Journal of Applied Psychology
2007, Vol. 92, No. 3, 812–819

0021-9010/07/$12.00 DOI: 10.1037/0021-9010.92.3.812

Can Training Improve the Quality of Inferences Made by Raters in Competency Modeling? A Quasi-Experiment

Filip Lievens
Ghent University

Juan I. Sanchez
Florida International University

A quasi-experiment was conducted to investigate the effects of frame-of-reference training on the quality of competency modeling ratings made by consultants. Human resources consultants from a large consulting firm were randomly assigned to either a training or a control condition. The discriminant validity, interrater reliability, and accuracy of the competency ratings were significantly higher in the training group than in the control group. Further, the discriminant validity and interrater reliability of competency inferences were highest among an additional group of trained consultants who also had competency modeling experience. Together, these results suggest that procedural interventions such as rater training can significantly enhance the quality of competency modeling.

Keywords: competency modeling, frame-of-reference training, job analysis, rating

In recent years, competency modeling has emerged as an alternative to traditional job analysis. Broadly defined, the primary aim of competency modeling is the identification of a set of core competencies required for successful performance across some or all jobs in the organization; these competencies in turn become the cornerstone of subsequent human resources (HR) practices. One of the key differences between competency modeling and traditional job analysis might be that competency modeling goes beyond the rigid boundaries of a job title by taking into account the organization's objectives, vision, and strategy in the formulation of staffing requirements. Traditional job analysis, in contrast, has largely neglected the role of these macro variables (Sanchez, 1994; Snow & Snell, 1992).

Despite the potential merits of competency modeling, Schippmann et al. (2000) warned that some competency modeling approaches might lack the methodological rigor inherent in traditional job-analytic techniques. Indeed, the few studies that have scrutinized the quality of the broad inferences made by raters in competency modeling have revealed troubling evidence with regard to the psychometric properties of competency ratings (Lievens, Sanchez, & De Corte, 2004; Morgeson, Delaney-Klinger, Mayfield, Ferrara, & Campion, 2004).

The possible lack of rigor in competency modeling has potentially serious practical and legal implications. For example, the most recent version of the *Principles for the Validation and Use of Personnel Selection Procedures* states that "any methods used to obtain information about work or workers should have reasonable psychometric characteristics . . . Lack of consensus about the information contained in the analysis of work should be noted and considered further" (Society for Industrial and Organizational Psychology, 2003, p. 11). Therefore, it is of key importance to examine procedural interventions that might enhance the quality of the inferences made by raters in competency modeling.

Recent research has investigated the beneficial effects of two such procedural factors, namely, using a variety of job experts and providing task information (Lievens et al., 2004). The purpose of this study was to investigate the effectiveness of still another potentially useful intervention, namely, providing training to those involved in making competency ratings. Specifically, we borrowed from prior research on rater training to develop a theoretically sound training program aimed at improving the quality of competency ratings. Quality was broadly operationalized in terms of interrater reliability, discriminant validity, and accuracy. As argued by Dierdorff and Wilson (2003) and Morgeson and Campion (1997), these are important criteria, as they reflect underlying issues of reliability and validity of work analysis data.

The Quality of Competency Ratings

To improve work-analytic outcomes, Sanchez and Levine (1994) noted that two primary components of the work-analytic process could be altered: the rating stimuli on which judgments are made and the rater's judgment skills. In regard to rating stimuli, Morgeson and Campion (2000) distinguished between direct and indirect methods of estimating knowledge, skill, and ability (KSA) requirements (see also Gatewood & Feild, 2001, pp. 367–380). Indirect estimation methods break down the complex inferential leap involved in specifying KSA requirements for the job into a series of more manageable steps, namely, identifying job tasks, judging the importance or criticality of these tasks, and making inferences about which KSAs are most important. The methodological rigor of this step-by-step approach lends credence to the final set of subject-matter experts' (SMEs) KSA ratings. Conversely, in direct estimation methods, SMEs are asked to directly

Filip Lievens, Department of Personnel Management and Work and Organizational Psychology, Ghent University, Ghent, Belgium; Juan I. Sanchez, Department of Management and International Business, Florida International University.

A previous version of this article was presented at the August 2006 annual meeting of the Academy of Management, Atlanta, GA.

Correspondence concerning this article should be addressed to Filip Lievens, Department of Personnel Management and Work and Organizational Psychology, Ghent University, Henri Dunantlaan 2, 9000, Ghent, Belgium. E-mail: filip.lievens@ugent.be

rate the importance of various KSAs for a given job, skipping the potential benefits of going through these intermediate steps. Clearly, this kind of holistic KSA judgment calls for a larger inferential leap than does the indirect estimation method (Sanchez & Levine, 2001; Schippmann et al., 2000). Lievens et al. (2004) argued that competency modeling falls in the direct estimation category because competencies are typically inferred from broad job descriptions as well as information about the organization's strategy.

We located only two studies that examined the quality of ratings made in competency modeling as a function of variations in the rating stimuli. Morgeson et al. (2004) reported that global judgments similar to those made in competency modeling were inflated as compared with task-level judgments. Lievens et al. (2004) found poor interrater reliability and poor discriminant validity among competency ratings made by inexperienced raters; nevertheless, the quality of competency ratings was higher among job experts, especially when competency modeling inferences were informed by task-related information.

In short, a conclusion that can be drawn from prior studies is that competency modeling ratings should not be taken for granted because they often have poor psychometric characteristics as a result of the rather large inferential leap required. From a practical standpoint, there appears to be a need to increase the rigor of current competency modeling efforts. However, the few studies that have examined potential improvements to the competency modeling process have focused on manipulations of the rating stimuli or the rating process. As Sanchez and Levine (1994) pointed out, interventions to improve the judgment skills of those in charge of rating the competencies should also be explored. One such intervention is rater training.

Enhancing the Quality of Competency Ratings Through Training

In the performance appraisal domain, frame-of-reference (FOR) training (Bernardin & Buckley, 1981) has emerged as the most promising rater training approach (Hauenstein, 1998; Woehr & Huffcutt, 1994). The general aim of FOR training is "to prime raters' use of an organizational category system for observing behavior and rating performance" (Bernardin, Buckley, Tyler, & Wiese, 2000, p. 227). FOR training intends to reduce idiosyncratic rating tendencies by establishing a framework for observing and evaluating performance. Such a framework carefully defines performance dimensions and provides a sample of behavioral incidents of these dimensions (together with their performance levels) through practice and feedback. The expectation is that raters will apply to their field ratings the behaviorally based framework instilled during training.

From a theoretical point of view, FOR training has been linked to schema-based theory (Cardy & Keefe, 1994; Day & Sulsky, 1995; Sulsky & Day, 1994). Specifically, it has been argued that FOR training instills in raters more appropriate schemata than their preexisting mental frameworks. Dorfman (1982) referred to these preexisting expectations regarding a particular job as a role schema. Although the use of such organizing schemata has advantages in terms of information processing efficiency, there are also drawbacks. For example, in the context of work analysis, role schemata have been found to direct rater attention consistent information only (Jacobs, Kulik, & Fichman, 1993; Kulik, 1989). In addition, role schemata might increase the overlap between competencies in raters' judgments because raters may not innately perceive their jobs as varying along a number of dimensions (Morgeson & Campion, 1997).

We were able to identify only two studies of rater training effects in the job analysis literature (i.e., Hahn & Dipboye, 1988; Sanchez & Levine, 1994). Although both studies found promising results, neither was anchored on the FOR approach. We believe that FOR training might be beneficial in the context of competency modeling because it attempts to counteract the possible drawbacks of schema-based processing by imposing more appropriate schemata on raters. For instance, FOR-trained raters should be more accurate in discerning essential from nonessential competencies for a given job because FOR training provides raters with a common mental framework regarding the extent to which specific tasks are relevant to competencies. In other words, raters develop a shared understanding of task–competency linkages (Morgeson et al., 2004). Thus, we predicted that the discriminant validity (Hypothesis 1a), interrater reliability (Hypothesis 1b), and accuracy (Hypothesis 1c) of competency ratings would be higher among FOR-trained raters than among raters in the control group.

Further, we expected that experience in competency modeling would augment the positive effects of training by enabling those raters who have already been trained to make even more fine-grained distinctions among competencies. Therefore, the positive effects of FOR training on discriminant validity, interrater reliability, and accuracy of competency ratings should be highest among FOR-trained raters who already have several years of experience in competency modeling. This prediction was based on prior research comparing experts and novices (Chi, Glaser, & Farr, 1988), which suggested that expert raters rely on well established cognitive structures when rating jobs. Specifically, expertise develops by abstracting from education (e.g., a degree in psychology or HR), training (e.g., a competency modeling training program), and experience (e.g., competency modeling experience). Thus, we expected that the shared framework of task–competency linkages acquired during training, combined with continued analysis and observation of jobs, would help these raters establish well rehearsed cognitive structures for each competency. Hence, we predicted that the discriminant validity (Hypothesis 2a), interrater reliability (Hypothesis 2b), and accuracy (Hypothesis 2c) of competency ratings would be highest among FOR-trained raters who already have competency modeling experience.

Method

Sample

The sample consisted of consultants working for a local branch of an international HR service firm. This firm provided a variety of HR services, including recruitment, selection, assessment, development, compensation, and career management. As determining competency modeling was seen as the foundation of these HR services, a training program was developed to increase the quality of competency ratings.

The group of HR consultants designated to participate in the training program in 2004 was randomly assigned to one of two conditions: a training condition and a control condition. None of these consultants had participated previously in competency mod-

eling. The control group consisted of 26 consultants (16 women, 10 men) with a mean age of 27.7 years ($SD = 4.2$) and a mean work experience of 4.6 years ($SD = 4.1$). The training group consisted of 25 consultants (15 women, 10 men) with a mean age of 28.8 years ($SD = 4.1$) and a mean work experience of 5.4 years ($SD = 4.1$). Although the consultants were randomly assigned to conditions, we checked for preexisting differences among the groups. No significant differences were found with regard to sex, age, work experience, or educational background.

Apart from these randomly composed groups, we also included a preexisting group of consultants. This third group of consultants (hereafter referred to as the expert group) had attended the same training program in 2002 or 2003 and had subsequently gained hands-on competency modeling experience. Their experience was verified through various indices (Quiñones, Ford, & Teachout, 1995), namely, the length of their competency modeling experience (between 1 and 5 years), the frequency of their competency modeling experience (between 6 and 25 times), and the diversity of their competency modeling experience (primarily in selection and compensation). This expert group consisted of 22 consultants (5 women, 17 men) with a mean age of 35.2 years ($SD = 9.6$) and a mean work experience of 11.8 years ($SD = 9.1$). Similar to the other groups, all expert consultants had a college degree, mostly in social sciences or business administration.

Competency Modeling Instrument

The specific competency modeling instrument used by the consulting firm consisted of 40 cards, each of which describe a competency (e.g., structuring work) in terms of behaviorally anchored definitions. These 40 competencies were grouped into 5 clusters: information management, leadership, interpersonal management, task management, and personal management. Due to proprietary factors, all competencies per cluster cannot be presented here. Raters used a Q-sort procedure to sort the 40 competencies into 3 rating categories: 1 (*essential*), 2 (*important*), and 3 (*not important*). This specific competency modeling instrument is similar to the portfolio sort cards of the LEADERSHIP ARCHITECT® (Lombardo & Eichinger, 2003), which Lievens et al. (2004) used in their competency modeling study.

Experimental Conditions

As noted above, there were two conditions. In the training condition, an attempt was made to embody the main concepts of FOR training (i.e., emphasizing the multidimensionality of work, defining competencies, providing sample behaviors for each competency, and using practice and feedback to instill the FOR on raters) while at the same time providing opportunities to apply these concepts in a work analysis context.

The training program started by defining competency modeling in the context of HR management. The trainers also discussed the history and applications of competency modeling and compared it with traditional job analysis. Consistent with the FOR training orientation (e.g., Bernardin & Buckley, 1981; Cardy & Keefe, 1994), consultants read a job description and were asked to identify the competencies that were necessary to carry out the job. The ensuing group discussion served to highlight the need to capture the multidimensionality of work. The specific competency modeling framework was introduced by presenting definitions of the various competencies, followed by a discussion of examples of behaviors representing these competencies. This process was repeated for each competency cluster. At the end of this phase, consultants were presented with a "retranslation" exercise (see Woehr, 1994, p. 529), in which consultants assigned behaviors to competencies or assigned competencies to competency clusters. The trainers then discussed the answers and provided feedback to the consultants. Next, the consultants were instructed to base their competency ratings on the tasks performed for the job at hand. In particular, they were asked to study the tasks included in the job description and to link them to the behaviorally based competencies. They were also taught how to use the Q-sort method to assign competencies to jobs. The remainder of the training session was dedicated to practicing the training concepts by rating two jobs (i.e., consultant and HR officer). Consultants received job descriptions and independently rated the competencies. Next, the trainers elicited a discussion of how the consultants determined their ratings and clarified any discrepancies. Finally, the trainers provided the consultants with feedback pertaining to their ratings. The training session lasted one full day. Two trainers were in charge of the delivery: 1 man (age = 30 years, work experience = 7 years) and 1 woman (age = 31 years, work experience = 7 years).

In the control condition, consultants did not attend the FOR training program. However, to prevent their resentful demoralization or any other form of the Hawthorne effect, they received a different type of training. Specifically, they attended a full day training in the STAR (situation-task-action-result) interviewing technique. Again, two trainers delivered this training: 2 men (age = 28 and 32 years, respectively; work experience = 5 and 7 years, respectively).

Competency Modeling Task

The focal competency modeling task asked consultants to determine the competencies required for the job of method engineer by independently assigning a rating to each of the 40 competencies. One week after the training program, consultants were provided with a one-and-a-half page description of the main responsibilities and tasks carried out by a method engineer. In accordance with competency modeling practices (Lievens et al., 2004; Schippmann et al., 2000), background information was also provided about a fictitious company, its history, products, and its business and HR objectives. To ensure realism, this information was retrieved from actual job and company materials. We chose a relatively uncommon job for two reasons. First, it might preclude the activation of job stereotypes (DeNisi, Cornelius, & Blencoe, 1987; Smith & Hakel, 1979). Second, none of the expert consultants had previously determined the competencies of this job.

Results

Overall Generalizability Analyses

To test our hypotheses regarding discriminant validity and interrater reliability, we used generalizability analysis (Brennan, 1992; Cronbach, Gleser, Nanda, & Rajaratnam, 1972). An advantage of generalizability over classical reliability theory is that it allows for the simultaneous estimation of various sources of variance. These so-called variance components capture each facet's

contribution to the total variance. The variance components in our overall generalizability analyses represented the variances of the mean job ratings attributable to competencies (object of measurement), to raters, and to the interactions among competencies and raters. However, estimated variance components are scale dependent, and therefore, we used the percent contribution of each variance component to interpret its relative magnitude (Shavelson & Webb, 1991). The percent contribution refers to the percentage of the sum of the variance components accounted for by each variance component.

Results of the overall generalizability analyses, broken down by condition, are presented in Table 1. Recall that Hypothesis 1a was related to discriminant validity, which was assessed by the variance due to competencies. The variance component due to competencies represents a desirable source of variance because it indicates discriminant validity across competencies. The variance due to competencies in the training group (38%) was twice the size of the variance due to Competencies in the control group (19%), thereby lending support to Hypothesis 1a.

Hypothesis 1b was related to interrater reliability. The Raters × Competencies interaction is related to interrater reliability because it gauges variation in competency ratings across raters. The variance component associated with the Raters × Competencies interaction was larger in the control group (74%) than in the training group (62%). These results provide support for Hypothesis 1b.

The next hypotheses dealt with differences between the training group and the preexisting expert group. The competencies variance component was larger in the expert group (45%) than in the training group (38%). The Raters × Competencies interaction variance component was also smaller in the expert group (54%) than in the training group (62%). These results provide support for Hypotheses 2a and 2b.

An inspection of the generalizability coefficients presented in Table 1 sheds light on the level of interrater reliability across conditions. A generalizability coefficient is an intraclass correlation defined as the ratio of the universe score variance to the expected observed score variance (Brennan, 1992). The highest value was found among the expert group (.95), followed by the trained group (.94) and the control group (.87). However, these coefficients are not directly comparable because they reflect generalizability over different numbers of raters per group (there were 26 raters in the control group, 25 in the training group, and 21 in the expert group). To make a comparison possible, we projected the generalizability coefficient under different numbers of raters (Table 2). Note that reliability estimates can be projected under different measurement conditions in generalizability analysis, thereby enabling prescriptions regarding ideal measurement conditions (Greguras & Robie, 1998). The projected generalizability coefficient exceeded .70 for as few as 3 raters in the expert group. In contrast, 3 raters in the control and training groups produced generalizability coefficients of .44 and .65, respectively. Table 2 also illustrates the practical impact of FOR training in competency modeling, because at least 9 untrained raters are needed to obtain a generalizability coefficient of .70, whereas only 4 trained raters are needed to achieve a similar coefficient (.71). When trained raters already have experience in competency modeling, only 3 of them are needed to achieve a coefficient of .71.

Table 1
Results of Overall Generalizability Analyses Broken Down by Condition

	Control		Trained		Expert	
Effect	VC	%	VC	%	VC	%
Raters	.03	6	.00	1	.01	2
Competencies	.11	19	.21	38	.27	45
Competencies × Raters	.42	74	.35	62	.33	54

Note. Due to rounding, percentages do not sum to 100. G-coefficients were computed on the total number of raters per condition. G-coefficients were .87 for the control group ($N = 26$), .94 for the trained group ($N = 25$), and .95 for the expert group ($N = 21$). VC = variance component.

Table 2
Generalizability Coefficients for Different Numbers of Raters Broken Down by Condition

No. of raters	Control	Trained	Expert
20	.84	.92	.94
15	.80	.90	.93
10	.72	.86	.89
9	.70	.85	.88
8	.68	.83	.87
7	.65	.81	.85
6	.61	.78	.83
5	.57	.75	.81
4	.51	.71	.77
3	.44	.65	.71

Within-Competency Cluster Generalizability Analyses

In a second series of analyses, we examined whether our overall results across competencies were replicated within specific competency clusters. As is often the case in competency modeling, the 40 competencies were grouped under 5 broad clusters, with each cluster containing 8 competencies. Table 3 presents the results of these within-competency cluster generalizability analyses. Note that this was a more stringent test of our hypothesis concerning discriminant validity because it enabled us to examine whether raters were able to make fine-grained distinctions between competencies within a specific cluster. The results described in Table 3 replicate the pattern obtained in the overall generalizability analyses; that is, raters in the expert group were best able to discriminate among competencies, followed by those in the training group, and then by those in the control group. In line with prior results, the variance components due to competencies in the expert and training groups were twice as large as the variance component due to competencies in the control group. An inspection of the generalizability coefficients for 3 raters also replicated the pattern observed in the overall generalizability analyses; that is, generalizability coefficients in the expert group were larger than those computed in the other two groups.

The within-cluster analyses also enabled an examination of differences among competency clusters. Discriminant validity results appeared less than satisfactory for the personal management cluster, which included competencies such as adaptability and self-development, with competencies explaining, at most, 22% of the variance. For this cluster, the highest generalizability coefficient obtained was .47.

Table 3
Results of Within-Competency Cluster Generalizability Analyses Broken Down by Condition

	Control		Trained		Expert	
Effect	VC	%	VC	%	VC	%
Cluster 1: Information management competencies						
Raters	.02	4	.00	0	.00	0
Competencies	.13	23	.33	49	.41	57
Competencies × Raters	.43	74	.35	51	.31	43
G-coefficient (3 raters)	.48		.74		.80	
Cluster 2: Task management competencies						
Raters	.00	0	.00	0	.00	0
Competencies	.17	26	.34	49	.41	58
Competencies × Raters	.51	74	.36	51	.30	42
G-coefficient (3 raters)	.51		.74		.80	
Cluster 3: Leadership competencies						
Raters	.09	17	.01	2	.01	1
Competencies	.06	12	.16	30	.22	39
Competencies × Raters	.39	71	.36	67	.33	60
G-coefficient (3 raters)	.33		.57		.66	
Cluster 4: Interpersonal competencies						
Raters	.08	13	.00	0	.00	0
Competencies	.14	23	.25	42	.31	47
Competencies × Raters	.38	64	.34	58	.35	53
G-coefficient (3 raters)	.52		.69		.72	
Cluster 5: Personal management competencies						
Raters	.05	11	.03	6	.02	4
Competencies	.03	8	.07	15	.12	22
Competencies × Raters	.36	81	.39	79	.40	74
G-coefficient (3 raters)	.22		.36		.47	

Note. Due to rounding, percentages do not sum to 100. VC = variance component.

Within-Competency Generalizability Analyses

Finally, we tested our hypotheses by conducting within-competency generalizability analyses. Forty generalizability analyses were conducted, one for each competency. The only available facet in these within-competency analyses was raters. Therefore, only hypotheses concerning interrater reliability (i.e., Hypotheses 1b and 2b) could be tested. Results of these within-competency generalizability analyses replicated those observed in our previous analyses. Across the 40 competencies, the mean variance component due to raters was .46 in the control group. In the training group, the mean variance component due to raters was .36. The expert group had the smallest variability among raters, with a mean variance component of .34. To examine whether these between-group differences were statistically significant, we conducted an analysis of variance (ANOVA) using the variance components due to raters as a dependent variable and condition as an independent variable. A significant main effect emerged, $F(2, 117) = 4.75$, $p < .05$, partial $\eta^2 = .08$. Post hoc tests revealed this effect was due to the significant difference ($p < .01$) between the control group and the other two groups, therefore lending support to Hypothesis 1b. Hypothesis 2b was not supported because there was no statistically significant difference between the training and expert groups.

Accuracy Analyses

We chose the eight "essential" competencies associated with the job of method engineer that were previously determined by the trainers and job incumbents as the standard against which competency ratings made by the consultants would be compared (for a similar field-based approach to accuracy, see Hahn & Dipboye, 1988). We used two accuracy measures. First, we used a signal-detection framework (Lord, 1985) wherein the 8 competencies deemed essential by the trainers and the job incumbents were considered to be "target" competencies, whereas the remaining 32 competencies were considered to be "noise" competencies. According to the formula given by Lord (1985), a standardized difference (or d') between the proportion of hits, which was defined as placing a given competency in the essential category when it was indeed a target competency, and the proportion of

false alarms, which involved placing a given competency in the essential category when it was a noise competency, was computed for each consultant. Thus, in the present study, d' refers to the extent to which individuals were accurate in discerning essential from nonessential competencies for a given job. The results are presented in Table 4. To examine statistically significant differences in accuracy among groups, we conducted an ANOVA where the d' index was the dependent variable, and condition was the independent variable. A significant main effect was found, $F(2, 70) = 13.44$, $p < .001$, partial $\eta^2 = .28$. Post hoc tests indicated that this effect was due to the significant difference ($p < .001$) between the control group and the other two groups. There was no statistically significant difference between the training and expert groups. These findings support Hypothesis 1c but fail to provide support for Hypothesis 2c.

Second, we computed an overall accuracy index (with lower scores indicating higher accuracy; see Table 4). To this end, we computed the sum of the squared distances between the consultants' ratings and the trainers/job incumbents' ratings, which were coded as 1 and 0 for essential and nonessential competencies, respectively. Next, we conducted an ANOVA where the overall accuracy index was the dependent variable, and condition was the independent variable. A significant main effect emerged, $F(2, 70) = 13.91$, $p < .001$, partial $\eta^2 = .28$. Post hoc tests indicated that this effect was due to the significant difference ($p < .001$) between the control group and the two other groups. There was no statistically significant difference between the training and expert groups. Again, these findings support Hypothesis 1c but not Hypothesis 2c.

Discussion

Despite its popularity among practitioners, competency modeling has been criticized for lacking the rigor necessary for making a valid determination of the competencies required for job performance (Lievens et al., 2004; Morgeson et al., 2004; Schippmann et al., 2000). This study provides continued support for the notion that competency modeling outcomes might lack acceptable psychometric properties when no methodological safeguards are in place. In fact, our findings reveal that inexperienced and untrained consultants displayed poor levels of interrater reliability and also had difficulty both distinguishing among the various competencies and discerning essential competencies.

A key contribution of this study is the identification of an important procedural factor impacting the quality of competency inferences. Specifically, the provision of training for those involved in competency modeling seemed a practical means to increase the quality of their competency ratings. Our training program incorporated the principles underlying FOR training, which has a proven record in the performance appraisal domain (Cardy & Keefe, 1994; Day & Sulsky, 1995; Sulsky & Day, 1994). The main training objective was instilling an appropriate organizing schema for rating competencies (i.e., the behaviorally based competency model) on raters so that they would use this categorization schema instead of their own preexisting schemata. Our results generally support the beneficial effects of the training program. Trained raters had higher levels of interrater reliability than did untrained raters. In addition, they made more fine-grained distinctions among competencies than did untrained consultants. Moreover, trained consultants were more accurate than untrained consultants in discerning essential competencies from nonessential ones. The return on investment derived from the relatively brief training format studied here promises to be large. For instance, we found that only 4 trained consultants were needed to obtain an interrater reliability of .70, whereas 9 untrained consultants were needed to attain the same outcome.

Also noteworthy are the findings that the discriminant validity and interrater reliability of the trained-expert group were higher than those of the training-only group, although the differences are less dramatic than those observed with the control group (e.g., there was no statistically significant difference in terms of accuracy). This finding suggests that the effects of FOR training are augmented by having prior experience in competency modeling. However, differences between the expert group and the other two groups should be interpreted with caution because, unlike the training and control groups, the expert group was a nonequivalent, preexisting group that was not formed through random assignment (note the differences in age, work experience, and sex ratios between the expert group and the other two groups). One might also argue that experts might have forgotten some of what had been imparted during training. Perhaps they developed shared schemata over time, which contributed to increases in reliability. In any case, our design did not allow for disentangling training effects from experience effects. The practical constraints imposed by the field setting (i.e., untrained experienced consultants were not available) precluded the addition of an expertise-only control group. Future studies that include Solomon-four group designs with experience-only and training-only groups may shed light on the separate and the combined effects of training and experience.

Table 4
Results of Accuracy Analyses Broken Down by Condition

	Control		Trained		Expert	
Effect	*M*	*SD*	*M*	*SD*	*M*	*SD*
Signal detection accuracy index (d')	.29	.32	.64	.38	.79	.32
Overall accuracy index	11.65	1.57	10.00	1.63	9.41	1.40

Note. d' = standardized proportion of hits − standardized proportion of false alarms (Lord, 1985). Higher d' values indicate higher accuracy. The overall accuracy index is the sum of the squared distances between the consultants' ratings and the trainers'/job incumbents' ratings. Lower overall accuracy values indicate higher accuracy.

Also interesting were the differences among competencies. Apparently, consultants had the most difficulty rating the cluster of competencies related to personal management, as indicated by their lower interrater reliabilities. A potential explanation might be that these competencies are possibly less observable (Salgado, Moscoso, & Lado, 2003). Of course, conclusions with regard to the extent to which specific competencies can be reliably rated are bound by the fact that consultants rated only one job. Training effectiveness may decrease as the training jobs increasingly diverge from the jobs being rated. Therefore, future research that includes different jobs is warranted.

This study is not without limitations. First, we focused on one specific competency modeling instrument used by a large HR service firm. Although this instrument conformed to the primary characteristics of competency modeling practices outlined by Schippmann et al. (2000), the generalizability of our results to other instruments calls for further examination. Similarly, our relatively small sample of HR consultants and the placement of the rating task approximately 1 week after the training session also bound the generalizability of our results. Even though our sample of HR consultants seems timely because external consultants often play a leading role in competency modeling, continued research using other populations (e.g., job incumbents or supervisors) and other timeframes should prove useful. Second, our evaluation focused on key outcomes of work analysis (discriminant validity, reliability, and accuracy). Given the time constraints imposed by our field setting, we did not have a chance to include other levels of training evaluation (e.g., the learning acquired and the cognitive processes triggered by the training).

In summary, our findings reveal a potentially cost-effective tool to improve the allegedly loose practice of competency modeling. They also extend the generalizability of the principles of FOR training to domains other than performance appraisal and to populations other than psychology students. Our extension of the FOR training format to the work analysis domain opens a fruitful avenue to those interested in increasing the rigor of competency modeling.

References

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review, 6,* 205–212.

Bernardin, H. J., Buckley, M. R., Tyler, C. L., & Wiese, D. S. (2000). A consideration of strategies in rater training. In G. Ferris (Ed.), *Research in personnel and human resource management* (Vol. 18, pp. 221–274). Greenwich, CT: JAI Press.

Brennan, R. L. (1992). *Elements of generalizability theory.* Iowa City, IA: American College Testing Program.

Cardy, R. L., & Keefe, T. J. (1994). Observational purpose and evaluative articulation in frame-of-reference training: The effects of alternative processing modes on rating accuracy. *Organizational Behavior and Human Decision Processes, 57,* 338–357.

Chi, M. T. H., Glaser, R., & Farr, M. J. (1988). *The nature of expertise.* Hillsdale, NJ: Erlbaum.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Day, D. V., & Sulsky, L. M. (1995). Effects of frame-of-reference training and information configuration on memory organization and rating accuracy. *Journal of Applied Psychology, 80,* 158–167.

DeNisi, A. S., Cornelius, E. T. III, & Blencoe, A. G. (1987). Further investigation of common knowledge effects on job analysis ratings. *Journal of Applied Psychology, 72,* 262–268.

Dierdorff, E. C., & Wilson, M. A. (2003). A meta-analysis of job analysis reliability. *Journal of Applied Psychology, 88,* 635–646.

Dorfman, P. W. (1982, August). *Schema and network representations of knowledge: Implications for performance appraisal.* Paper presented at the annual convention of the American Psychological Association, Washington, DC.

Gatewood, R. D., & Feild, H. S. (2001). *Human resource selection.* Orlando, FL: Harcourt.

Greguras, G. J., & Robie, C. (1998). A new look at within-source interrater reliability of 360-degree feedback ratings. *Journal of Applied Psychology, 83,* 960–968.

Hahn, D. C., & Dipboye, R. L. (1988). Effects of training and information on the accuracy and reliability of job evaluations. *Journal of Applied Psychology, 73,* 146–153.

Hauenstein, N. M. A. (1998). Training raters to increase the accuracy and usefulness of appraisals. In J. Smither (Ed.), *Performance appraisal: State-of-the-art methods for performance management* (pp. 404–444). San Francisco: Jossey Bass.

Jacobs, S. L., Kulik, C. T., & Fichman, M. (1993). Category-based and piecemeal processes in job evaluations. *Journal of Applied Social Psychology, 23,* 1226–1248.

Kulik, C. T. (1989). The effects of job categorization on judgments of the motivating potential of jobs. *Administrative Science Quarterly, 34,* 68–80.

Lievens, F., Sanchez, J. I., & De Corte, W. (2004). Easing the inferential leap in competency modeling: The effects of task-related information and subject matter expertise. *Personnel Psychology, 57,* 881–904.

Lombardo, M., & Eichinger, R. (2003). *The LEADERSHIP ARCHITECT® norms and validity report.* Minneapolis, MN: Lominger.

Lord, R. G. (1985). Accuracy in behavioral measurement: An alternative definition based on raters' cognitive schema and signal detection theory. *Journal of Applied Psychology, 70,* 66–71.

Morgeson, F. P., & Campion, M. A. (1997). Social and cognitive sources of potential inaccuracy in job analysis. *Journal of Applied Psychology, 79,* 475–480.

Morgeson, F. P., & Campion, M. A. (2000). Accuracy in job analysis: Toward an inference-based model. *Journal of Organizational Behavior, 21,* 819–827.

Morgeson, F. P., Delaney-Klinger, K. A., Mayfield, M. S., Ferrara, P., & Campion, M. A. (2004). Self-presentation processes in job analysis: A field experiment investigating inflation in abilities, tasks, and competencies. *Journal of Applied Psychology, 89,* 674–686.

Quinones, M. A., Ford, J. K., & Teachout, M. S. (1995). The relationship between work experience and job performance: A conceptual and meta-analytic review. *Personnel Psychology, 48,* 887–910.

Salgado, J. F., Moscoso, S., & Lado, M. (2003). Test-retest reliability of ratings of job performance in managers. *International Journal of Selection and Assessment, 11,* 98–101.

Sanchez, J. I. (1994). From documentation to innovation: Reshaping job analysis to meet emerging business needs. *Human Resource Management Review, 4,* 51–74.

Sanchez, J. I., & Levine, E. L. (1994). The impact of raters' cognition on judgment accuracy: An extension to the job analysis domain. *Journal of Business and Psychology, 9,* 47–57.

Sanchez, J. I., & Levine, E. L. (2001). The analysis of work in the 20th and 21st centuries. In N. Anderson, D. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work, and organizational psychology: Vol. 1. Personnel psychology* (pp. 71–89). London/New York: Sage.

Schippmann, J. S., Ash, R. A., Battista, M., Carr, L., Eyde, L. D., Hesketh, B., et al. (2000). The practice of competency modeling. *Personnel Psychology, 53,* 703–740.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Smith, J. E., & Hakel, M. D. (1979). Convergence among data sources, response bias, and reliability and validity of a structured job analysis questionnaire. *Personnel Psychology, 32,* 677–692.

Snow, C. C., & Snell, S. A. (1992). Staffing as strategy. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 71–98). San Francisco: Jossey-Bass.

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

Sulsky, L. M., & Day, D. V. (1994). Effects of frame-of-reference training on rater accuracy under alternative time delays. *Journal of Applied Psychology, 79,* 535–543.

Woehr, D. J. (1994). Understanding frame-of-reference training: The impact of training on the recall of performance information. *Journal of Applied Psychology, 79,* 525–534.

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67,* 189–205.

Received January 29, 2006
Revision received July 7, 2006
Accepted August 15, 2006 ■