioned previously, a MNSQ infit or outfit value of 1 is ideal by Rasch specifications (Hong & Wong, 2005). Values greater than 1.4 indicate a lack of construct homogeneity with other items in a

scale (Doble & Fisher, 1998; Green, 1996), whereas values smaller than 0.6 may indicate item redundancy. As explained previously, 27 items (excluding Item 16) showed MNSQ infit and outfit statistics within the range of 0.6 to 1.4. In addition, all the items showed fairly high correlations with the total score (from .61 to .78), as presented in Table 1. Even Item 16 showed a moderate correlation with the total score (r = .61), indicating substantial item homogeneity. This may justify the use of the Rasch model in this study. In general, Rasch item-fit statistics supported the SLR scale's unidimensionality.

Local dependency. Local independence specifies that the response to one item has no influence on the response to another, after accounting for the underlying variable (e.g., Rasch ability dimension or the first principal component). Potentially locally dependent pairs of items have high positive or negative residual correlations after partialing out the Rasch dimension. As shown in Table 3, only 10 of the 378 possible pairs showed substantial standardized residual correlations due to item content similarity (e.g., *maintains high ethical standards* versus *demonstrates honesty and sincerity*). Because three items of the professionalism and maturity scale showed relatively large residual correlations among them (.33 to .63), the professionalism and maturity aspect might form a secondary (or independent) dimension of the SLR scale.

Residual principal component analysis. We performed a residual PCA under the Rasch model to determine the SLR's dimensionality using WINSTEPS. The residual PCA decomposes the matrix of item correlations based on standardized residuals to identify possible other dimensions that may affect response patterns. Residuals are the differences between what the Rasch model predicts and what is observed. If unidimensionality holds, then all recovered components should be at the noise level.

After partialing out the Rasch dimension, the first component from the matrix of residuals revealed a pattern among the 11 cognitive items that is in opposition to the pattern among the 12 noncognitive items. This implies that it is possible to obtain two measures from the SLR rather than a single composite measure. A plot of the first residual component clearly depicts the relationships among the SLR items (see Figure 2). In Figure 2, the X-axis ranges from items that are easy to difficult to endorse, whereas the first residual component loading (Y-axis) distinguishes cognitive from noncognitive items. As shown in Figure 2, a total of 11 items (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, and 11) could be represented by a cognitive component and another 12 items (13, 14, 16, 18, 19, 21, 22, 23, 25, 26, 27, and 28) could be represented by a noncognitive

component. Assignment of the 5 items (17, 24, 20, 12, and 15) that were generally easy to endorse (except Item 15) to a component was unclear. In general, it appeared that the SLR might be composed of multiple components.

Table 3

Largest Standardized Residual Correlations

No.	Items	No.	Items	*r*
25.	Maintains high ethical standards	26.	Demonstrates honesty and sincerity	.63
5.	Produces novel ideas	8.	Is among the most creative persons I know	.47
1.	Has sufficient knowledge of the field	2.	Has appropriate skills to perform effectively	.40
13.	Shares ideas easily	14.	Supports the efforts of others	.40
1.	Has sufficient knowledge of the field	3.	Has a broad perspective on the field	.37
26.	Demonstrates honesty and sincerity	27.	Is dependable	.35
25.	Maintains high ethical standards	27.	Is dependable	.33
10.	Speaks in a clear/organized/logical manner	11.	Writes with precision and style	.33
5.	Produces novel ideas	6.	Generates multiple solutions to problems	.32
6.	Generates multiple solutions to problems	26.	Demonstrates honesty and sincerity	–31

Note. *r* is a partial correlation with Rasch dimension removed.

Principal component analysis (PCA). The PCA under CTT was applied to test the unidimensionality assumption using the SAS statistical program. The number of largest eigenvalues (for eigenvalues greater than 1.4, see below) and the cumulative proportion of the total variance accounted for by the components (> 50%) were used to assess unidimensionality. The interpretation of the PCA results depends on the choice of the critical value for the eigenvalues. According to a simulation study (Smith, 1996), eigenvalues greater than 1.40 did not occur for the second factor in simulated unidimensional dichotomous and rating-scale data that were generated based on the RSM. The second eigenvalues fell mostly within the 1.20 to 1.30 range. Accordingly, we used the same criterion (eigenvalue in the 1.2 to 1.3 range) to

determine the presence of a second component on the SLR data, so as not to extract too many negligible/nuisance components.

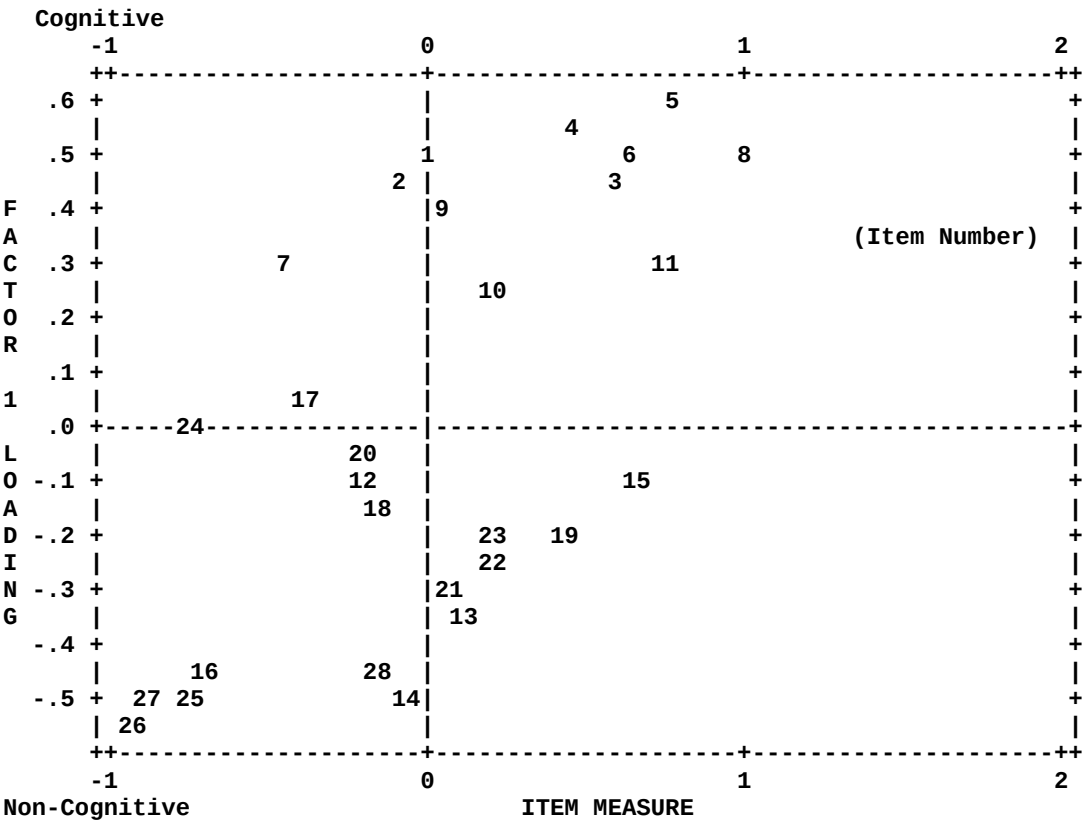

Figure 2. **Principal components (standardized residual) factor plot.**

The PCA results suggested two common components in the SLR data based on the greater-than-1.40 criterion. The first two components explained large amounts of variance for the SLR (e.g., eigenvalues of components 1 to 3 = 14.86, 2.01, and .99; the proportions of the total variance accounted by each component = 53%, 7%, and 4%). The first two components extracted over 60% of the total variance.

Factor analysis. We also conducted exploratory factor analyses (EFA) to determine the factor structure of the SLR. When using factor analytic techniques in this study, we initially examined a single factor model to see if all of the items loaded at least above .40 on the factor. As presented in the second column of Table 4, factor loadings of the 28 items ranged from .59 to .80, indicating unidimensionality in the SLR. As the magnitude of the first eigenvalue (14.495) indicates, a strong common factor captured substantial common variance among the 28 items

(82%). When we extracted two common factors based on eigenvalue-greater-than-1 criterion (2nd eigenvalue = 1.658), this factor model yielded results almost identical with those of the Rasch principal component analysis. The same five items (12, 15, 17, 20, and 24) demonstrated complex loadings for both the cognitive and noncognitive factors, as shown in the third and fourth column of Table 4. The factor loadings presented in those columns of Table 4 can be mapped with those in Figure 2. The memberships of the 28 items were the same as shown in the residual PCA plot under the Rasch model. The 28 items, however, work together well enough to define a single variable, as the eigenvalue and its proportion of variance accounted for indicate. In addition, the correlation between the two factors was moderate (r = .66).

Table 4

Factor Loadings from Exploratory Factor Analyses (N = 427)

Form/item	Single-factor model	Two-factor model	
		Cognitive	Noncognitive
Knowledge & skill			
1. Has sufficient knowledge of the field	.759	.765	.061
2. Has appropriate skills to perform effectively	.771	.736	.104
3. Has a broad perspective on the field	.701	.741	.022
4. Is among the brightest persons I know	.719	.811	–.029
Creativity			
5. Produces novel ideas	.679	.854	–.118
6. Generates multiple solutions to problems	.719	.770	.012
7. Is intensely curious about the field	.728	.621	.174
8. Is among the most creative persons I know	.728	.799	–.007
Communication skills			
9. Demonstrates clear and critical thinking	.784	.711	.144
10. Speaks in a clear, organized, and logical manner	.701	.584	.182
11. Writes with precision and style	.664	.635	.089
12. Listens well and responds appropriately	.777	.388	.465

(Table continues)

Table 4 (continued)

Form/item	Single factor model	Two factor model	
		Cognitive	Noncognitive
Teamwork			
13. Shares ideas easily	.727	.150	.653
14. Supports the efforts of others	.664	–.052	.788
15. Makes decisions easily	.729	.359	.441
16. Is rarely hostile or distrustful	.590	–.046	.699
Motivation			
17. Maintains high standards of performance	.797	.464	.410
18. Can overcome challenges and setbacks	.738	.327	.485
19. Sets realistic goals	.771	.312	.534
20. Has an unusually high levels of energy	.635	.323	.375
Self–organization			
21. Organizes work and time effectively	.733	.224	.584
22. Shares ideas and findings with others	.686	.181	.575
23. Makes good decisions	.767	.323	.520
24. Can work independently of others	.735	.437	.369
Professionalism & maturity			
25. Maintains high ethical standards	.714	–.019	.810
26. Demonstrates honesty and sincerity	.698	–.084	.858
27. Is dependable	.728	.033	.773
28. Regulates own emotions appropriately	.662	.011	.722

As mentioned previously, the SLR scale was designed to define seven aspects of applicant characteristics using the 28 items with the current score reporting procedure. Students get seven separate scores and a composite score, based on the theoretical framework of the SLR. The empirical evidence (e.g., two secondary factors and a higher order factor) could possibly be used to adapt the current scoring format in a practically meaningful way. For example, reporting two section scores (that reflect cognitive and noncognitive aspects, respectively) and a composite score could prove to be more useful in practice than reporting seven separate scores and the composite.

Discussion

The conversion of a traditionally qualitative process into a quantitative measure will in the future allow the validity of nonacademic factors to be evaluated in predicting success in higher education. This has both theoretical and applied implications. The theoretical implication concerns documenting the importance of nonacademic factors to key academic outcomes, such as attrition, grade point average, thesis quality, and type of job secured after graduation. The applied dimension involves the usefulness of an instrument such as the SLR as part of the admissions process in higher education.

Previous research (Kyllonen & Kim, 2005) examined the SLR's psychometric qualities, such as its reliability and dimensionality, using factor analytic techniques under a classical test theory model. This approach posed some limitations (e.g., sample dependency, interval scale requirement). The Rasch measurement model was used in this study to overcome these limitations and to confirm the psychometric quality of the SLR scale. Despite some differences between CTT and IRT, however, the two techniques can generally be expected to produce similar results when the items form a unidimensional scale and when sets of items within a larger item pool form strong, coherent cores.

The current results from the Rasch model agreed with ones previously obtained under CTT (Kyllonen & Kim, 2005). The item- and person-separation reliabilities were high, indicating excellent psychometric quality for the SLR. All items were highly correlated with the total score, suggesting substantial item homogeneity. The item-fit measures indicated that the 28 items represented well the specified contents of the SLR; thus, additional items may not be needed. As in the previous study, noncognitive items tended to elicit higher endorsements than did cognitive items. However, this phenomenon can be interpreted as a general tendency in human assessment rather than a bias or content issue.

The response format provides a number of possible answers to each question and it requires all respondents to use the same stimuli when formulating their responses. Despite scale developers' best intentions, several problems with this approach may occur. Respondents may not use the rating scale as it was intended, or they may interpret it according to their own understanding of the response labels. Lack of clear definitions of labels may lead to idiosyncratic responses. Providing many category choices can introduce more noise than information by requiring respondents to make their choices somewhat haphazardly. Because of that, we chose to

employ a 5-point Likert-type rating scale, and then we examined rating-scale diagnostics under IRT. In general, the 5-point Likert-type rating scale worked effectively for the SLR. Although the below average category had quite a low frequency of use, producing a negatively skewed distribution, the average endorsement increased monotonically across the rating scale. Accordingly, revision of the SLR's response set may not be necessary.

In the present study, principal component analysis of Rasch residuals indicated the possible existence of a second factor or a noncognitive factor to go along with a cognitive factor. EFA techniques based on PCA methods also indicated two factors on the SLR. These are consistent with item content and conceptual frameworks. The two methodologies also support the existence of a general common factor on the SLR scale. As mentioned previously, Rasch and factor analyses are quite comparable as long as a strong common factor exists despite multidimensional structural details.

It is important to distinguish between theoretical unidimensionality and practical, or functional, unidimensionality. From a theoretical perspective, it is possible to argue that the SLR is composed of seven dimensions. In the previous study (Kyllonen & Kim, 2005), we concluded that faculty members differentiated between cognitive and noncognitive qualities, and that within the noncognitive realm they reliably differentiated teamwork, professionalism, motivation, and communication skills. These conclusions were mainly derived from the perceived need to reduce the differences between the theoretical (seven factors) and functional dimensionality (e.g., two factors) of the SLR. For most purposes, however, we can view the two closely related factors as measuring a single theme, such as examinees' overall personal qualities. Hence, a single score that combines those seven aspects is reported. This concept of practical or functional unidimensionality is often used in constructing achievement tests.

Some researchers prefer Rasch methodology to classical techniques when assessing the psychometric qualities of their instruments/scales, emphasizing its statistical superiority (e.g., sample independence, creation of a scale with interval properties). It appears that the Rasch model offers a promising approach for solving a variety of problems encountered in personality assessment. Accordingly, the findings from the current study can be used more extensively to revise or elaborate the SLR's format if required. We believe we can capture through quantitative ratings essentially the same judgments about candidates that conventional letters address, but in a form for which dimensionality and reliability can be evaluated.

In general, we conclude based on this study that the SLR is well-constructed psychometrically, showing high reliability and separation indices and adequate item fit statistics. However, limitations of the study and some caveats must be noted. First, ratees' and raters' personality characteristics are confounded in the SLR rating situation. Second, it is not possible to examine interrater agreement because only one rater was available for each student. These are the fundamental limitations of the SLR data used in the present study that may prevent the current findings from being generalized. Future studies should be conducted to resolve these limitations using a more adequate data set (e.g., at least two raters for each student). Accordingly, in future efforts we will investigate interrater agreement (i.e., how closely two raters agree), and validity (i.e., how accurately do ratings predict outcomes such as attrition and grade point average). In addition, based on the results of the two studies (this study and Kyllonen & Kim, 2005), the following questions can be raised: What is the minimum number of items needed to achieve reasonably reliable results? Can fewer than 28 items be used while still maintaining an acceptable level of reliability and the same content coverage as the 28-item SLR? The shorter version of the SLR would require a shorter time to complete and might prove useful in practice, particularly for college professors with tight time schedules.

References

Andrich, D. (1978). Rating formulation for ordered response categories. *Psychometrika, 43*, 561–573.

Bond, T. G., & Fox, C. M. (2001). Applying the Rasch model: Fundamental measurement in the human sciences. Mahwah, NJ: Erlbaum.

Briel, J., Bejar, I., Chandler, M., Powell, G., Manning, K., Robinson, D., et al. (2000). *GRE Horizons Planning Initiative*. Unpublished manuscript, ETS, Princeton, New Jersey.

Chang, C.-H. (1996). Finding two dimensions in MMPI-2 depression. *Structural Equation Modeling, 3*, 41–49.

Doble, S. E., & Fisher, A. G. (1998). The dimensionality and validity of the older Americans resources and services (OARS) activities of daily living (ADL) scale. *Journal of Outcome Measurement, 2*, 4–24.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Green, K. E. (1996). Dimensionality analyses of complex data. *Structural Equation Modeling, 3*, 50–61.

Gustafson, J. E. (1980). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology, 33,* 220.

Hong, S., & Wong, E. (2005). Rasch rating-scale modeling of the Korean version of the Beck depression inventory. *Educational and Psychological Measurement, 65,* 124–139.

Hough, L. M. (2001). I/Owes its advances to personality. In B. W. Roberts & R. Hogan (Eds.), *Personality psychology in the workplace. Decade of behavior* (pp. 19–44). Washington, DC: American Psychological Association.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the graduate record examination: Implications for graduate student selection and performance. *Psychological Bulletin, 27*, 162–181.

Kyllonen, P. C. (2005). *The case for noncognitive assessments.* Princeton, NJ: ETS.

Kyllonen, P. C., & Kim, S. (2005, April). *Personal qualities in higher education: dimensionality of faculty ratings of students applying to graduate school.* Paper presented at the annual meeting of the American Educational Research Association, Montréal, Canada.

Kyllonen, P. C., Walters, A. M., & Kaufman, J. C. (2005). Noncognitive constructs and their assessment in graduate education: A review. *Educational Assessment, 10,* 153–184.

Linacre, J. M. (1999). Investigating rating-scale category utility. *Journal of Outcome Measurement, 3*(2), 103–122.

Linacre, J. M., & Wright, B. D. (2000). WINSTEPS: Multiple-choice, rating scale, and partial credit Rasch analysis [Computer software]. Chicago: MESA Press.

Schmidt, F. L. (1994). The future of personnel selection in the U.S. Army. In M. G. Rumsey & C. B. Walker (Eds.), *Personnel selection and classification* (pp. 333–350). Hillsdale, NJ: Erlbaum.

Smith, R. M. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling, 3*, 25–40.

Smits, N., Mellenbergh, G. J., & Vorst, H. C. M. (2002). Alternative missing data techniques to grade point average: Imputing unavailable grades. *Journal of Educational Measurement, 39*, 187–206.

Walpole, M. B., Burton, N. W., Kanyi, K., & Jackenthal, A. (2001). *Selecting successful graduate students: In-depth interviews with GRE users.* Princeton, NJ: ETS.

Walters, A. M., Kyllonen, P. C., & Plante, J. W. (2004). *Preliminary research to develop a standardized letter of recommendation.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Willingham, W. W. (1985). *Success in college*. New York: College Entrance Examination Board.

Wright, B. D. (1996). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling, 3*, 3–24.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Yuan, Y. C. (2000, April). *Multiple imputation for missing data: Concepts and new development.* Paper presented at the 25th annual SAS User Group International Conference. Indianapolis, IN.

Notes

[1] This analysis involves computing the Rasch residuals, that is, the observed responses minus their expected values under the Rasch model. These residuals are subjected to a principal component analysis. If unidimensionality holds, then all recovered components should be at the noise level.