Short, Simple, and Specific: The Influence of Item Design Characteristics in Multi-Source Assessment Contexts

Stéphane Brutus
John Molson School of Business
Concordia University

and

Jeffrey Facteau
Organizational Consultant
Georgia

The present study investigates the impact of item characteristics on multi-source performance assessment. Three item characteristics (syntax, double-barreledness, & behavioral specificity) were linked to the psychometric properties of items used by self, subordinates, peers, and supervisors as operationalized by the relationship between the item and the performance dimension it is intended to measure. Results show that syntax, a linguistic index that pertains to the length of items, is related to the psychometric properties of all rating sources except subordinates. The implications of this effect for the design of multi-source assessment instrument are discussed.

Introduction

"[DONALD] ACTS DECISIVELY IN TIMES OF CRISIS"

Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree
☐	☐	☐	☐	☐

Every year, a countless number of individuals are presented with similar statements in order to guide their evaluation of a coworker. Without question, the use of performance items in combination with a rating scale represents the most common methodology to assess individual performance in organizations (Murphy & Cleveland, 1995). Accordingly research on the topic is abundant (e.g., Austin & Villanova, 1992). Typically, performance evaluation surveys are composed of items pertaining to on-the-job behaviors (e.g., Austin & Villanova, 1992; Kavanaugh, 1971; Murphy & Cleveland, 1995). These items represent written stimuli, connoting relevant behaviors for the organization, that cue raters and enable them to judge the frequency or the quality of those particular behaviors for a ratee.

Funding for this paper was provided by the Social Sciences and Humanities Research Council of Canada (grant #410-2000-0818) whose support is gratefully acknowledged.
The authors thank John Fleenor, Cindy McCauley, Manuel London and Jean Leslie for their support of this research project.

In the past decade, multi-source assessment (MSA) has brought significant changes to the manner by which performance items are used (Brutus & Derayeh, 2002). The traditional performance appraisal processes calls for supervisors to evaluate the performance of their subordinates. In MSA, those same items are also used to rate one's own performance, the performance of a supervisor, a peer, and even that of someone from outside the organization (London & Smither, 1995). Of interest in this paper is the impact of linguistic properties of performance items on their psychometric quality when completed by different types of raters.

From an organizational standpoint, there exist many reasons behind the adoption of MSA such as breaking down organizational barriers and increasing communications between employees (Waldman & Atwater, 1998). However, most of the appeal of MSA stems from the expected gains in reliability and validity that ensue from using multiple raters (Borman, 1997; Murphy, Cleveland, & Mohler, 2001). To date, the presence of these gains is still debated as research on the reliability and validity of MSA

has not yielded consistent results. Core to this issue is the appropriateness of aggregating performance ratings derived from multiple raters; a computation inherent to the purported gains in reliability and validity achieved via MSA. Facteau and Craig (2001) state that "one could expect that a given manager's ratings on any one of the items has the same meaning regardless of whether the item was rated by a peer, superior, subordinate, or the actual manager" (p.15). Some researchers have uncovered evidence that the relationship between items and underlying performance constructs are the same across rating sources or equivalent (Maurer, Raju, & Collins, 1998 for peer and subordinates; Facteau & Craig, 2001 for all four sources) while others have not (Lance & Bennett, 1997 for all four sources).

The issue of measurement equivalence across all rating sources is critical for MSA and, for this reason, much effort is being deployed to achieve it. In the *Handbook of Multisource Feedback*, VanVelsor and Leslie (2001) prescribe a very careful approach to the development of MSA instruments to ensure the reliance on performance items that 'work' well for all sources. This approach, however useful for practitioners, is purely inductive and does not provide any insights as to the cause of item equivalence (or lack thereof). One of the keys to resolving this issue is to look closely at the stimulus used to cue individual ratings. Performance items are rooted in written language and they can be analyzed in terms of their linguistic properties. In his methodological review, Schwarz (1999) stated that self-reports "…are a fallible source of data and minor changes in *question wording* (emphasis added), question format, or question context can result in major changes in the obtained results" (p.93). Surprisingly, no research has systematically addressed the impact of linguistic properties of performance items on individual ratings. In this paper we examine the role of linguistic properties of performance items in determining the psychometric quality of the items when completed by different groups of raters (i.e., self, subordinates, peers, and supervisors). Although the operationalization of psychometric quality could take many formats we conceptualized it as the relationship between an item and the performance dimension it is intended to measure. In a confirmatory factor analysis (CFA) framework such relationships are reflected by each item's loading on their intended factor or R. As the main dependent variable of this study, R is determined by the mean and variance of intercorrelations of items sharing the same factor. Thus, we pose that R, as a dependent variable, would be sensitive to variations in item characteristics because these characteristics are likely to influence these intercorrelations. According to this perspective, a poor item is one that is only weakly related to the factor it was written to measure, whereas a good item is one that bears a strong relationship to its intended factor. In the following sections, we describe several different linguistic properties of performance items and how they may relate to the psychometric quality of the items when completed by the individual himself or herself, a subordinate, a peer, and a supervisor.

Item Syntax

The systematic analysis of the properties of language has been the focus of the generative linguistic theory developed by Noam Chomsky (1965; 1986). The principles, proposed by Chomsky, are also referred to as the Universal Grammar and contain several components, the simplest of which is syntax. Syntax is concerned with the structure of sentences, the most basic elements of which are constituents. Constituents represent strings of words with a certain degree of internal cohesion, and with certain formal properties. In other words, syntax is the manner by which words are ordered and grouped together into constituents.

The syntactic analysis of performance items involves the decomposition of the item into its base constituents. For example:

Item (a):[This manager] [[has[personal warmth]]

C1 C2 C3

Item (b):[This manager] [has[[[solid relationships

C1 C2 C3

[with higher management [in times of crisis]]]].

C4 C5

Item (a) has three constituents: a subject (C1), a verb (C2) and a modifier (C3); while item (b) has five constituents: a subject (C1) a verb (C2), and three modifiers (C3, C4 & C5). Accurate responses to performance items are contingent upon the understanding of the item as intended by the item designer. Syntax is related to the complexity of sentences and the ease by which they can be deciphered. The syntax of sentences has been found to have many psychological correlates such as language acquisition (e.g., Robert, 1998) and recall (e.g., Rychlak, Stilson, & Rychlak, 1993). Many have argued for the use of syntactically simple survey items in order to enhance the quality of ratings (e.g., Sudman & Bradburn, 1982; Rogelberg & Waclawski, 2001). Concise items are more likely to be processed accurately by raters because they increase the likelihood that the respondents' understanding of the item matches what the researcher had in mind (Schwarz, 1999). In other words, we pose that the syntax of items will have a direct bearing on each item's loading on their intended factor or R. Concise items will be more closely related to the underlying performance dimension than long items.

Another syntactic element that may have an effect on the psychometric properties is the number of behavioral referents present in the performance item. The literature on survey designs argues for the use of items that pertain to

one, and only one, behavior. In the present study we also investigate the effects of "double-barreled" items, as these types of items are often referred to, because they can obviously lead to different interpretations by raters. For example, the following item: "This individual can adequately resolve conflicts and establish strategic goals" clearly evokes two distinct behaviors. The direct empirical demonstration of the effect "double-barreledness" has, to our knowledge, never been made despite the widespread acceptance of this effect in the literature.

Behavioral Specificity

So far, we have discussed two characteristics of performance items related to the structure of sentences. In this section we address the deeper properties of language or the meaning of performance items. As stated earlier, performance items aim to communicate relevant job behaviors to the rater. However, the expansiveness of language allows for multiple linguistic representation of any particular behavior and a key question in designing items is whether the linguistic formulation chosen describes the targeted behavior with enough precision. The prescription of behavioral specificity is widely imparted in the performance appraisal literature (e.g., Murphy & Cleveland, 1995; Villanova & Bernardin, 1991). A behaviorally specific item is one that not only narrowly identifies the behavior to be evaluated but also provides, when possible, a contextual frame within which the target behavior is expected to occur. In reference to behavioral ratings, Aiken (1997) states that: "In general, errors in rating are smaller if each characteristic or behavior being rated is described as objectively as possible with reference to some actually observed activity. Rating statements should be short and clear, avoiding abstract, general terms such as honest, loyal, superior, and average" (p. 50). Cognitive models of performance ratings also suggest that a behavioral stimulus that is narrowly defined yields more accurate performance information than one that is vague and ambiguous (Feldman, 1981; 1986; DeNisi & Williams, 1988). Survey methodologists have found that survey items targeting specific constructs are less likely to be influenced by contextual effects such as the content of the preceding item in the survey (Harrison & McLaughin; 1995; Schwarz, 1999). In sum, the main argument for behavioral specificity is that it reduces idiosyncratic variations in the interpretation of the item. For example:

Item (c): This manager inspires **others**.
Item (d): This manager inspires **subordinates**.

C1 C2 **C3**

Item (c) is less specific than item (d) because its constituent #3 (i.e., "others") refers to an unspecified group of individuals while the same constituent in item (d) explicitly targets a group of individuals. The interpretation of item (c), then, is more likely to vary across raters. In sum, behavioral specificity increases the consistency of interpretation of performance items across raters and, as a result, reduces error found in the relationship between items and the underlying performance dimensions. Our first set of hypotheses pertain to the main effects of item syntax, double-barreledness, and specificity on their psychometric properties.

Hypothesis 1a: There is a negative relationship between the syntax of performance items and their psychometric properties.

Hypothesis 1b: There is a negative relationship between the double-barreledness of performance items and their psychometric properties.

Hypothesis 1c: There is a negative relationship between the specificity of performance items and their psychometric properties.

Rater influences on item interpretation

As stated earlier, the introduction of MSA in many organizations has resulted in the use of performance items by various rating groups. It is hypothesized that the syntax and specificity of performance items will have different effect for those different groups. Naturally, the process of rating a self, a supervisor, a peer, or a subordinate differs substantially (Borman, 1997). Here we make a distinction between those individuals rating subordinates and those rating themselves, a peer, or a supervisor (non-traditional raters) and pose that the effect of syntax will be greater for non-traditional raters than it is for supervisors. We offer two factors that may explain this differential effect. First, downwards evaluations (i.e., a supervisor evaluating a subordinate) are still the most common in organizations and individuals have greater expertise in the evaluation of subordinates than in the evaluation of any other target (i.e., self, peer, or supervisor). Performance evaluations fall within the formal duties of supervisors and, as a result, these raters are experienced in rating subordinates—they are often trained to do so (Smith, 1986). Accordingly, we expect the negative effects stemming from item complexity to be mitigated by the expertise of supervisors. Another factor is that of rater accountability. London, Smither and Adsit (1997) discussed the importance of rating accountability to obtain valid ratings. In MSA, supervisors are the only rater group truly accountable for their ratings and not protected by anonymity in MSA (Antonioni & Woehr, 2001; Brutus & Derayeh, 2002) and, thus, are probably the most motivated to decipher complex performance items. Conversely, non-traditional raters who posses lower levels of expertise and lesser accountability are more likely to be negatively influenced by item complexity. Our second set of hypotheses pertain to an interactive effect of item

characteristics (syntax and behavioral specificity) and rating source on the psychometric properties of items.

Hypothesis 2a: Rating source will moderate the negative relationship between the syntax of performance items and their psychometric properties in that the relationship will be stronger for self, peers and subordinates than it is for supervisors.

Hypothesis 2b: Rating source will moderate the negative relationship between the specificity of performance items and their psychometric properties in that the relationship will be stronger for peers and subordinates than it is for supervisors.

Method

Sample

Ratings from three different MSA surveys were used for this study (labeled instrument A, B, and C). Data from the three surveys was gathered from individuals participating in developmental activities in various organizations.

Ratings from 2,120 self-raters, 5,865 subordinates, 6,995 peers, and 1,829 supervisors were available for instrument A. These data were provided by a leadership development institute that uses this instrument as part of its activities. This instrument contained 106 items that measured 16 performance dimensions.

Data for instruments B and C were obtained from two separate utility companies in the southeastern United States. For instrument B, ratings from 270 self-raters, 969 subordinates, 1,323 peers, and 335 supervisors were available. The managers who received ratings represented all major areas of the company's business (e.g., power generation, transmission, corporate functions, etc.). This instrument contained 38 items that measured eight performance dimensions.

For instrument C, ratings were obtained from 1,555 self-raters, 1,517 subordinates, 5,704 peers, and 1,795 supervisors in a second utility organization. The instrument contained 44 items that measured eight performance dimensions

It is important to note that the procedures that were used for administering these three instruments were nearly identical. In all cases, managers rated their own performance, and distributed MSA surveys to their peers, subordinates and supervisors. Ratings obtained from peers and subordinates were anonymous.

The extent to which items in these instruments were submitted to psychometric screening varied. Instrument A was developed using factor analysis and validation procedures using an external criterion. Items included in instrument B and C, on the other hand, were not submitted to item analysis and were only reviewed for content and readability.

Measures

Two groups of experts were used to assess the characteristics of the items used in this study. The first group of experts was composed of two students in the linguistic department of the university of one of the authors. The second group was composed of nine individuals, all of whom with a Ph.D. in Industrial-Organizational psychology, with an average of 11.6 years of experience in performance assessment. Each expert was presented with the 188 performance items and asked to rate each one on a specific dimension. The two linguistic experts were asked to assess the syntax dimension while the nine psychologists were asked to assess the "double-barreledness" dimension. Specificity was operationalized in linguistic and psychological terms. Hence, both groups of experts rated the specificity of the performance items; each group using a framework relevant to its discipline (linguistics and psychology) to create separate specificity indices. The measurement process is described in more detail in the following section.

Syntax. The syntax index represents the number of constituents present in each item. Syntax ratings were made for each performance item according to a protocol developed by Chomsky (1965). A rating of 1 was assigned to items that are restricted syntactically to a verb followed by an object which is obligatorily selected by that verb. For example, "this manager [[$_{v}$emphasizes] [$_{Obj}$ cooperation]]". A rating of 2 was attributed to items with one optional complement or verbal modifier. These modifiers add supplementary information about the event/state/process described by the verb but their absence does not result in ungrammaticality. For example, "this manager responds *with insensitivity* to the feelings and views of others." This protocol assumes that the syntactic complexity of items increases with each modifier or adjunct (optional complement) that is added to the structure. Thus, an item with one obligatory object and two optional ones is assigned a rating of 3, and so on.

The two raters exhibited an adequate level of agreement for this index ($r_{xy} = .79$). A composite index of syntax was composed by averaging the ratings of the two experts ($M = 2.97$; $SD = 1.20$). The five items with the highest and lowest syntax scores are displayed in Table 1.

Double-barreledness. The psychologists were asked to rate whether each item represented one or more than one behavior. A dichotomous response scale was used for each item. The interrater agreement for this index, measured by $r^{*}_{WG(J)}$ (Lindell, Brandt, & Whitney, 1999), was .61. Forty-six percent of the items were judged by all nine raters to be unidimensional. A composite double-barreledness index was computed by aggregating the ratings of the nine raters ($M = 1.25$; $SD = .32$). Examples of items judged to be unidimensional and multi-dimensional are displayed in Table 2.

Table 1

Sample items
Five items with the highest average ratings on syntax

Mean	*Item*
7.5	Uses good timing and common sense in negotiating; makes his/her points when the time is ripe and does it diplomatically.
7	Focuses more on managing other people to accomplish a task than on personally finishing everything the work group does.
6.5	Works effectiveley with higher management (e.g., presents to them, persuades them, and stands up to them if necessary).
6	Is able to fire or deal firmly with loyal but incompetent people without procrastinating.
6	When working with peers from other functions or units, gains their cooperation and support.

Five items with the lowest average ratings on syntax

Mean	*Item*
1	Recognizes the kind of challenges he/she likes most.
1	Is action-oriented.
1	Supports team decisions.
1	Celebrates good tries.
1	Pursues continuous learning.

Table 2

Sample items
A sample of five items judged to be unidimensional

Item
Is widely counted on by peers.
Is willing to help an employee with personal problems.
Expresses ideas which are well-organized and easily understood by others.
Develops original, creative, innovative approaches to work situations.
Works effectively with higher management (e.g., presents to them, persuade them, and stands up to them if necessary).

A sample of five items judged to be multi-dimensional

Item
Is able to fire or deal firmly with loyal but incompetent people without procrastinating.
Acts fairly and does not play favorite.
Makes timely decisions in situations of uncertainty or limited information.
Presents a consistently positive image to employees at all times.
Has a warm personality that puts people at ease.

Specificity (linguistic). The linguistic operationalization of specificity was obtained by analyzing each item constituents identified by the syntactic analysis. This measurement process was based on widely accepted views on the semantic specificity of nominal expressions (e.g., Chomsky, 1965). Each constituent was assessed in term of its utilization of certain linguistic elements. The specificity index for each item represents the sum of these ratings for all of the constituents that composed the performance items. The linguistic elements that were used were, in order: generic nouns, weakly quantified objects, strongly quantified objects, personal objects, proper nouns, and sequential complements or objects. Generic nominal expressions (rating of 1) do not refer to a particular individual, but to a kind or type of individuals, or to some abstract entity. In "…celebrates successes" for example, no specification is given as to the type of successes, how they are celebrated and so forth. Weakly quantified nominals (rating of 2) are nominals preceded by indefinite determiners (e.g., some, many, others). For example, the typical example of an indefinite determiner is 'a(n)' as in 'a manager'. Strongly quantified nominals (rating of 3) refer

Table 3

Sample items

Five items with the highest average ratings on specificity (linguistic)

Mean	*Item*
25	Is able to build work and management systems that are self-monitoring and can be managed effectively by remote control.
23	Is able to transfer principles and knowledge, such as that from coursework and seminars, to the job at hand.
22	Acts decisively when faced with a tough decision such as laying off workers, even though it hurts him/her personally.
21	Supports a working environment which values a broad range of experiences, backgrounds, and points of view.
20.5	Pushes decision making down to the lowest level.

Five items with the lowest average ratings on specificity (linguistic)

Mean	*Item*
1	Is action-oriented.
1	Encourages balance in personal and work life.
1	Celebrates success.
1	Follows through on commitments.
1	Can lead and let others lead.

to those nominals preceded by a more direct referent (e.g., the, each, every, most, or all). For example, "this manager treats *all coworkers* with dignity and respect" points exhaustively to a well defined set of individuals. A rating of 4 was attributed to personal pronouns (e.g., me, you, he, she, or it), demonstratives (e.g., this, that, these, or those), and possessives (e.g., my, your, his, or her) are even more specific than strongly quantified nominal expressions (e.g., this manager encourages *us* to be open to different viewpoints). Finally, sentential complements or subjects (rating of 5) represented the highest degree of semantics allowed and was attributed to items describing complete situations—events, state, and processes. For example, "this manager demonstrates a lack of urgency *in meeting customer expectations*".

The two raters exhibited a high level of agreement for this index ($r_{xy} = .98$). A composite specificity index was composed by averaging the ratings of the two experts ($M = 7.38$; $SD = 4.67$). The five items with the highest and lowest linguistic specificity scores are displayed in Table 3.

Specificity (psychology). The psychological operationalization of specificity required asking the nine psychologists to judge the behavioral specificity of each item. After reading a brief definition of specificity, the behavioral experts rated each item on a 5-point scale (1 = not specific at all, 2 = slightly specific, 3 = moderately specific, 4 = quite specific, 5 = extremely specific). The interrater agreement for this index, measured with $r^*_{WG(J)}$ (Lindell, Brandt, & Whitney, 1999), was .64. A composite specificity index was computed by aggregating the ratings of the nine raters ($M = 3.23$; $SD = .53$). The five items with the highest and lowest specificity scores are displayed in Table 4.

Item quality. Psychometric quality was operationalized as the relationship between an item and the performance dimension it was intended to measure. The most logical operationalization of an item's quality is its loading on its intended factor. However, using factor loadings as our dependent variable would have resulted in a scale that violated the statistical requirement for equal intervals. This is because the same unit difference between low factor loadings and high factor loadings translates into different amounts of variance accounted for by the factor. For example, factor loadings of .10 and .20, when squared, reveal that a factor accounts for 1% and 4% of the items' variance, respectively. However, loadings of .80 and .90 translate into 64% and 81% of the items' variance. Even though a .10 difference in the magnitude of the loadings is involved in each comparison, it translates into a difference of 3% of the variance in the first case and a difference of 17% in the second.

For each instrument, we estimated one confirmatory factor model for each rating source (self, peer, subordinate, and supervisor) in the sample that provided data for that instrument. We specified these models by identifying how the items on each instrument are allocated to performance dimensions. Three measures of psychometric quality were computed: R, factor loading, and uniqueness. The three indices were highly correlated and the results of the subsequent analyses were identical across all three. For the sake of brevity we only report R, or the amount of variance in an item that was accounted for by the factor it

Table 4

Sample items

Five items with the highest average ratings on specificity (psychological)

Mean	*Item*
4.4	Has solid working relationships with higher management.
4.3	Once the more glaring problems in an assessment are solved, can see the underlying problems and patterns that were bscured before.
4.2	Makes timely decisions in situations of uncertainty or limited information.
4.2	Would become cynical toward the organization during hard times.
4.2	Effectively presents "bad" news to executives.

Five items with the lowest average ratings on specificity (psychological)

Mean	*Item*
2	Does an honest self-assessment.
2	Supports efforts to make [organization] a world class organization.
2.1	Takes charge of his/her career.
2.1	Is widely counted on by peers.
2.1	Seeks corrective feedback to improve himself/herself.

was intended to measure as the main dependent variable in this study. We obtained R values for each item by estimating confirmatory factor models for each of the three MSA instruments. Table 5 provides a summary of the dimensions measured by each MSA instrument, as well as the number of items assessing each dimension. Using this table, one can see the basic structure of the confirmatory factor models that we estimated for each instrument. For example, we estimated a model containing eight latent performance factors for instrument C. The number of items loading on these factors ranged from 3 to 8 for each MSA instrument, the same confirmatory factor model was estimated on each group of raters. The models differed only across the three instruments.

For each confirmatory factor model, we set the scale of the latent factors by fixing the loading of one item on each factor to unity. We allowed the latent factors to intercorrelate. Finally, we conducted these analyses on the covariance matrix for each sample. We used the EQS (Bentler, 1995) structural modeling program to conduct our confirmatory factor analyses.

The intercorrelations among the dimensions of the three instruments for the four sources are presented in Table 6 and 7 while the results of the confirmatory factor analyses for each instrument by rating source are presented in Table 8. We present several fit indices in the table. The normed fit index (NFI; Bentler & Bonnet, 1980) and the comparative fit index (CFI; Bentler, 1990) provide information about the proportion in fit achieved by a model relative to a null model (in which all observed variables are treated as independent). The Tucker-Lewis index (TLI; Tucker & Lewis, 1973) has the same interpretation, but it contains an adjustment for model complexity. Values of these three indices that are greater than or equal to .90 typically are interpreted as representing a good fit to the data. We also present the root mean square error of approximation (RMSEA; Browne & Cudeck, 1993). RMSEA is a measure of lack of fit per degree of freedom for a model. Values between 0 and .05 can be interpreted as reflecting a close fit of the model in the population, whereas values of about .08 or less reflect reasonable fit (Browne & Cudeck, 1993).

The fit indexes presented in Table 8 suggest that the a priori models for each instrument provided a reasonable fit to the data. While some of the fit indexes fell below conventional cut-offs, at least one of the indexes for each model points to the conclusion that the model does a reasonable job of representing the data.

Results

Intercorrelations among the main variables of interest in the study are presented in Table 9. The relationship between the two measures of specificity was not significant indicating a lack of convergence between the linguistic and psychological frameworks. This was expected as the two operationalizations of specificity stemmed from different disciplines. Because the items originated from three different instruments, a MANOVA was performed to investigate test effects for the four indices. An overall effect was found; $F\ (2,\ 366) = 5.19$; $p < .001$. Significant differences between the three tests were found for the linguistic indices; syntax [$F(2, 185) = 17.07$; $p < .001$] and behavioral specificity (linguistics) [$F(2, 185) = 11.04$; $p < .001$]. Sheffe's post-hoc comparisons ($p < .01$) indicated that instrument A was higher on syntax than instrument B and C. For behavioral specificity (linguistics), instrument A was found to be higher than instrument C.

Hierarchical regression analysis (Baron & Kenny, 1986) was used to determine the incremental variance in R explained by item characteristics, over and above the mean

Table 5. Dimensions and number of items for each MSA instrument

	Performance Dimensions	# Items
Instrument A	Resourcefulness	17
	Doing whatever it takes	14
	Being a quick study	4
	Decisiveness	4
	Leading employees	13
	Setting a developmental climate	5
	Confronting problem employees	4
	Workteam orientation	4
	Hiring talented staff	3
	Building and mending relationships	11
	Compassion and sensitivity	4
	Straightforwardness and composure	6
	Balance between life and work	4
	Self-awareness	4
	Putting people at ease	4
	Acting with flexibility	5
	Total	106
Instrument B	Cooperation	4
	Organizational Support	4
	Valuing Diversity	4
	Innovation	6
	Leadership	6
	Integrity	6
	Flexibility	4
	Communication	4
	Total	38
Instrument C	Ethical Behavior	8
	Customer Focus	5
	Business Understanding	4
	Cooperation	3
	Problem Solving	8
	Responsibility	5
	Motivating Others	7
	Openness	4
	Total	44

differences found across the three instruments. Dummy codes were assigned to the three instruments; this variable was entered first in the equation, followed by all four item characteristics. The results are shown in Table 10. The array of item characteristics was found to increase the explained variance in R for self, subordinates, peers, and supervisors.

Syntax was found to be a significant predictor of the psychometric quality of performance items. For self, peer, and supervisor ratings, a negative relationship was found between syntax and R. In other words, the addition of syntactic components to performance items resulted in a decrease in the items' R values. These results lend partial support to hypothesis 1a. Double-barreledness and both specificity indices were not significantly related to item R values for any of the rating sources. Thus, Hypothesis 1b and 1c were not supported.

The second set of hypotheses pertained to an interaction between source of ratings and the psychometric properties of the items. More specifically, stronger relationships between item characteristics and the psychometric properties of the same item were expected for self, peers, and subordinates than for supervisors. The results show that this effect did not occur. The regressions weights for the supervisor ratings do not differ significantly from those of the other sources.

Discussion

This study was aimed at investigating the effects of item characteristics on the psychometric properties of MSA items when used by various rating sources. Despite the substantial advances made in performance appraisal research, prescriptions pertaining to the design of performance items have been quite consistent through the years: make them concise, specific, and target only one behavior. However, the increasing popularity of MSA has brought fundamental changes in the manner by which performance items are used. The presentation of these items to an increasingly wide range of raters raises some questions as to the influence of item characteristics on their responses.

The need to investigate the impact of item characteristics is further motivated by the surprising variation found in the three instruments used in this study. Some items were found to be very concise—the shortest used three words—while others were quite verbose. Wide variations were uncovered in terms of the specificity and the number of behaviors targeted. This variation is certainly not unique to the instruments used in the study. In their survey of MSF instruments, Leslie and Fleenor (1998) report sample items from 24 instruments with similar variability in item characteristics.

Our results indicate that the syntactic properties of performance items influence ratings for all sources except subordinates. Long items did not function as well as short ones for self, peers, and supervisors. As stated earlier, accurate responses to performance items are contingent upon the understanding of the item as intended by the item designer. Less variations in this understanding takes place for simple items. This result poses a dilemma for item designers that need to describe more complex or subtle behaviors. As noted by an anonymous reviewer, the fact that concise items are more likely to be processed accurately clashes with the idea of providing more

Table 6. Correlation tables for instrument A

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	–	.92	.77	.69	.85	.85	.71	.60	.72	.84	.64	.58	.19	.78	.49	.87
2	.89	–	.77	.86	.77	.86	.76	.51	.71	.72	.58	.49	.09	.71	.44	.81
3	.78	.72	–	.59	.56	.61	.50	.30	.53	.55	.39	.39	.10	.52	.29	.59
4	.62	.53	.55	–	.51	.63	.75	.35	.53	.45	.58	.28	.00	.45	.20	.55
5	.80	.60	.47	.43	–	.96	.65	.79	.73	.88	.86	.63	.31	.86	.67	.95
6	.80	.84	.56	.61	.94	–	.68	.71	.80	.79	.82	.56	.20	.80	.57	.91
7	.64	.70	.43	.69	.61	.65	–	.49	.64	.51	.40	.38	.11	.55	.24	.62
8	.58	.45	.26	.28	.81	.71	.51	–	.63	.59	.60	.41	.40	.59	.45	.68
9	.69	.66	.49	.48	.71	.79	.58	.60	–	.60	.57	.41	.40	.59	.39	.68
10	.77	.59	.43	.29	.83	.67	.40	.57	.50	–	.82	.66	.27	.86	.77	.94
11	.54	.48	.29	.22	.81	.74	.34	.55	.52	.75	–	.57	.35	.79	.80	.87
12	.53	.43	.31	.20	.55	.45	.31	.36	.33	.61	.45	–	.32	.67	.49	.67
13	.18	.04	.09	.03	.34	.18	.14	.46	.22	.30	.32	.32	–	.25	.34	.30
14	.77	.69	.48	.43	.79	.72	.52	.53	.54	.77	.66	.61	.25	–	.65	.90
15	.48	.40	.25	.17	.63	.52	.21	.41	.41	.76	.74	.44	.28	.56	–	.73
16	.87	.80	.53	.52	.95	.88	.63	.69	.68	.89	.78	.59	.30	.85	.60	–

Correlations above the diagonals were computed from supervisor ratings.
Correlations below the diagonals were computed from peer ratings.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	–	.91	.67	.62	.79	.81	.61	.42	.65	.76	.55	.42	.16	.68	.35	.87
2	.91	–	.64	.79	.73	.83	.65	.38	.62	.63	.51	.31	.06	.65	.32	.79
3	.67	.64	–	.48	.43	.47	.29	.12	.38	.45	.30	.22	.12	.43	.17	.50
4	.62	.79	.48	–	.46	.58	.63	.30	.43	.33	.17	.20	.04	.37	.07	.51
5	.79	.73	.43	.46	–	.94	.66	.69	.68	.79	.79	.49	.25	.73	.54	.91
6	.81	.83	.47	.58	.94	–	.58	.57	.75	.65	.71	.43	.12	.67	.40	.87
7	.61	.65	.29	.63	.66	.58	–	.46	.46	.42	.32	.34	.10	.41	.17	.59
8	.42	.38	.12	.30	.69	.57	.46	–	.57	.37	.40	.34	.26	.36	.24	.52
9	.65	.62	.38	.43	.68	.75	.46	.57	–	.46	.56	.26	.13	.47	.30	.61
10	.76	.63	.45	.33	.79	.65	.42	.37	.46	–	.74	.47	.24	.68	.69	.87
11	.55	.51	.30	.17	.19	.71	.32	.40	.56	.74	–	.38	.26	.68	.73	.79
12	.42	.37	.22	.20	.49	.43	.34	.34	.26	.47	.38	–	.28	.43	.26	.49
13	.16	.06	.12	.04	.25	.12	.10	.26	.13	.24	.26	.28	–	.19	.24	.24
14	.68	.65	.43	.37	.73	.67	.41	.36	.47	.68	.68	.43	.20	–	.41	.77
15	.35	.32	.17	.07	.54	.40	.17	.24	.30	.69	.73	.26	.24	.45	–	.58
16	.87	.79	.51	.51	.91	.87	.59	.52	.61	.87	.79	.49	.24	.77	.58	–

Correlations above the diagonals were computed from subordinate ratings.
Correlations below the diagonals were computed from self ratings.

information enhances accuracy by providing a common frame of reference. Our results indicate that conciseness has more influence on item quality then specificity. It may be useful to pair the use of complex items with careful instructions and enhanced accountability mechanisms in order to mitigate the influence of complexity.

However, it is also important to interpret these results in light of the broader utility of performance items. First, in addition to guiding raters in the evaluation process, performance items also serve as guide to the ratee when used in a developmental context. When used as feedback material, the use of specific items is advantageous as they provide helpful guidance for individual development (Kluger & DeNisi, 1995; Murphy & Cleveland, 1995). Secondly, the effect sizes uncovered were small. A large amount of variance in R is left unaccounted for. Contextual factors such as the purpose of the appraisal, for example, have been shown to greatly influence the properties of performance ratings (Jawahar & Williams, 1997). The fact is that a wealth of factors influence the quality of items and

Table 7. Correlation tables for instrument B and C

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
1	–	.69	.88	.75	.87	.89	.77	.85
2	.60	–	.69	.77	.70	.70	.78	.67
3	.87	.65	–	.73	.87	.90	.75	.85
4	.74	.66	.73	–	.88	.83	.88	.81
5	.86	.68	.86	.93	–	.96	.84	.95
6	.92	.65	.89	.84	.97	–	.85	.96
7	.75	.74	.72	.89	.84	.81	–	.81
8	.79	.62	.81	.83	.92	.90	.76	–

Correlations above the diagonals were computed from supervisor ratings.
Correlations below the diagonals were computed from peer ratings.

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
1	–	.64	.85	.72	.86	.89	.75	.83
2	.78	–	.66	.74	.63	.63	.77	.58
3	.95	.76	–	.77	.88	.88	.69	.88
4	.88	.82	.86	–	.91	.84	.82	.78
5	.94	.79	.92	.95	–	.96	.82	.96
6	.93	.76	.92	.91	.97	–	.81	.96
7	.85	.85	.82	.92	.89	.89	–	.75
8	.90	.73	.88	.87	.95	.96	.86	–

Correlations above the diagonals were computed from subordinate ratings.
Correlations below the diagonals were computed from self ratings.

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
1	–	.78	.52	.92	.78	.91	.77	.86
2	.60	–	.70	.88	.96	.92	.87	.93
3	.38	.58	–	.64	.74	.65	.70	.76
4	.87	.77	.50	–	.91	.93	.91	.97
5	.59	.89	.67	.79	–	.94	.90	.97
6	.78	.88	.56	.87	.91	–	.86	.94
7	.58	.76	.63	.82	.83	.77	–	.99
8	.74	.86	.73	.93	.93	.90	.97	

Correlations above the diagonals were computed from supervisor ratings.
Correlations below the diagonals were computed from peer ratings.

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
1	–	.73	.43	.86	.70	.88	.63	.81
2	.79	–	.56	.79	.90	.83	.75	.84
3	.60	.73	–	.49	.60	.44	.56	.61
4	.92	.89	.67	–	.81	.85	.82	.92
5	.83	.93	.72	.94	–	.84	.81	.91
6	.94	.89	.70	.97	.95	–	.69	.87
7	.81	.87	.67	.95	.92	.91	–	.99
8	.91	.90	.74	.99	.94	.97	.99	–

Correlations above the diagonals were computed from subordinate ratings.
Correlations below the diagonals were computed from self ratings.

Table 8. Fit of a priori measurement models for each of the three MSA instruments by rating source

Instrument/ Source	Chi-square	df	NFI	CFI	RMSEA
Instrument A					
Self	23,355.12	5,339	.74	.79	.04
Subordinate	57,174.32	5,339	.85	.86	.04
Peer	70,469.02	5,339	.84	.85	.04
Supervisor	24,487.68	5,339	.80	.83	.04
Instrument B					
Self	1,326.08	637	.80	.89	.06
Subordinate	4,961.96	637	.94	.95	.06
Peer	4,171.46	637	.90	.91	.07
Supervisor	1,677.76	637	.83	.89	.07
Instrument C					
Self	4,443.95	874	.84	.86	.05
Subordinate	4,686.95	874	.90	.92	.06
Peer	16,386.46	874	.91	.91	.06
Supervisor	5,745.09	874	.87	.89	.06

Note: All chi-square values are statistically significant ($p < .05$). NFI = normed fit index, CFI = comparative fit index, RMSEA = root mean square error of approximation.

any prescriptions pertaining to the design of items should be considered within the broader context of the appraisal.

Interestingly, the other syntactic index, double-barreledness, was not found to affect ratings. It is difficult to argue that an item targeted at two behaviors will perform as well as an item that targets a single one. However, a key element for double-barreledness, one not tapped by this study, is the degree of convergence between the behaviors targeted. For example, the discrepancy between the behaviors targeted by "is action-oriented and delegates to subordinates" is larger than that of "is action-oriented and opportunistic". In the sample, it is probable that the double-barreled items had both barrels pointed in the same direction! Still, taken at face value, this finding indicates that double-barreled items are not inherently poor.

Limitations

The main limitation of this study is that the utilization of existing instruments may have suppressed the results. In the development of these instruments, item selection was performed resulting in a reduction in the variance of the psychometric indices. For example, truly double-barreled items were not present in our sample. Hence, our results are probably a conservative estimate of the effects of item characteristics. Another related limitation is our reliance on a strict psychometric assessment of item quality. As noted by an anonymous reviewer, the best strategy for evaluating an item is to bear in mind its statistical and psychometric properties in combination with the importance of the content the item addresses. Although we chose to focus on psychometric criteria for evaluating an item, conceptual and content related considerations are important in determining the quality of an item.

Also, our hypotheses were based on assumed characteristics of the different rating sources (i.e., variations in expertise and accountability). In terms of experience with performance appraisal for example, it is likely that many peers and subordinates participating in this study had supervisory duties and had to rate their own subordinates at one point. A direct measurement of these factors would provide a better test of our hypotheses. On a related point, other individual differences probably play a major role on the influence of item design characteristics. Scullen, Goff, and Mount (2000) demonstrated that rater idiosyncratic tendencies accounted for most of the variance in ratings. As noted by an anonymous reviewer, raters must certainly vary in their capacity to decipher and interpret syntactically complex items. Perhaps the effects could vary depending on the job complexity levels on the assumption that people will gravitate to jobs with cognitive levels similar to their own. Finally, our reliance on ratings obtained for

Table 9. Correlations between the main variables

	N	Mean	SD	1	2	3	4	5	6	7	8
1. R self	188	.61	.11	–							
2. R peer	188	.72	.09	.89**	–						
3. R subordinate	188	.73	.11	.85**	.94**	–					
4. R supervisor	188	.70	.09	.89**	.90**	.82**	–				
5. Behavioral specificity (p)	195	3.23	.53	−.03	−.01	−.05	−.01	(.64)			
6. Double barreledness	195	1.25	.32	−.07	−.07	−.08	−.03	.26**	(.61)		
7. Syntax	195	2.97	1.30	−.30**	−.30**	−.31**	−.24	−.02	−.03	(.79)	
8. Behavioral specificity (I)	195	7.38	4.67	−.20**	−.23*	−.28*	−.16	−.04	−.03	.61**	(.98)

Note: $r^{*}_{WG(J)}$ is used for the reliability of behavioral specificity (p) and double-barreledness is estimated while a correlation coefficient is used for syntax and behavioral specificity (I).

Table 10. Results of the regression analyses for each rating group

	Self ratings			Subordinate ratings			Peer ratings			Supervisor ratings		
	R^2	ΔR^2	Beta	R^2	ΔR^2	Beta	R^2	ΔR^2	Beta	R^2	ΔR^2	Beta
Instruments												
	.12	.12**		.24	.24**		.15	.15**		.05	.05**	
Syntax			−.20**			−.09			−.16**			−.18**
Double-b			.04			−.05			−.04			.05
Specificity (I)			.01			−.09			−.04			.01
Specificity (p)			−.08			−.10			−.04			−.05
	.16	.04**		.28	.04**		.18	.03**		.09	.04**	

Notes: N = 188; **$p < .01$; *$p < .05$

developmental purposes limits generalization to ratings obtained for other purposes.

Conclusion

MSA designers are confronted with the following paradox: although the unique information contributed by each rater adds validity to the assessment of individual performance that same information also implies variation in its psychometric properties which, in turn, often prevents meaningful comparisons across raters. Item design characteristics represent a simple way to tackle part of this problem. Yet, despite widespread knowledge that behavior-based performance item should follow these prescriptions, a close look at off-the-shelf MSA instruments reveals much variation in item characteristics. This study offers empirical support for at least one of these prescriptions: keep it short. However, this study represents only a first step. Additional research on item characteristics is likely to lead to a greater understanding of the effects of item characteristics on performance assessment.

References

Aiken, L.R. (1997) *Rating scales and checklists: Evaluating behaviors, personality, and attitudes*. New York: Wiley.

Antonioni, D. and Woehr, D.J. (2001) Improving the quality of multisource rater performance. In D.W. Bracken, C.W. Timmreck and A.H. Church (Eds.). *The Handbook of Multisource Feedback*. San Francisco: Jossey-Bass.

Austin, J.T. and Villanova, P. (1992) The Criterion Problem: 1917–1992. *Journal of Applied Psychology*, **77**, 836–875.

Baron, R.M. and Kenny, D.A. (1986) The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, **54**, 1173–1182.

Bentler, P.M. (1995) *EQS structural equations program manual*. Encino, CA: Multivariate Software, Inc.

Bentler, P.M. (1990) Comparative fit indexes in structural models. *Psychological Bulletin*, **107**, 238–246.

Bentler, P.M. and Bonnet, D.G. (1980) Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, **88**, 588–606.

Borman, W.C. (1997) 360 ratings: An analysis of assumptions and a research agenda for evaluating their validity. *Human Resource Management Review*, **7**, 299–315.

Browne, M.W. and Cudeck, R. (1993) Alternative ways of assessing model fit. In K.A. Bollen and J.S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Brutus, S. and Derayeh, M. (2002) Multisource assessment programs in organizations: An insider's perspective. *Human Resource Development Quarterly*, **13**, 187–203.

Chomsky, N. (1965) *Aspects of the Theory of Syntax*. Cambridge: MIT Press.

Chomsky, N. (1986) *Knowledge of Language. Its Nature, Origin and Use*. New York: Praeger.

DeNisi, A.S. and Williams, K.J. (1988) A cognitive approach to performance appraisal. In K. Rowlandand and G. Ferris (Eds.), *Research in personnel and human resource management* **Vol. 6** (pp. 109–156).

Facteau, J.D. and Craig, S.B. (2001) Are performance appraisal ratings obtained from different rating sources comparable? *Journal of Applied Psychology*, **86**, 215–227.

Feldman, J.M. (1981) Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology*, **66**, 127–148.

Feldman, J.M. (1986) Instrumentation and training for performance appraisal: A perceptual-cognitive viewpoint. In K. Rowlands and G. Ferris (Eds.), *Research in personnel and human resource management* **Vol. 4** (pp. 148–216).

Harrison, D., McLaughlin, M.E. and Coalter, T.M. (1995) Do context effects really matter?: Psychometric and cognitive evidence in organizational justice perceptions. *Academy of Management Proceedings*, **1995**, 375–379.

Jawahar, I.M. and Williams, C.R. (1997) Where all the children are above average: The performance appraisal purpose effect. *Personnel Psychology*, **50**, 905–925.

Kavanaugh, M.J. (1971) The content issue in performance appraisal: A review. *Personnel Psychology*, **24**, 653–668.

Kluger, A.N. and DeNisi, A. (1996) Effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theroy. *Psychological Bulletin*, **119**, 254–284.

Lance, C.E. and Bennett, W. (1997). *Rater source differences in cognitive representations of performance information*. Paper presented at the annual meeting of the Society for Industrial and organizational Psychology, St. Louis, MO.

Leslie, J.B. and Fleenor, J.W. (1998) *Feedback to managers: A review and comparison of multi-rater instruments for management development (3rd ed)*. Greensboro, NC: Center for Creative Leadership.

Lindell, M.K., Brandt, C.J. and Whitney, D.J. (1999) A revised index of interrater agreement for multi-item ratings of a single target. *Applied Psychological Measurement*, **23**, 127–135.

London, M. and Smither, J.W. (1995) Can multi-source feedback change self-evaluations, skill development, and performance? Theory-based applications and directions for research. *Personnel Psychology*, **48**, 803–839.

London, M., Smither, J.W. and Adsit, D.J. (1997) Accountability: The Achilles' heel of multisource feedback. *Group and Organization Management*, **22**, (2), 162–184.

Maurer, T.J., Raju, N.S. and Collins, WC. (1998) Peer and subordinate performance appraisal measurement equivalence. *Journal of Applied Psychology*, **83**, 693–702.

Murphy, K.R. and Cleveland, J.N. (1995) *Understanding performance appraisal*. Thousand Oaks, CA: Sage Publications.

Murphy, K.R., Cleveland, J.N. and Mohler, C.J. (2001) Reliability, validity, and meaningfulness of multisource ratings. In D.W. Bracken, C.W. Timmereck and A.H. Church (Eds.). *The Handbook of Multisource Feedback*. San Francisco: Jossey-Bass.

Robert, F. (1998) Structural Complexity and the Time Course of Grammatical Development. *Cognition*, **66**, 249–301.

Rogelberg, S.G. and Waclawski, J. (2001) Instrumentation design. In D.W. Bracken, C.W. Timmereck and A.H. Church (Eds.). *The Handbook of Multisource Feedback*. San Francisco: Jossey-Bass.

Rychlak, J.F., Stilson, S.R. and Rychlak, L.S. (1993) Testing a Predicational Model of Cognition: Cueing Predicate Meanings in Sentences and Word Triplets. *Journal of Psycholinguistic Research*, **22**, 479–503.

Schwarz, N. (1999) Self reports: How the questions shape the answers. *American Psychologist*, **54**, 93–105.

Scullen, S.E., Goff, M. and Mount, M.K. (2000) Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, **85**, 956–970.

Smith, D.E. (1986) Training Programs for Performance Appraisal: A Review. *Academy of Management Review*, **11**, 22–41.

Sudman, S. and Bradburn, N.M. (1982) *Asking questions*, San Francisco: Jossey Bass.

Tucker, L.R and Lewis, C. (1973) A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, **38**, 1–10.

VanVelsor, E. and Leslie, J.B. (2001) Selecting a multisource feedback instrument. In D.W. Bracken, C.W. Timmereck and A.H. Church (Eds.). *The Handbook of Multisource Feedback*. San Francisco: Jossey-Bass.

Villanova, P. and Bernardin, H.J. (1991) Performance appraisal: The means, motive, and opportunity to manage impressions. In R.A. Giacalone and P. Rosenfeld (Eds.), *Applied impression management: How image-making affetcs managerial decisions* (pp. 81–96). Thousand Oaks, CA: Sage.

Waldman, D.A. and Atwater, L.E. (1998) *The power of 360-degree feedback*. Houston, TX: Gulf Publishing.