Research Report

The Standardized Letter of Recommendation: Implications for Selection

Ou Lydia Liu

Jennifer Minsky

Guangming Ling

Patrick Kyllonen

Research & Development

August 2007
RR-07-38

The Standardized Letter of Recommendation: Implications for Selection

Ou Lydia Liu, Jennifer Minsky, Guangming Ling, and Patrick Kyllonen
ETS, Princeton, NJ

August 2007

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

In an effort to standardize academic application procedures, the Standardized Letter of Recommendation (SLR) was developed to capture important cognitive and noncognitive qualities of graduate school candidates. The SLR consists of seven scales (*knowledge, analytical skills, communication skills, motivation, self- organization, professionalism and maturity,* and *teamwork*) and was applied to an intern-selection scenario. Both professor ratings (N = 414) during the application process and mentor ratings of the selected students (N = 51) after the internship was completed were collected using the SLR. A multidimensional Rasch investigation suggests that the seven scales of the SLR displayed satisfactory psychometric properties in terms of reliability, model fit, item fit statistics, and discrimination. The two cognitive scales, *knowledge* and *analytical skills*, were found to be the best predictors for intern selection. The professor ratings and mentor ratings had moderate to high correlations, with the professor ratings being systematically higher than the mentor ratings. Possible reasons for the rating discrepancies are discussed. Also, implications for how the SLR can be used and improved in other selection situations are suggested.

Key words: Rasch multidimensional model, standardized letter of recommendation

Acknowledgments

We would like to give special thanks to Brent Bridgeman, Matthias von Davier, and Dan Eignor at ETS for their helpful comments, suggestions, and edits on an earlier draft of this report. The authors also gratefully thank Kim Fryer for her editorial assistance.

Table of Contents

List of Tables

List of Figures

Page

Introduction

Background

Letters of recommendation are an indispensable component of many selection endeavors. They are required for almost all undergraduate and graduate school applications, academic award and fellowship applications, and job applications. The purpose of these letters is to provide information about an applicant that helps decision makers ascertain whether the candidate's background fits an available position and whether the candidate is likely to be successful in fulfilling the expectations of the role. Letters of recommendation allow consideration of personal, relevant information that may otherwise be difficult to acquire (McCarthy & Goffin, 2001). Letters sometimes consist of a few brief, open-ended and/or Likert-type items that ask about the competence of an applicant. Items may or may not guide the recommender to write about specific areas such as academic performance or degree-work potential. Recommenders may be asked to place applicants at a level (e.g., top 5%, bottom half) compared to their peers.

Within the college and university sector, letters, as well as academic transcripts, standardized test scores, and personal statements, are the main selection criteria on which admissions decisions are based; thus letters have a significant impact on the selection process (Range, Menyhert, Walsh, Hardin, Ellis et al., 1991; Walpole, Burton, Kanyi, & Jackenthal, 2002). For example, in their examination of 416 American Psychological Association-accredited internship programs, Lopez, Oehlert, and Moberly (1996) found that letters of recommendation, along with clinical experience and the interview process, were rated as the most important selection criterion. As critical as these letters are to the selection process, however, several points of contention raised by research call into question their reliability.

Issues With Reliability and Utility

Typically letters of recommendation are not standardized, which consequentially poses threats to their reliability and utility (Range et al., 1991). Issues such as leniency (e.g., applicants choose people who are likely to write complimentary letters), the recommender's inadequate knowledge of an applicant's comprehensive behaviors or skill sets, and the fact that no training or standardized guidelines exist for writing letters of recommendation confound the legitimacy and usefulness of unstandardized letters (Aamont, Bryan, & Whitcomb, 1993; Range et al., 1991). In addition, with the rapid increase of international applicants, misinterpretations of

comments or descriptions in traditional reference letters are likely to occur due to differing social and cultural connotations or imperfect mastery of English.

For applicants, the chances of being viewed favorably depend partially on the focus, style, and specificity of recommendation letters rather than solely their own competence. A student applying to a particular graduate program may have to provide three letters of recommendation. Each letter is likely to be distinct in terms of the quality and quantity of information provided its writer. Writers may address different applicant characteristics or may differ with respect to the level of specificity they provide, making it difficult to evaluate applicants on a common ground. Indeed, research has indicated that letters containing specific examples of applicants' qualities are rated more highly than those that contain only generalizations, and letters written by one person for different applicants show more agreement than those written by different people for one applicant (Baxter, Brock, Hill, & Rozelle, 1981; Knouse, 1983). Thus, a professor who provides a general endorsement without appropriate documentation to support an observation may inadvertently jeopardize an applicant's chance of admission, compared with a professor who writes a much more convincing letter. In addition, whether applicants can have access to the recommendation is likely to influence how the recommenders write the letters. Ceci and Peters (1984, p. 31) found that students who did not waive their right to view a recommendation were rated higher than those who did waive their right, thus leading the authors to propose "…the need for departmental admissions committees to adopt a uniform policy governing letters of recommendation."

Under most circumstances, a letter of recommendation is treated as qualitative information that receives a subjective interpretation from a reviewer. Misunderstandings and misinterpretations may arise from the language nuances and subtleties in the letters, especially in cross-cultural situations (Kim & Kyllonen, 2006; Kyllonen & Kim, 2005). Additionally, because the majority of letters are written so positively, discriminating among them becomes burdensome and decision makers tend to search them for any suggestions of negative evidence (Miller & Van Rybroek, 1988; Range et al., 1991). Aamont, Nagy, and Thompson (1998) found that decision makers seem to be more approving of letters that are longer than of those that are concise. Thus, it is apparent that unstandardized letters grant unwarranted advantages to applicants with fluent, coherent, longer, and/or convincing reference letters. To make letters of recommendation more

useful and reliable as selection tools, there is a pressing need to standardize them so that they can provide consistent and systematic information.

SLR Development

As increasing attention has been devoted to the development of new types of recommendation letters, the Center for New Constructs at ETS initiated the SLR by quantifying important traits and characteristics related to successful graduate school outcomes. Letters of recommendation can vary in content, but in academia, applicants' knowledge, skills, and personality traits are usually appraised (McCarthy & Goffin, 2001). Research has found that the assessment of noncognitive qualities as part of the graduate school selection process is important to faculty, who believe that factors such as motivation, interests, and communication should be considered along with cognitive factors. These noncognitive skills are thought of as essential for success in graduate programs (Enright & Gitomer, 1989; Kyllonen, Walters, & Kaufman, 2005). Moreover, the inclusion of factors like these as graduate school selection criteria may help predict student retention (Kuncel, Hezlett, & Ones, 2001). Thus, measuring candidates' noncognitive qualities that are important to academic success could be extremely beneficial in predicting student retention; for example, students who are very motivated and are good communicators might be more likely to persist and graduate.

The SLR, consisting of 28 items (4 items per scale) covering multiple aspects of an applicant's cognitive and noncognitive characteristics, was therefore designed to provide the same kinds of judgments as conventional letters of recommendation, but using a quantitative approach. The seven scales that constitute the SLR—*knowledge and skill, creativity, communication skills, teamwork, motivation, self-organization,* and *professionalism and maturity*--were developed based on a comprehensive review of the existing literature and extensive interviews with faculty members and focus groups (Kyllonen et al., 2005; Walpole et al., 2002). The individual items on each scale were developed from the same sources. The personality items were reviewed in terms of their technical qualities (International Personality Item Pool, 2001). Expert opinion was collected on the overall coverage and clarity of the items.

Compared to other available standardized reference letters in the market (e.g., the Common Application developed by the Common Application membership organization), the SLR developed by ETS has three advantages. First, for most standardi ed letters that incorporate noncognitive factors, the items do not go beyond a couple of words such as *motivation* or

intellectual promise; thus the recommenders are asked to record their judgment without any supporting context. The item statements can be too vague to guide the ratings, and the reliability of the judgments may therefore be questionable. In contrast, the SLR provides a unique framework that allows those assessing applicants to respond to specific and more contextualized items. Studies note that the predictive validity of reference letters increases with the level of specificity and accuracy within the letter (Aamont et al., 1993). The SLR maintains this view for each scale; a distinct aspect of the construct is represented in each of the four items presented. Furthermore, each item is clearly stated to solicit an evaluation (e.g., has knowledge of behavioral science research methods). Second, the SLR was developed empirically with rigorous, methodological research to support its development. Insufficient information has been found about the development methodology of other letters of recommendation. Evidence of their psychometric properties should be provided to support their value (Muchinsky, 1979). Third, the amount of response time and accessibility to feedback were also considered when developing the SLR at ETS. Most traditional recommendation letters need to be downloaded, filled out manually, and then mailed to each college or university students are applying to. The entire SLR process, by contrast, can be completed online and is therefore expected to be more time-efficient.

Previous Studies Using the SLR

An earlier version of the SLR was investigated within a graduate school application scenario. Most items were similar to the version utilized in the current study and contained seven scales: *knowledge and skill, creativity, communication skills, teamwork, motivation, self-organization,* and *professionalism & maturity*. Faculty members (N = 430) were asked to complete the SLR for students for whom they had recently written a traditional letter of recommendation. The data under classical test theory (CTT) and Rasch modeling revealed some key findings: (a) The SLR is able to differentiate applicants from various academic backgrounds; (b) noncognitive variables tend to elicit higher endorsement than do cognitive variables; and (c) a five-point Likert scale functioned effectively for the SLR (Kim & Kyllonen, 2006; Kyllonen & Kim, 2005).

Despite the useful information provided by previous investigations, a few concerns remain. First, although the Kim and Kyllonen (2006) and the Kyllonen and Kim (2005) studies analyzed the same data, they reached inconsistent conclusions about the structure of the SLR items. The Kim and Kyllonen study concluded that the 28 items work together well to define a

single dimension based on a Rasch model analysis, exploratory factor analysis (EFA), and principal component analysis (PCA). However, the authors mentioned that the PCA and EFA suggest two common components or two factors in the SLR data with a correlation of .66. Their argument for a unidimensional model was guided mainly by a practical consideration that a composite score might prove to be more useful than seven separate scores. The Kyllonen and Kim study used EFA and confirmatory factor analysis (CFA) and compared the fit among four models (one-, two-, five-, and seven-factor). They found that both five- and seven-factor models fit better than the one- and two-factor models and that the five- and seven-factor models were similar in terms of fit. The authors concluded that there was at least some evidence that the recommenders used all five or seven factors in their ratings. No data have been available to examine the predicting power of the SLR, therefore its practical utility remains unknown. However, each scale's predicting power may provide additional information about the dimensionality of the SLR data. If each scale functions very differently in predicting outcomes, then the scales may measure different traits. Finally, more efforts are called for to ensure the generalizability of the SLR; it is important to investigate how the SLR functions in contexts other than graduate school admissions.

Objectives of This Study

This study continued the effort to explore the utility of the SLR by utilizing it for a summer intern program at ETS. In addition to SLR ratings from the candidates' professors, ratings from ETS mentors using the same SLR were collected after internships were completed. The mentor ratings were then used as a criterion for the intern application data. The previous version of the SLR was adapted with a few modifications. The *creativity* scale was replaced by a scale called *analytical skills*, which assesses a candidate's quantitative and qualitative analysis skills. This is because proficiency in quantitative or qualitative analyses is a critical factor for the candidates to successfully complete their internship projects. In this sense, the *creativity* scale was less relevant for the purpose of this program than the *analytical skills* scale. In addition, the wording of a few other items was changed to increase their clarity, as informed by previous analysis (Kyllonen & Kim, 2005; Kim & Kyllonen, 2006).

The primary purpose of this study was to address the following three questions: (a) How do the SLR items and scales function under the seven-factor design; (b) how do the SLR ratings affect ETS's intern selection; and (c) how do the SLR ratings predict intern performance at ETS?

The seven scales emerged from an extensive literature review, reflecting both theoretical values and empirical importance. The previous studies provided some clue that multiple dimensions underlie the SLR data. Hence, this study took a confirmatory approach to examine the psychometric properties of the SLR items under the seven-factor assumption.

Although the items used in this study largely followed the patterns of those used in the Kyllonen and Kim (2005) study and the Kim and Kyllonen (2006) studies, two distinctions may limit the comparability of the results from this study and the previous ones. First, the items and scales were modified. As stated earlier, the *creativity* scale was completely replaced by the *analytical skill* scale. In addition, a few items were modified in terms of wording. Second, the SLR was implemented under different circumstances. In the two previous studies using the same data, faculty members from a variety of universities in the United States were asked to rate a student for whom they had recently written a letter of recommendation, using the SLR. In this study, professors, most in measurement programs, used the SLR to evaluate students who applied to the ETS internship program. Thus, due to the difference in the impact the ratings have for the students, the recommenders may differ greatly in their motivation to evaluate the students. Furthermore, the intern applicants are normally competitive graduate students who are highly motivated to succeed academically, while the background of the students who were rated in the previous studies is unknown. These differences may affect the pattern and range of ratings, thus leading to different SLR structures.

In this study, a multidimensional model was selected to analyze the data collected during the intern application process. Under the item response theory (IRT) framework, if the observed responses fit the measurement model, the item parameter estimation can be sample-independent and can produce unbiased estimates (Embretson & Reise, 2000). The multidimensional model used in this study is described in detail in the Methods section.

Methods

Instrument

The SLR used to select ETS interns comprises seven scales with four items on each scale. The seven scales are: *knowledge, analytical skills, communication skills, motivation, self-organization, professionalism & maturity,* and *teamwork*. The SLR was designed to capture important cognitive and noncognitive characteristics deemed essential for satisfactory performance in a variety of professional settings (Walpole et al., 2002). It was developed as a

Web-based instrument that was refined through focus groups and usability studies (Walters, Kyllonen, & Plante, 2003). Data are automatically stored and can be retrieved as soon as a recommendation is completed. Recommenders are asked to specify the institution they are affiliated with, their title, and how long and in what capacity they have known the applicant. When responding to the SLR items, they are instructed to indicate how well the applicant they are rating compares to other students at their institution. All the items were rated using the following five-point Likert-type response scale: 1 (below average), 2 (average), 3 (above average), 4 (outstanding), and 5 (truly exceptional). An additional sixth category represented the "Do not know" situations, which were treated as missing data in the analyses. Therefore, the raw scores for each scale range from 0 to 20, with 140 being the maximum total score for all 28 items. After the items are completed, recommenders are given the opportunity to provide a narrative in which they can elaborate on each scale, offering examples and comments. Overall, 253 out of 11,592 (414 × 28) possible ratings were missing, a rate of 2.2%. Table 1 presents the 28-item SLR and descriptive statistics. An example of the actual assessment is provided in the appendix.

Participants

Every summer ETS offers students enrolled in doctoral programs research opportunities in different fields such as psychometrics, linguistics, and educational technology. Selected students spend two months at ETS mentored by senior staff as they work on projects and attend seminars and workshops. As part of the application process, candidates from 2004 to 2006 (N = 414) were required to ask two people who were familiar with their work to complete the SLR . All applicants had at least one faculty member from their home institution complete the SLR; 363 applicants received two SLR ratings. During these years, 51 applicants were selected as summer interns. After the program was completed in 2006, the ETS mentors who participated from 2004 to 2006 were asked to nominate an additional staff member who was familiar with each past intern's work. Both mentors and nominated staff were then asked to complete the SLR. As with the faculty ratings, all interns were rated by at least one ETS mentor, and 20 of the 51 interns received two SLR ratings. Since the number of completed SLRs varied for each applicant and selected intern, to be consistent, maximum ratings were used in the analyses.[1]

Table 1

Descriptive Statistics for the Professor Ratings of the SLR

Scale/Item	*N*	Mean	SD	Scale mean	Scale SD
Knowledge					
1. Has foundational knowledge in some branch of behavioral science (e.g., educational theory, psychology, economics).	403	3.70	.68	3.63	.59
2. Has knowledge of behavioral science research methods.	406	3.79	.66		
3. Has knowledge of educational or psychological assessment.	394	3.77	.74		
4. Has knowledge of program evaluation.	354	3.25	.90		
Analytical skills					
5. Has quantitative research methods and data analysis skills.	410	3.97	.75	3.75	.63
6. Has qualitative research methods and data analysis skills.	360	3.27	1.00		
7. Has demonstrated skill in measurement and assessment techniques.	388	3.80	.74		
8. Is skilled in statistical methods.	402	3.88	.79		
Communication skills					
9. Demonstrates clear and critical thinking.	414	4.05	.63	3.83	.64
10. Speaks in a clear, organized, and logical manner.	414	3.74	.80		
11. Writes with precision and style.	410	3.51	.79		
12. Listens well and responds appropriately.	414	4.02	.69		
Motivation					
13. Maintains high standards of performance.	414	4.22	.61	4.25	.52
14. Can overcome challenges and setbacks.	408	4.15	.61		
15. Seek out opportunities to learn.	414	4.38	.59		
16. Has a high level of energy.	413	4.25	.60		
Self-organization					
17. Organizes work and time effectively.	411	4.06	.64	4.02	.54
18. Set realistic goals.	408	3.85	.63		
19. Makes good decisions.	407	3.95	.62		
20. Can work independently of others .	412	4.21	.60		
Professionalism & maturity					
21. Maintains high ethical standards.	408	4.30	.58	4.28	.52
22. Demonstrates honesty and sincerity.	412	4.34	.57		
23. Is dependable.	414	4.35	.59		
24. Regulates own emotions appropriately.	407	4.17	.61		
Teamwork					
25. Shares ideas easily.	412	4.02	.63	4.15	.55
26. Supports the efforts of others.	407	4.12	.63		
27. Works well in group settings.	408	4.15	.65		
28. Behaves in an open and friendly manner.	414	4.32	.61		

The MRCML Model

The multidimensional random coefficient multinomial logit (MRCML) model (Adams, Wilson, & Wang, 1997) was used to perform the item calibration and person estimation. This model is a multidimensional extension of the Rasch family of item response models, and its applications have been used extensively in many settings (Briggs & Wilson, 2003; Wang, 1999; Wang, Wilson, & Adams, 1997). The software program *ConQuest* (Wu, Adams, & Wilson, 1997) was used to perform the analyses.

The MRCML model can be formally expressed as

$$P(X_{nik}=1|\boldsymbol{\theta}_n,\boldsymbol{\xi})=\frac{\exp[\mathbf{b}'_{ik}\boldsymbol{\theta}_n+\mathbf{a}'_{ik}\boldsymbol{\xi}]}{\sum_{k=1}^{K_i}\exp[\mathbf{b}'_{ik}\boldsymbol{\theta}_n+\mathbf{a}'_{ik}\boldsymbol{\xi}]}\ ,(1) \qquad (1)$$

where X_{nik} is a random variable and assumes a value of 1 when person *n*'s response to item *i* falls into category *k,* or zero otherwise ($1\leq i\leq \mathrm{I}$, $1\leq k\leq \mathrm{K}_i$, $1\leq n\leq N$). X_{ni1} is constrained to zero as a reference category for model identification purposes; $\boldsymbol{\theta}_n$ is a *D* x 1 parameter vector of a person *n* ($1 \leq d$ [dimension] $\leq \mathrm{D}$), indicating the multidimensional trait level. Note that the dimensions in the MRCML model are allowed to be non-orthogonal. In the case of the SLR, there are seven dimensions based on the seven scales. Therefore, each candidate will have an estimate on each of the seven dimensions (i.e., *knowledge*, *teamwork*, etc.). $\mathbf{b}'_{ik}$ is a 1 x *D* scoring vector for category *k* of item *i* that specifies the performance level associated with each score category (e.g., for a multiple-choice item of four categories, a score of 1 is assigned to the key category, and a score of zero is assigned to the rest of the categories). $\boldsymbol{\xi}$ is a *p* x 1 item parameter vector; and $\mathbf{a}'_{ik}$ is a 1 x *p* vector, specifying the hypothesized mapping of items to dimension (Briggs & Wilson, 2003). In this case, the between-item multidimensionality was assumed, meaning that each item falls into one dimension only. The design matrix consists of 28 item parameters and 28 × 3 step parameters. There are also seven person parameters, one for each dimension.

The marginal maximum likelihood method is used to estimate the item parameters and the population means and variances. The expectation-maximization (EM) algorithm is used to obtain the maximum-likelihood estimates and asymptotic standard errors (Bock & Aitken, 1981). For model identification purposes, the population mean of applicant trait estimates ($\boldsymbol{\theta}_n$) is

constrained to zero. Monte Carlo approximations are used in *ConQuest* when the models have high dimensionality. A number of nodes (e.g., $N = Q$) are randomly drawn from the standard multivariate normal distribution. For each iteration in the estimation process, the Q nodes are rotated so that they become random draws from a multivariate normal distribution with a mean and variance. The weight for all nodes is set to be $1/Q$ in *ConQuest*.

By using this model, the ratings on each item are treated as distinct information about each applicant. The loss in reliability when selecting a multidimensional model instead of a unidimensional one is overcome by incorporating the correlation between the latent dimensions.

Procedure

To investigate the dimensionality of the 28-item SLR, a seven-dimensional model was fitted on the basis of the seven scales using the MRCML model. Then the psychometric properties of the seven dimensions were examined with regard to scale reliability, item fit, item discrimination, and Wright map. The correlation between dimensions was also provided.

In the next step, the predicting power of the seven dimensions for intern selection was analyzed. The estimates produced in the multidimensional analyses above were used for the regression. For Rasch-type models, the observed scores are sufficient statistics for person estimates, and a standard error of measurement is provided for each distinct response vector. Since the criterion variable is a binary variable (selected or not), logistic regression was performed individually for the seven scales using the person estimates obtained from the IRT analyses. The reason for using separate logistic regressions instead of a multiple regression is that the regression coefficients might be distorted when there is high collinearity between scales. The probability of being selected as an intern is calculated based on the estimated regression coefficient for each response category by scale. Its logit form can be expressed as

$$\text{logit}(p_{dj}) = \log_e\left(\frac{p_{dj}}{1-p_{dj}}\right) = a_d + b_d j \qquad (2)$$

where p_{dj} stands for the probability of being selected given a score j (j = 1, 2, 3, 4, 5) on scale d (e.g., *knowledge, analytical skills*). a_d and b_d are the estimated regression intercept and slope, respectively, for scale d. Here the slope b_d indicates the difference in the logit of probability for two adjacent scores.

Finally, the professor ratings and the ETS mentor ratings for the selected interns were compared for the seven scales. Since only 51 applicants were selected as ETS interns, the sample size was not large enough for Rasch calibrations. Therefore, analyses involving ETS mentor ratings were based on student-observed scores instead of Rasch estimates. Correlations between the two sets of ratings and *t*-tests were used to examine the consistency of the professor ratings and ETS mentor ratings.

Results

Dimensionality Investigation

Under the multidimensional model, the reliability ranged from .84 (*analytical skills* scale) to .91 (*self-organization* scale) for the seven scales, reasonably high given the fact that each scale only has four items.

Table 2 presents the item-difficulty estimate and its standard error, fit statistics, and discrimination index for the 28 SLR items.

Item difficulty. An item was said to be difficult if the recommenders tended to give low ratings of students and easy if the recommenders tended to give high ratings. In Table 2, the items are listed by descending difficulty parameters; the larger the item-difficulty parameter, the more likely the applicants were rated low on this item. The lowest rated item is Item 2 on the *analytical skill* scale, which asks whether an applicant "has qualitative research methods and data analysis skills." Since most intern applicants came from educational or psychological measurement programs that emphasize quantitative skills, it is not surprising that this item had low overall ratings.

Note that the *knowledge* and the *analytical skills* items are the lowest-rated items overall and tend to have lower ratings than the items measuring noncognitive skills. A likely factor is that these cognitive skill items ask about more objective, observable traits than the noncognitive items do. The three items with the highest ratings all come from the *professionalism and maturity* scale. These items elicit information about fundamental human qualities such as honesty, ethical standards, and sincerity. The high ratings could result from the fact that (a) the majority of the candidates are honest and sincere; and/or (b) the raters are inclined to give high scores on these items out of kindness, since implying people are dishonest or insincere sounds

Table 2

Difficulty Estimate, Standard Error, Fit Statistic, and Discrimination for SLR Items

No.	Item	Difficulty	SE	Fit	Discrimination
A6	Has qualitative research methods and data analysis skills	-.77	.07	1.48	.47
K4	Has knowledge of program evaluation	-.82	.07	1.27	.58
K2	Has knowledge of behavioral science research methods	-1.53	.09	.83	.63
A8	Is skilled in statistical methods	-2.02	.08	.90	.53
K3	Has knowledge of educational or psychological assessment	-2.05	.08	.96	.57
A7	Has demonstrated skill in measurement and assessment techniques	-2.05	.08	.85	.56
K1	Has foundational knowledge in some branch of behavioral science (e.g., educational theory, psychology, economics).	-2.15	.08	.93	.66
A5	Has quantitative research methods and data analysis skills	-2.16	.08	.90	.57
C10	Speaks in a clear, organized, and logical manner	-2.34	.10	.94	.70
C11	Writes with precision and style	-3.34	.10	1.23	.64
C12	Listens well and responds appropriately	-4.09	.11	.94	.75
M14	Can overcome challenges and setbacks	-4.31	.12	1.04	.70
C9	Demonstrates clear and critical thinking	-4.48	.11	.96	.77
O19	Makes good decisions	-4.58	.13	.94	.79
T25	Shares ideas easily	-4.58	.13	1.15	.69
O17	Organizes work and time effectively	-4.78	.13	1.19	.75
T27	Works well in group settings	-4.85	.13	.96	.70
M13	Maintains high standards of performance	-4.89	.12	.90	.75
T26	Supports the efforts of others	-4.95	.13	.93	.70
M16	Has a high level of energy	-5.02	.12	1.03	.67
M15	Seek out opportunities to learn	-5.12	.13	1.00	.64
O18	Set realistic goals	-5.71	.13	1.05	.76
P24	Regulates own emotions appropriately	-6.27	.14	1.13	.73
O20	Can work independently of others	-6.38	.14	1.1	.76
T28	Behaves in an open and friendly manner	-6.42	.14	1.15	.60
P21	Maintains high ethical standards	-6.86	.15	.85	.71
P23	Is dependable	-7.29	.15	1.19	.70
P22	Demonstrates honesty and sincerity	-7.30	.15	.93	.66

Note. A=analytical skills, K=knowledge, C=communication, M=motivation, O=self organization, P=professionalism & maturity, and T=teamwork. The numeral in the far left column corresponds to the item number in Table 1.

fairly harsh. Although most applicants obtained high ratings on these items, this does not suggest that the scale is not informative. Its potential effectiveness lies in that it may help to screen out candidates with low ratings on this scale.

Item fit. For *ConQuest*, the item fit is indicated by the weighted fit mean square (WFMS) statistic. WFMS detects discrepancies between the multidimensional model adopted for our analysis and the observed responses (Wright & Masters, 1982). WFMS is expected to have a value of one when there is satisfactory agreement between the models and the data (Wu et al., 1997). Adams and Khoo (1996) have defined that a value between .75 and 1.33 can be considered to fit a model reasonably well. An item with a fit index larger than one tends to have more variability in the data than the model predicts and is often associated with a low discrimination index. An item with a fit index less than one indicates less observed variability than the model predicts and is often associated with a high discrimination index.

The fit statistics (see Table 2) for all the SLR items are within a reasonable range of .75-1.33, except for Item 6, which is also the lowest rated item in the SLR item pool. As discussed above, this item asks about candidates' *qualitative* research skills, while most of the candidates have been trained in graduate programs emphasizing *quantitative* skills. The mismatch between this item and the analytical scale contributes to the large misfit. Thus, this item probably should be removed from the SLR for the purpose of intern selection, since quantitative skills are what are valued for the ETS intern program.

Discrimination index. The discrimination index measures the ability of an item to distinguish between candidates with high and low ratings. This index comes from the CTT framework since the Rasch-type models assume a constant discrimination parameter across all items. For each item, the correlation between the candidate's score on the item and the aggregate total score is used as an index of discrimination. If p_{ni} is the proportion of score levels that candidate n achieved on item i, and $p_n = \sum_i p_{ni}$ is the sum of the proportions of the maximum score achieved by candidate i, then the discrimination is calculated as the product-moment correlation between p_{ni} and p_n for all candidates. If the items are scored dichotomously, this index will be the usual point-biserial index of discrimination. By conventional rule the discrimination should be higher than .25. Table 2 shows that all the SLR items have a satisfactory discrimination index, ranging from .47 to .79.

Wright map. The Wright map depicts the person trait distribution and the item difficulty distribution. For the Rasch-type model, the probability of answering an item correctly, or in this case, rating a student, is conceived as a function of the candidate trait level and the item difficulty level (Wright & Stone, 1979). A Wright map is a visual representation of this relationship in which the candidate estimates and the item estimates are put onto the same interval scale to allow direct comparisons.

Figure 1 shows the Wright map for each of the seven dimensions. The far left column represents the logit value, which is the metric for both the candidate estimate and the item estimate. The larger the logit value, the higher the candidate's trait level, and the less likely the items will be endorsed. Each x represents an applicant, and the numerical numbers represent item numbers. For example, Item 4 in Figure 1 is the same item as Item 4 in Table 1, which is an item on the *knowledge* scale. The relative position of the candidate estimate and the item estimate determines the probability of this candidate's rating on this item. For example, if an applicant estimate is above an item estimate on the Wright map, the probability that she or he receives a high rating on this particular item is more than .5. The more the applicant estimate exceeds the item estimate, the higher the probability of a high rating on a given item.

Items on the *knowledge* scale and the *analytical skills* scale are above all the other items on the Wright map (see Figure 1), which suggests that the applicants were rated lower on cognitive items than on noncognitive items. This phenomenon can be interpreted as a general tendency in a reference-writing context rather than a bias of the SLR design per se (Kim & Kyllonen, 2006). Most of the items clustered together reasonably well within each scale, suggesting that they have similar difficulties.

In general, the candidate estimates are well above the item estimates, suggesting that it is easy to get high ratings on these items, including the relatively less endorsed *knowledge* and *analytical skills* items. This is consistent with previous findings that there is a general inclination for raters to rate applicants highly on these kinds of items (Kim & Kyllonen, 2006; Kyllonen & Kim, 2005).

Correlation. The correlations between dimensions are provided in Table 3.These scales are moderately to highly correlated, with the lowest correlation (r = .39) between the *analytical skills* and the *teamwork* scale and the highest correlation (r =.78) between the *motivation* and the *self-organization* scale. High correlations possibly occur due to the *halo effect*, in which ratings on previous scales influence subsequent scales, or there might be some substantive relationship

Logit	Knowledge	Analytical	Communication	Motivation	Organization	Professionalism	Teamwork
6							
5							
4	X						
	X	X	X	X		X	
3	XX	XX	XXX	XX	X	X	XX
	XX	XX	X XX	XXX	XX	X	XX
2	XXX	XXXX	XXX	XXX	XXX	XX	XXX
	XXXX	XXXXX	XXXXX	XXXX	XXX	XXX	XXX
1	XXXXXX	XXXXXX	XXXXX	XXXX	XXXX	XXXX	XXX
0	XXXXXXXX	XXXXXXXX	XXXXXX	XXXXX	XXXX	XXXX	XXXX
	XXXXXXXX	XXXXXXXXX	XXXXXXX	XXXXX X	XXXXXX	XXXXX	XXXXX
-1	XXXXXXXX **4**	XXXXXXXXX **6**	XXXXXX	XXXXX	XXXXX	XXXXX	XXXXXX
	XXXXXX **2**	XXXXXX	XXXXX	XXXX	XXXX	XXXXX	XXXXX
-2	XXXX **1 3**	XXX **5 7 8**	XXXX	XXX	XXXX	XXXX	XXXX
	XX	XX	XXXX **10**	XXX	XXX	XXX	XXX
-3	X	X	XXX	XX	X	XXXX	XXX
			XX **11**	X	XX	XXX	XX
-4			X **12 9**	**14**	XX	XX	XX
					X **17 19**	X	XX **25 27**
-5				**13 15 16**		X	XX **26**
					18		X
-6					**20**	**24**	**28**
-7						**21**	
						22 23	

***Figure 1*. Wright map of the seven scales.**

Note. For each scale, X stands for the person and the numbers stand for the item number. Each X represents 7.7 cases.

between being organized and being motivated. It is plausible that well organized people are relatively highly motivated. Conversely, there is no reason to expect unmotivated people to make the effort to be organized.

Table 3

Pearson Correlations Between Scales Rated by Professors

Scale	N	Knowledge	Analytical	Communication	Motivation	Organization	Professionalism
Analytical	414	.67					
Communication	414	.57	.46				
Motivation	414	.53	.45	.66			
Organization	414	.59	.52	.70	.78		
Professionalism	414	.49	.40	.60	.71	.72	
Teamwork	414	.49	.39	.67	.68	.70	.74

SLR Prediction of ETS Intern Selection

To examine the SLR prediction of intern selection, the scale estimates produced by the multidimensional model were used as independent variables in the logistic regression models. Results from separate logistic regressions by scale are provided in Table 4. Two cognitive scales showed statistical significance in terms of predicting intern selection.

As described earlier, the probability of being selected as an intern is calculated based on the logistic regression coefficients. Since the SLR items are represented by a five-point response set, we selected the five integer score points (1, 2, 3, 4, and 5) to illustrate the prediction of the selection.

Figure 2 shows the probability of being selected by mean score on a given scale. For example, for candidates with a mean score of 1 on the *analytical skills* scale, their estimated probability of being selected as an intern is .11. The probability increases monotonically as the mean score increases, and the probability becomes .48 if they achieve the maximum mean score of 5 on this scale. Fifty-one interns were selected out of 414 applicants, so the random probability of being selected is about .12 (51/414). The fact that .48 is four times greater than .12 suggests that the *analytical skills* scale is a fairly substantial predictor. In addition, the differential predicting power of the scales provides some evidence for a multidimensional model.

Table 4

Logistic Regression Coefficients, Standard Errors, and Significance

	Intercept	B	SE	Wald
Knowledge	-2.07	0.52	0.22	5.77*
Analytical	-2.10	0.68	0.25	7.48**
Communication	-2.09	0.21	0.11	3.65
Motivation	-2.06	0.20	0.11	3.30
Organization	-2.08	0.16	0.08	4.10
Professionalism	-2.04	0.15	0.08	3.48
Teamwork	-2.00	0.14	0.09	2.45

Note. Each of the seven scales score was used to predict the selection status separately. B stands for the regression coefficient and SE is the standard error of the coefficient. Wald test denotes the significance of the logistic regression coefficient.

* $p < .05$. ** $p < .01$.

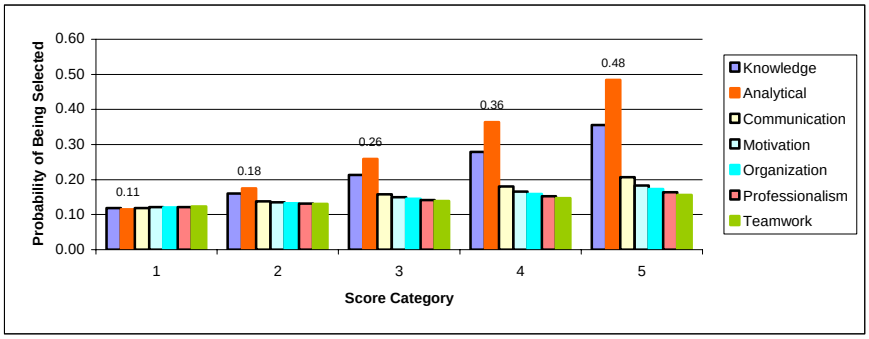

***Figure 2*. Probability of being selected as an intern for each score category.**

Overall, the *knowledge* and *analytical skills* scales appear to be the best predictors for intern selection. Intern applicants who scored high on these two scales had a much higher probability of being selected as a summer intern. The noncognitive measures demonstrated less predicting power, with the *communication* scale the best noncognitive predictor. Note that the low prediction ability of the noncognitive measures could be particularly related to the intern selection scenario. Since the interns were only present for about two months, the selection committee might

have focused on what candidates could produce during a limited time frame rather than on other qualities such as *teamwork* or *organization* that might be viewed as having higher value in a longer term position. The stakes are also lower for temporary positions such as internships than they are for regular professional positions or academic placements. Therefore, the role of noncognitive measures could be underestimated in the process of selecting summer interns. There are very likely other circumstances where *teamwork* or *professionalism* is the essence of the job, so these measures could be considered highly desirable in those selection scenarios.

ETS Performance Evaluation Using the SLR

After the 51 interns completed their internship, each of them was evaluated by their ETS mentor(s) using the same SLR. The descriptive statistics of the ETS ratings are provided in Table 5. The *knowledge* and *analytical skills* scales yielded the lowest scale means, which is consistent with previous findings that cognitive measures tend to elicit lower ratings from professors (Kim & Kyllonen, 2006; Kyllonen & Kim, 2005). These two scales likely invite the most objective ratings because of the specificity of the items, as opposed to the more general nature of the noncognitive measures. Furthermore, the recommenders probably are aware that a low rating on noncognitive items asking about fundamental qualities such as honesty could have a substantially negative impact on the applicant. This tendency to be lenient makes it likely that the ratings on some of the noncognitive variables will be higher than those on cognitive variables.

Table 6 shows the correlations between scales rated by ETS mentors. As discussed earlier, given the small sample size ($N = 51$), analyses involving ETS mentors were based on observed scores instead of Rasch estimates. The correlations between dimensions ranged from moderate to high (i.e., .42 to .86). The correlations between noncognitive variables were relatively higher than those between cognitive variables, which was also true for professor ratings.

Paired-sample *t*-tests were used to compare the professor and ETS ratings within each scale for the selected interns. Results in Table 7 show that the ratings the interns obtained from their professors were consistently higher than those received from their ETS mentors. Each comparison showed statistical significance at $\alpha = .01$ level. An effect size indicated by Cohen's *d* (Cohen, 1988) was calculated for each comparison. All the effect sizes are above .8, which is considered large in the social science context. The *professionalism and maturity* and *motivation* scales showed the largest effect size.

Table 5

Descriptive Statistics for the ETS Mentor Ratings of the SLR

Scale/Item	*N*	Mean	SD	Scale mean	Scale SD
Knowledge					
1. Has foundational knowledge in some branch of behavioral science (e.g., educational theory, psychology, economics).	51	3.31	.73	3.15	0.70
2. Has knowledge of behavioral science research methods.	51	3.33	.87		
3. Has knowledge of educational or psychological assessment.	51	3.11	.91		
4. Has knowledge of program evaluation.	51	2.70	.98		
Analytical skills					
5. Has quantitative research methods and data analysis skills.	51	3.49	.84	3.30	0.78
6. Has qualitative research methods and data analysis skills.	51	3.42	1.03		
7. Has demonstrated skill in measurement and assessment techniques.	51	3.12	.97		
8. Is skilled in statistical methods.	51	3.25	1.01		
Communication skills					
9. Demonstrates clear and critical thinking.	51	3.44	1.01	3.39	0.91
10. Speaks in a clear, organized, and logical manner.	51	3.48	.91		
11. Writes with precision and style.	51	3.19	.85		
12. Listens well and responds appropriately.	51	3.50	1.09		
Motivation					
13. Maintains high standards of performance.	51	3.58	.86	3.55	0.92
14. Can overcome challenges and setbacks.	51	3.48	1.03		
15. Seek out opportunities to learn.	51	3.58	.98		
16. Has a high level of energy.	51	3.58	1.03		
Self-organization					
17. Organizes work and time effectively.	51	3.34	1.04	3.35	0.89
18. Set realistic goals.	51	3.29	.92		
19. Makes good decisions.	51	3.30	.84		
20. Can work independently of others.	51	3.52	1.02		
Professionalism & maturity					
21. Maintains high ethical standards.	51	3.57	.89	3.54	0.89
22. Demonstrates honesty and sincerity.	51	3.58	.90		
23. Is dependable.	51	3.50	1.03		
24. Regulates own emotions appropriately.	51	3.60	.91		
Teamwork					
25. Shares ideas easily.	51	3.33	1.01	3.43	0.85
26. Supports the efforts of others.	51	3.35	.94		
27. Works well in group settings.	51	3.45	.86		
28. Behaves in an open and friendly manner.	51	3.67	.92		

Table 6

Correlations Between Scales Rated by ETS Mentors

Scale	*N*	Knowledge	Analytical	Communication	Motivation	Organization	Professionalism
Analytical	51	0.79					
Communication	51	0.61	0.54				
Motivation	51	0.61	0.58	0.82			
Organization	51	0.76	0.73	0.81	0.82		
Professionalism	51	0.68	0.58	0.80	0.74	0.85	
Teamwork	51	0.56	0.42	0.82	0.78	0.83	0.86

Table 7

Comparison of Professor and ETS Mentor Ratings

		Professor		ETS		*t*	Effect Size *d*
	N	Mean	SD	Mean	SD		
Knowledge	51	3.77	.57	3.15	.70	6.07**	.97
Analytical	51	4.04	.56	3.30	.78	7.05**	1.09
Communication	51	4.10	.62	3.39	.91	6.36**	.91
Motivation	51	4.41	.38	3.55	.92	6.07**	1.22
Organization	51	4.14	.47	3.35	.89	6.23**	1.11
Professionalism	51	4.44	.49	3.54	.89	6.63**	1.25
Teamwork	51	4.21	.63	3.43	.85	6.45**	1.04

** $p < .01$

The significant *t* values suggest that on average, the ETS ratings are lower than the professor ratings. The correlation between the two sets of ratings inform us whether the difference in ranking is systematic, since the two sets of ratings can have the same mean yet correlate poorly with each other. The means and correlations are presented in Figure 3. The *motivation* and *professionalism and maturity* scales have fairly low correlations between professor and ETS ratings, .20 and .27 respectively. Obviously the ETS mentors had different opinions from professors on whether the interns maintained high standards of performance and had been dependable. Also, the amount of interaction varies between interns and their ETS mentors across

projects as some mentors may have had more contact with their interns than others. Ratings on other scales are moderately correlated, with coefficients ranging from .41 to .60. The *communication skills* scale has the highest correlation of .60, probably because of the relative objectivity of the abilities measured by the items, such as speaking, writing, and listening.

Discussion

Through rigorous research grounded in the literature and information accumulated from focus groups and interviews with faculty members, the SLR was created in order to significantly improve the way traditional letters of recommendation are utilized (Kyllonen et al., 2005; Walpole et al., 2005). As the literature reveals, traditional letters are fraught with complications, specifically with respect to issues of reliability (Aamont et al., 1993; Range et al., 1991). The SLR is standardized, an essential feature that allows for individuals to be readily compared to one another and a key component for decision makers who need to assess numerous, diverse candidates. The SLR is unique in that it translates qualitative information found in traditional letters into quantifiable assessments. This allows the reader of the SLR to gain a coherent interpretation of a candidate's strengths and weaknesses in terms of the different dimensions represented, thus decreasing the possibility of misinterpretations. Evaluators are also provided the opportunity to elaborate on each of the seven scales by writing narratives containing specific examples or comments concerning the candidate's qualities and performance.

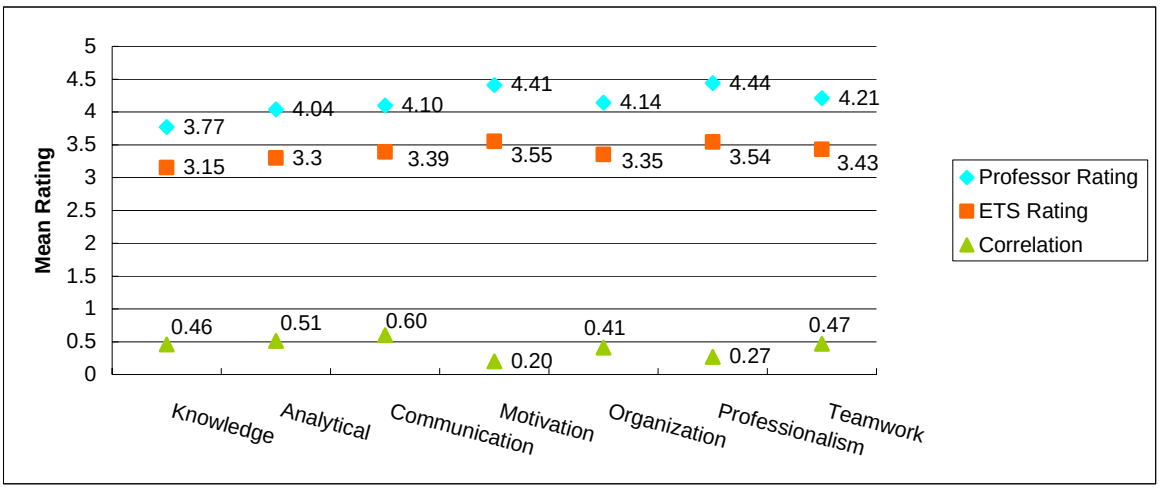

Figure 3. **Means and correlations of professor rating and ETS rating.**

The main objective of transferring qualitative reference letters to quantitative measures is to standardize the application procedure and avoid misinterpretation of the letters by decision makers. For example, in some cultures it is against the societal convention to comment negatively on candidates' qualities: in order to be honest and impartial, recommenders often only vaguely imply the less competitive aspects of the candidates. If the review committee does not fully understand these "secret codes," misinterpretations are likely to occur. For example, a comment like "this person spent a long time working on this project" could suggest that this candidate works hard and is persistent. However, it could also indicate that this person is not competitive and takes a long time to accomplish a task. The SLR possibly reduces the ambiguity by providing a common ground for comments and evaluation.

Noncognitive variables such as motivation and personality have been investigated by numerous conceptual and empirical studies (Briel et al., 2000; Hough, 2001; Kuncel et al., 2001; Schmidt, 1994), and a largely agreed-upon conclusion is that they have been important predictors for academic and professional achievement. The SLR described in this study represents important dimensions of candidate abilities and qualities, covering a balanced combination of both cognitive and noncognitive factors.

Previous research (Kim & Kyllonen, 2006; Kyllonen & Kim, 2005) examined the quality of the SLR using CTT and Rasch models. Besides finding that the items generally showed satisfactory psychometric properties, the authors raised the point of distinguishing between theoretical and practical dimensionality of the SLR. Based on the seven-dimensional IRT model, there is no compelling evidence against the original design of seven scales. At the individual-item level, all but one item showed reasonable fit statistics, and all items showed good discrimination power. The reliability was high for all scales even with only four items per scale, indicating strong internal consistency.

The current study found some scales to be highly correlated. However, there are two reasons to leave the scales as they are instead of collapsing them. First, the use of the SLR is not limited to academic applicants. It can be extended to business selections, for which one scale might be weighted more than another. For example, for jobs that require confidentiality, whether the candidate "maintains high ethical standards" could be a first priority to consider, while other jobs may value most whether the employee "has a high level of energy." Collapsing the scales eliminates the possibility of differentiating between the different qualities assessed by the items on

each scale. Reporting seven scale scores also gives users the flexibility to decide which of the scales they want to stress when evaluating candidates, depending on the nature of the selection. Second, in the intern application situation, the high correlations might occur due to the *halo effect*, which might have partially disguised the distinctions between scales. As long as a sufficient number of records are available for each recommender, it is possible to monitor and adjust the rater effect through IRT models (Patz, Junker, Johnson, & Mariano, 2002; Wang & Wilson, 1996, 2005).

In terms of predicting intern selection, the cognitive scales showed better prediction than the noncognitive counterparts did. As discussed earlier, this could be due to the fact that intern positions were short-term so the mentors tended to focus more on objective, cognitive measures rather than on indirect measures such as *motivation* and *teamwork*. Also, cognitive abilities are relatively easy to identify and measure. It could have been difficult for mentors to distinguish and comment on the noncognitive scales due to the short time period in which the interns were observed.

This study also examined the consistency between prediction and criterion ratings. The comparisons revealed that ETS mentors systematically gave lower ratings to the interns than the professors did. Three possible reasons are speculated to account for this discrepancy. First, the level of familiarity with the interns is different. Background data indicate that most of the professors have known the applicants for more than two years, yet the mentors have only about eight weeks to interact with the interns. Depending on the nature of their projects, some interns may have limited contact with mentors during that short period. Second, the stakes are different for the professors and the ETS mentors. Since normally professors want to place students at ETS as an intern, they are motivated to present their students in a good light. However, for most of the ETS mentors, there is no such incentive. Finally, and probably most importantly, the professors and ETS mentors evaluated the interns under different circumstances. The professors could have rated their students highly because the students asked insightful questions in classes, completed course projects satisfactorily, and demonstrated competence in other assigned tasks. However, the ETS mentors might find that the students did not have sufficient knowledge or hands-on experience to tackle real projects at ETS, since these tasks usually require a different level of sophistication. In a sense, students might have been rated in different contexts by people with different criteria on their mind, which might account for the discrepancies observed in the ratings. Bear in mind, however,

that these findings were observed from a fairly small sample (N = 51) and may not represent the true relationship between professor ratings and mentor ratings. If more data become available, the discrepancy between the two sets of third-party observations may vary in magnitude and direction for some of the scales.

In both previous research and in this study, recommenders tended to rate the applicants higher than warranted as compared to the ETS mentor ratings, perhaps out of kindness or possibly fear that low ratings might be damaging to the applicant's career (Robiner, Saltzman, Hoberman, Semrud-Clikeman, & Schirvar, 1998). To date, no effective mechanism has been developed to completely eliminate this tendency. However, a few things have been suggested to reduce the variation of potential bias caused by leniency in ratings. Ceci and Peters (1984) noted that when letters of recommendation are kept confidential, there tended to be less leniency bias than when the applicants waived their access to the letters. Therefore, the SLR used in this study was kept confidential from the intern applicants. Another viable solution is to monitor and adjust for the rater effect (e.g., applicants rated by harsh raters should be compensated). In addition, some brief comments or examples from recommenders may help the selection committee gauge the accuracy of the ratings. For example, recommenders could be asked to provide supportive, convincing evidence for an "outstanding" rating or explain why they gave the lowest rating to an applicant.

Although most recommenders tend to overrate the applicants, they still vary in their degree of leniency. Hence, for future research, more efforts should be devoted to addressing the issue of inter-rater reliability. Using modern measurement models, the rater effect can be monitored and the ratings adjusted for rater leniency or harshness (Wu et al., 1998). For the current study, although most intern applicants had SLR ratings from two professors, individual professors did not rate enough students for us to track their overall harshness or leniency. If ETS continues to use the SLR for intern selection, more records from professors should become available as they complete additional SLRs for their students. With sufficient data, the rater effect can be monitored and adjusted to contribute to the validity of the selection process.

Finally, selection committees need to be aware that subjectivity is highly likely to affect the ratings in a letter of recommendation and perhaps unavoidable. The SLR should therefore be used in conjunction with other kinds of evaluation criteria (e.g., GPA, standardized test scores) deemed important for admission or selection. Evidence from multiple sources of different types is needed to ensure a fair selection outcome.

References

Aamont, M. G., Bryan, D. A., & Whitcomb, A. J. (1993). Predicting performance with letters of recommendation. *Public Personnel Management, 22,* 81–90.

Aamont, M. G., Nagy, M. S., & Thompson, N. (1998, June). *Employment references: Who are we talking about?* Paper presented at the annual meeting of the International Personnel Management Association Assessment Council, Chicago.

Adams, R. J., & Khoo, S. (1996). *ACER Quest - The interactive test analysis system*. Camberwell, Australia: ACER Press.

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1–23.

Baxter, J. C., Brock, B., Hill, P. C., & Rozelle, R. M. (1981). Letters of recommendation: A question of value. *Journal of Applied Psychology, 66,* 296–301.

Bock, R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika, 46,* 443–459.

Briel, J., Bejar, I., Chandler, M., Powell, G., Manning, K., Robinson, D., et al. (2000). *GRE horizons planning initiative.* Unpublished manuscript, ETS, Princeton, NJ.

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement, 4,* 87–100.

Ceci, S. J., & Peters, D. (1984). A naturalistic study of the effects of confidentiality. *American Psychologist, 39,* 29–31.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Enright, M. K., & Gitomer, D. H. (1989). *Toward a description of successful graduate students* (GRE Board Professional Rep. No. 89-09). Princeton, NJ: ETS.

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.

International Personality Item Pool. (2001). *A scientific collaboratory for the development of advanced measures of personality traits and other individual differences.* Retrieved November 28, 2006 from http://ipip.ori.org/.

Kim, S., & Kyllonen, P. C. (2006). *Rasch rating scale modeling of data from the standardized letter of recommendation* (ETS Research Rep. No. RR-06-33). Princeton, NJ: ETS.

Knouse, S. B. (1983). The letter of recommendation: Specificity and favorability of information. *Personnel Psychology, 36,* 331–341.

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., & 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, Montreal, Canada.

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

Lopez, S. J., Oehlert, M. E., & Moberly, R. L. (1996). Selection criteria for American Psychological Association-accredited internship programs: A survey of training directors. *Professional Psychology: Research and Practice, 27,* 518–520.

McCarthy, J. M., & Goffin, R. D. (2001). Improving the validity of letters of recommendation: An investigation of three standardized reference forms. *Military Psychology, 13,* 199–222.

Miller, R. K., & Van Rybroek, G. J. (1988). Internship letters of recommendation: Where are the other 90%? *Professional Psychology: Research and Practice, 19,* 115–117.

Muchinsky, P. (1979). The use of reference reports in personnel selection: A review and evaluation. *Journal of Occupational Psychology, 52,* 287–297.

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27*(4), 341–384.

Range, L. M., Menyhert, A., Walsh, M. L., Hardin, K. N., Ellis, J. B., & Craddick, R. (1991). Letters of recommendation: Perspectives, recommendations, and ethics. *Professional Psychology: Research and Practice, 22,* 389–392.

Robiner, W. N., Saltzman, S. R., Hoberman, H. M., Semrud-Clikeman, M., & Schirvar, J. A. (1998). Psychology supervisors' bias in evaluations and letters of recommendation. *Clinical Supervisor, 16*, 49–72.

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.

Walpole, M. B., Burton, N. W., Kanyi, K., & Jackenthal, A. (2002). *Selecting successful graduate students: In-depth interviews with GRE users* (GRE Board Research Rep. No. 99-11). Princeton, NJ: ETS.

Walters, A. M., Kyllonen, P. C., & Plante, J. W. (2003). *Preliminary research to develop a standardized letter of recommendation.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Wang, W. (1999). Direct estimation of correlations among latent traits within IRT framework. *Methods of Psychological Research Online, 4*(2), 47–70.

Wang, W., & Wilson, M. (1996). Comparing multiple-choice items and performance-based items using item response modeling. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory into practice* (vol. III; pp. 167-194). Norwood, NJ: Ablex.

Wang, W., & Wilson, M. (2005). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement, 29*, 296–318.

Wang, W., Wilson, M., & Adams, R. J. (1997). Rasch models for multidimensionality between and within items. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice* (vol IV; pp. 139–154). Norwood, NJ: Ablex.

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

Wright, B. D., & Stone, M. (1979). *Best test design.* Chicago: MESA Press.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1997). ConQuest: Generalized item response modeling software [Computer software]. Melbourne, Australia: ACER.

Notes

[1] Analyses were also conducted using the mean score, and the results were similar to those yielded using the maximum score.

Appendix

Example of Instructions, Items, and Response Choices on the Standardized Letter of Recommendation

SLR *Standardized Letter of Recommendation*

Evaluator Form 2 – Knowledge

Evaluator Information	Knowledge	Analytical Skills	Communication Skills	Motivation	Self-Organization	Professionalism and Maturity	Teamwork	Project Fit

Instructions: Below is a series of statements pertaining to a student quality. Based on your experiences, please indicate how the applicant compares to other students at your institution using the rating system and comment field. For items in which you do not have appropriate knowledge of the student comment, please select, **"I don't know."**

Items

1.	Has foundational knowledge in some branch of behavioral science (e.g. educational theory, psychology, economics).	Outstanding ▽
2.	Has knowledge of behavioral science research methods.	Average ▽
3.	Has knowledge of educational or psychological assessment.	I don't know ▽
4.	Has knowledge of program evaluation.	Outstanding ▽

Add examples and general comments on Knowledge, if appropriate.