Psychology Science Quarterly, Volume 51, 2009 (1), pp. 3 - 15

Quality of the Critical Incident Technique in practice: Interrater reliability and users' acceptance under real conditions

ANNA KOCH[1], ANJA STROBEL[2], GÜLER KICI[3], & KARL WESTHOFF[2]

Abstract

The Critical Incident Technique (CIT) is a widely used task analysis method in personnel psychology. While studies on psychometric properties of the CIT so far primarily took into account relevance ratings of task-lists or attributes, and hence, only a smaller or adapted part of the CIT, little is known about the psychometric properties of the complete CIT in its most meaningful and fruitful way. Therefore, the aim of the present study was to assess interrater reliability and the participants' view of the CIT under real conditions and especially to provide data for the key step of the CIT: the classification of behavior descriptions into requirements. Additionally, the cost-benefit-ratio and practicability were rated from the participants' views as an important indicator for the acceptance of the task analysis approach in practice. Instructors of German Institutions for Statutory Accidents Insurance and Prevention as well as their supervisors took part in a job analysis with the CIT. Moderate interrater reliability for the relevance rating was found while the classification step yielded unexpectedly low coefficients for interrater reliability. The cost-benefit-ratio and practicability of the complete CIT were rated very positive. The results are discussed in relation to determinants that facilitate or impede the application of task analysis procedures.

Key words: critical incident technique, interrater reliability, acceptance, practical aspects, task analysis

[1] Correspondence should be addressed to Anna Koch, M.Sc., Dresden University of Technology, Department of Psychology, Assessment and Intervention, 01062 Dresden, Germany; e-mail: anna.koch@tu-dresden.de.

[2] Dresden University of Technology, Department of Psychology, Assessment and Intervention, 01062 Dresden, Germany

[3] BG Institute Work and Health, Königsbrücker Landstraße 2, 01109 Dresden, Germany

Introduction

The Critical Incident Technique (CIT, Flanagan, 1954) is a widely used task analysis method in personnel psychology (cf. Anderson & Wilson, 1997). The CIT consists of "a set of procedures for collecting direct observations of human behavior in such a way as to facilitate their potential usefulness in solving practical problems and developing broad psychological principles (Flanagan, 1954, p. 327)". These direct observations, or incidents, are called Critical Incidents (CI), because they are critical in the sense that they describe (a) crucial situational demands in a given job and (b) behaviors in these situations which discriminate between successful and less successful job performance. The main steps of the CIT *sensu* Flanagan (1954) are: (1) gathering CI by interviewing or observing subject-matter experts, (2) rating the relevance of these CI, and (3) classifying these CI into job requirements.

Thus, the CIT provides a valuable basis for the development of selection interviews such as the Situational Interview (Latham, Saari, Pursell, & Campion, 1980), the Behavior Description Interview (Janz, Hellervik, & Gilmore, 1986), or the Multimodal Employment Interview (Schuler, 1992) as well as for the construction of Assessment Center (AC) tasks or standardized assessment tests as Situational Judgement Tests (McDaniel & Nguyen, 2001). Moreover, the CIT has also been applied in various areas beyond the field of personnel psychology, e.g. to determine reasons for success and failure of university students (Schmelzer, Schmelzer, Figler, & Brozo, 1987) or to analyze aspects of service quality (cf. Gremler, 2004).

As the CIT is an only partially structured procedure for collecting qualitative data and can be adapted to specific application demands (Anderson & Wilson 1997; Chell, 1998; Gremler, 2004), its psychometric properties as well as its economic and practical aspects largely depend on these specific demands and need to be re-determined in each new field of application. This lack of possible generalization from one finding to another in a different field (Gremler, 2004) has prompted the development of standardized task analysis procedures, e.g. requirement lists (see Lievens, Sanchez, & De Corte, 2004; Morgeson, Delaney-Klinger, Mayfield, Ferrara, & Campion, 2004), where a partial sample of job holders or experts is asked for CI which then are used to develop job analysis questionnaires.

This approach considerably differs from the CIT in its original form, because the main advantage of the CIT – getting differentiated behavior-related explanations for each requirement and using them to build up e.g. interview guides or AC tasks – is lost by quasi moving backwards to lists. Nevertheless, the majority of studies in this field examined the psychometric properties (primarily the interrater reliability) and practical aspects of job-/task-lists based on a CIT-derived approach to identify work attributes (see Dierdorff & Wilson, 2003; Voskuijl & van Sliedregt, 2002), whereas only a few studies analyzed the CIT in its original and elaborated form (Andersson & Nielsson, 1964; Gremler, 2004; Ronan & Latham, 1974).

Studies of the first group analyze the interrater reliability of the job/task-lists and possible determinants of interrater reliability of job analysis. There are two recent meta-analyses giving a review of diverse studies on the reliability of these ratings (Dierdorff & Wilson, 2003; Voskuijl & van Sliedregt, 2002). Over all studies included in their meta-analysis, Voskuijl and van Sliedregt (2003) reported a high mean interrater reliability (for the rating of tasks, behavior, attributes, or other job dimensions) of $r = .56$ (Pearson's r, unweighted) and

r_w = .59 (Pearson's r, weighted by the number of jobs that were analyzed). Dierdorff and Wilson (2003) reported a high mean interrater reliability for relevance ratings of tasks over all studies of r = .77 (Spearman-Brown-corrected, sample-size weighted).

In contrast to these comprehensive results for standardized CIT-based ratings, the studies of the second group examining the original CIT *sensu* Flanagan are sparse and fragmentary. Systematic results only can be found in elder studies, namely Andersson and Nielsson (1964) and Ronan and Latham (1974). They systematically analyzed interrater reliability of the CIT. For relevance, rated by a questionnaire (six-point scale), moderate correlations between r = .27 and .42 were found, depending on the subgroup examined (Andersson & Nielsson, 1964). A high average rank correlation of r_s = .83 was found for the classification of CI into predefined requirements in workshops of two persons (Andersson & Nielsson, 1964). Ronan and Latham (1974) found a high 79 % concordance between three independent raters. However, these authors rearranged the steps of the CIT compared to Flanagan (1954), and in contrast to Flanagan's approach, the relevance rating was performed at the end of the studies, only considering the rating of the categories and subcategories, not the rating of the CI itself (Andersson & Nielsson, 1964; Flanagan, 1954).

Hence, apart from neglecting the rating of the CI itself, the meta-analyses only considered relevance ratings, that is, only the second of Flanagan's three steps of the CIT – i.e. (1) gathering, (2) relevance rating, and (3) classifying of CI (Flanagan, 1954) – a pattern which can be also found in studies of applied economics (Gremler, 2004). Not surprisingly, Gremler (2004) concludes that there is a lack of empirical studies analyzing the original CIT. This conclusion mainly refers to studies of step 3 (classifying CI) according to Flanagan's approach rather not to step 1 (gathering CI) because the basis for this step are diverse and broad descriptions of job situations and job behavior.

Taken together, despite of decades of CIT research, evidence is still missing regarding the interrater reliability of the CIT in its comprehensive original form (Gremler, 2004; Schuler, 2002). Nevertheless, due to the promising results of the available studies we expected moderate to high concordance for step 2 and 3 of the CIT in its comprehensive original form, too.

The interrater reliability of step 3 (classifying CI) still remains to be determined. Additionally, as Gremler (2004) pointed out, even the available evidence on the interrater reliability of step 2 (relevance rating) of the CIT is difficult to compare between studies, as different methodologies have been used or have not been adequately described at all. Hence, the primary objective of the present study was to obtain clearly defined interrater reliability coefficients for the crucial steps of the CIT as described by recognized experts in the field (Anderson & Wilson, 1997). In addition, we were interested in the perceived cost-benefit-ratio and practicability of the CIT, because they are important determinants of acceptance of a selection method by practitioners (Klingner & Schuler, 2004; Schuler, Hell, Trapmann, Schaar, & Boramir, 2007; Terpstra & Rozell, 1997). The basis for analyzing the perceived quality of selection procedures is the concept of social validity which includes all determinants that make a selection process a fair and acceptable social situation (Schuler, 1993). This definition includes also task analysis as the main basis for all forms of selection procedures. The perceived cost-benefit-ratio and practicability as components of social validity, together with the main psychometric properties, are important indicators for the quality of task analysis methods like the CIT, and also for its success in application to become used within companies and organizations (Carson, Becker, & Henderson, 1998; Klingner & Schuler, 2004).

Despite the fact that the CIT is currently a widely accepted task analysis method, only 5 % of German companies (Stephan & Westhoff, 2002) and 0 % of a German, Austrian and Swiss sample (Krause & Gebert, 2003) actually use it for task analysis. This is in sharp contrast to the 45 % of companies in the U.S. which use the CIT (Krause & Gebert, 2003).

Therefore, our study also addressed the issue whether there are reasons not to use the CIT that refer to its cost-benefit-ratio and practicability – an issue which so far has not been examined in CIT research. Based on existing questionnaires for selection instruments (Carson et al., 1998; Klingner & Schuler, 2004) and building on a self-developed systematic version of the CIT to pay equal tribute to the technique's theoretical, economic, and practical aspects, the aim of the present study was going beyond the existing evidence and to assess interrater reliability as well as the perceived cost-benefit-ratio and practicability of the CIT for the key steps of this method, that is the relevance rating and the classifying of the CI.

Methods

Setting and participants

A task analysis was conducted for instructors from the German Institutions for Statutory Accidents Insurance and Prevention. The requirement profile was used to develop an interview guide for assessing teachers and trainers in occupational health. This interview was the basis for a training where the collected job situations are simulated and the coping of the situations is practiced. Because of this focus, the use of the CIT was the best method for getting diverse descriptions of job situations as well as diverse and clear behavior descriptions from the job. This was the basis for behavior-related requirement categories.

Voluntary participants were recruited from the group of teachers and trainers of the organizations by presenting the aims of the study at supervisor meetings and by contacting potential participants via e-mail and telephone. The total number of participants was 45, not all of them participated in all steps of the CIT (see Table 1).

45 employees (63 % job holders and 24 % supervisors) participated in collecting the CI (step 1). 40 employees (45 % job holders and 19 % supervisors) took part in rating the relevance of the CI (step 2). In both steps, e-mail questionnaires were used. 15 employees (10 job holders and 5 supervisors) took part in the final classification of the CI into job requirements (step 3). In step 3, the participants rated 219 behavior descriptions in three five-hour workshops with five employees from three different locations of the organization each in order to obtain a broader data base and to allow a comparison of different group-results. At every step, the participating groups were representative samples in terms of age, gender, position, experience, and different locations of the organization (Table 1).

Data collection

Table 2 outlines the different steps of the study.

Table 1:
Sample Characteristics

	Step 1: Gathering CI	Step 2: Relevance rating	Step 3: Classification of CI
n	45	40	15
Age			
M	45	45	45
SD	9	9	9
Status			
Job holders	63 %	45 %	66 %
Supervisors	24 %	19 %	33 %
Job experience			
< 1 year	7 %	2 %	0 %
1-5 years	24 %	17 %	0 %
> 5 years	60 %	49 %	100 %

Table 2:
Methods and Results for Each Step of the CIT

	Step 1: Gathering CI	Step 2: Relevance rating	Step 3: Classification of CI
Method/Material	E-mail questionnaire	E-mail questionnaire	Workshops
Instruction	Report a job situation that you have experienced or observed in the past and that was handled effectively or ineffectively by a job holder	Rate all CI gathered from step 1 according to whether they are important for the success of the task concerned (5-point scale)	Classify the CI into requirement categories developed within the workshop
n_{CI}	144 resulting CI: 67 CI present- 77 CI future-oriented	reduced to 109 CI: 51 CI present- 58 CI future-oriented	219 behavior-descriptions
Interrater Reliability		Kendall $W = .32$ ($p < 0.01$)	Kappa $\kappa = .09$-$.14$ ($p > 0.05$)

Step 1: Gathering the CI. Standardized interviews and questionnaires accompanied by a cover letter describing well the background and purpose of the study are comparable concerning the quality and scope of the collected data (Jonassen, Hannum & Tessmer, 1989). Hence, the CI were collected by means of a questionnaire sent by e-mail to job holders and supervisors of the organization all over Germany. Therefore, a questionnaire was used for the gathering of CI similar to the CI report form of Anderson and Wilson (1997). In the beginning, the participants were given comprehensive instructions including a description of the CIT and the CI as well as some examples for CI from a similar job context (school teaching). Then the participants were asked to recall critical incidents by following a general instruction and six more specific questions (see Table 3).

To follow the recommendations for task analysis, the participants were asked for two present-oriented CI (bottom-up) and two future-oriented CI (top-down) (Heider-Friedel, Strobel, & Westhoff, 2006; Landis, Fogli, & Goldberg, 1998). The instructions for the future-oriented CI were similar to those for the present-oriented CI. The main difference was that the participants were asked to *imagine* what situations might result from the future trends for the tasks of instructors. The comprehensible examples given in the instructions for the present and future oriented CI assured that the participants completed the questionnaire in the intended way. In addition, the general instruction for the CI report and the six specific questions were presented at each page to be completed so the participants could easily follow them. At the end of the questionnaire, the participants had the opportunity to give open feedback about step 1.

Table 3:
CI report form

Recall a job situation which you have experienced or observed and which had been handled effectively or ineffectively by a job holder.
(1) What was the situation leading up to the event?
(2) What preceded the event?
(3) What did the instructor do?
(4) What was the result of the instructor's action?
(5) What was the result for the whole working process?
(6) What would have been ineffective behavior on the part of the instructor in this situation?

Note: instructors mean job holders dealing with vocational training and adult education in the field of occupational safety and health.

Step 2: Relevance rating. In line with other job analysis studies (Dierdorff & Wilson, 2003; Gremler, 2004; Voskuijl & van Sliedregt, 2003), a questionnaire was used for the relevance rating. On a 5-point relevance-scale, the participants were asked to rate each CI from step 1 in terms of whether it was quite relevant for the task of the instructors. The grades of the scale accordingly to present studies were: 1 = "very relevant", 2 = "rather relevant", 3 = "rather irrelevant", 4 = "irrelevant", and there was an additional fifth category: "same content as another CI". This category, as an additional form of data reduction, ensured that only incidents with different content were selected for the next step. As in step 1, the questionnaire included a section for open feedback at the end.

Step 3: Classification of behavior descriptions. The classification of the behavior descriptions in step 3 took place in four independent workshops. Each workshop consisted of three to five participants. In contrast to previous studies (Andersson & Nielsson, 1964; Ronan & Latham, 1974), the participants were asked to classify behavior descriptions into requirement categories that were *not* predefined (Anderson & Wilson, 1997), to get behavior-related as well as company-specific requirement categories.

The participants' task in each of the workshops was to sort the behavior descriptions into categories by consensus. The participants were therefore asked to decide which requirement label would be the best for the behavior described in a CI. At the beginning, the first behavior description defined the first requirement category. For each subsequent behavior description, the participants had to assess whether it fitted to the first behavior or not. If not, they had to create a new category, etc. Behavior descriptions not fitting into any category had to be classified in all workshops to the category "miscellaneous". Each workshop resulted in a requirement profile.

At the end of the workshop, the participants were asked to fill in a 9-item-feedback-questionnaire on the cost-benefit-ratio and practicability which would act as an indicator for economic and practical aspects of the task analysis conducted. Only those who participated in all three steps of the study were asked to answer the questionnaire, as it consisted of questions relating to all steps. The participants assessed cost-benefit of the task analysis in terms of four items and practicability in terms of five items (Table 4).

The items were verbalized as statements (e.g. "The cost-benefit ratio of questionnaire 1 is appropriate."), to which the participants could agree or disagree on a 4-point scale ranging from 1 = "agree", 2 = "slightly agree", 3 = "slightly disagree" to 4 = "disagree".

Merged requirement profile. The four requirement profiles from step 3 were merged together to prepare the result of the task analysis. If a behavior description was classified by three or four workshop groups into the same requirement category, this behavior description was assigned to the merged requirement profile. Behavior descriptions classified by different workshops into two or more different requirement categories were excluded from the merged requirement profile. At the end of this classification process the participants labeled the categories with a term which reflected the content of the behaviors in the respective category. In the result there was the final requirement profile, representing the completion of the task analysis.

Table 4:
Ratings of Economic and Practical Aspects (n = 15)

	M	*SD*
Items for economic aspects		
1. The cost and benefit of the 1st questionnaire is appropriate	1,73	0.59
2. The cost and benefit of the 2nd questionnaire is appropriate	2,07	0.88
3. The cost and benefit of the workshop is appropriate	1,93	0.80
4. I would use this method for conducting task analysis again	2,13	0.99
Items for practical aspects		
1. The aim of the 1st questionnaire was clear to me	1,57	0.65
2. The aim of the 2nd questionnaire was clear to me	2,00	0.85
3. The aim of the workshop was clear to me	1,67	0.72
4. I could bring my own view into the task analysis process	1,47	0.64
5. There is a practical benefit of the resulting requirement profile	1,79	0.70

Note. Ratings were made after behavior classification in Step 3. A 4-point rating scale was used (1 = "agree", 2 = "slightly agree", 3 = "slightly disagree", to 4 = "disagree")

Data analysis

Interrater reliability. Interrater reliability was determined in steps 2 and 3. In step 2, interrater reliability was determined by means of Kendall *W* (Wirtz & Caspar, 2002). The ratings of the participants on a 5-point relevance scale ("very relevant" to "irrelevant") and an additional category ("same content as another critical incident") were analyzed with regard to concordance of the individual ratings.

In step 3, interrater reliability was determined by frequency ratings in percentages according to corresponding studies (Andersson & Nielsson, 1964; Ronan & Latham, 1974). Additionally, owing to the category data the Kappa coefficient κ for $m > 2$ raters (Fleiss, 1971) was determined. This additional aspect of interrater reliability was determined in two ways: (1) whether the respective workshop groups developed comparable requirement profiles; and (2) whether they classified the different behavior descriptions into comparable requirement categories in the merged profile of the four workshops.

Perceived cost-benefit ratio and practicability. Perceived cost-benefit-ratio and practicability as indicators for economic and practical aspects were determined at the end of step 3. Participants of all four workshops were asked to rate the cost-benefit-ratio and practicability of the CIT via questionnaire (see table 2). The mean *M* of the ratings for each item was the estimate for the perceived cost-benefit-ratio, practicability and therewith the acceptance of the method by the participants.

Results

Table 2 (bottom row) gives the interrater reliability coefficients. Table 4 summarizes the results regarding perceived economic and practical aspects.

Interrater reliability

Concerning the relevance ratings (step 2) moderate and significant interrater reliability of Kendall $W = .32$ ($p < 0.01$) was found. A "slight" (Landis & Koch, 1977) interrater reliability of $\kappa = .09$ ($p > 0.05$; concordance probability 56 %, random probability 52 %) was found for concordance between the four workshop groups with regard to the requirement categories generated (table 2). With regard to concordance in terms of the classification of the CI into the merged profile of the four workshops, again "slight" interrater reliability ($\kappa = .14$, $p > 0.05$; concordance probability 24 %, random probability 14 %) was found.

Perceived cost-benefit ratio and practicability

The rating of statements concerning the cost-benefit-ratio and practicability of the CIT on a 4-point rating scale (1 = "agree", 2 = "slightly agree", 3 = "slightly disagree" to 4 = "disagree") yielded the following results: As presented in table 4, the perceived positive cost-benefit ratio of steps 1 to 3 was rated with means from $M = 1{,}73$ to $M = 2{,}07$ ($SD_{step1} = 0.59$; $SD_{step2} = 0.88$; $SD_{step3} = 0.80$). That implies the majority of participants affirmed the positive statements concerning cost-benefit of CIT. The participants rated the question of whether they would use the CIT again as well with a mean of $M = 2{,}13$ ($SD = 0.99$).

Similarly, all practical aspects were rated positively or very positively (Table 4). The participants' mean ratings concerning the questions as to how well the aims of steps 1 to 3 had been defined ranged from $M = 1{,}57$ to $M = 2{,}00$ ($SD_{step1} = 0.65$; $SD_{step2} = 0.85$; $SD_{step3} = 0.72$). The question as to how far the participants were able to bring their own view into the CIT was rated with a mean of $M = 1{,}47$ ($SD = 0.64$). The practical benefit of the resulting CIT requirement profile was rated with a mean of $M = 1{,}79$ ($SD = 0.70$).

In addition, the participants had the opportunity to give open feedback at the end of each step. After step 1 and step 2 of the CIT, they commented on insufficient clarity of aims. At the end of step 3, however, they frequently stated that everything had been clear.

Discussion

This study focused on the interrater reliability of the CIT as well as on participants' evaluations of its cost-benefit ratio and practicability. Therefore, a task analysis based on the CIT sensu Flanagan (1954) including further developments (Anderson & Wilson, 1997; Heider-Friedel et al., 2006) was conducted for instructors of German Institutions for Statutory Accidents Insurance and Prevention.

Interrater reliability of the CIT

Concerning the relevance rating, a moderate (Rindermann & Hentschel, 2003) interrater reliability of .32 (Kendall *W*) was found for the concordance of participants' relevance ratings. Because of using CI as items for the questionnaire in contrast to attributes or tasks, our study corresponds with the approach used by Andersson and Nielsson (1964) who also observed interrater reliabilities from .27 to .42 for step 2. Compared to meta-analyses (Dierdorff & Wilson, 2003; Voskuijl & van Sliedregt, 2003), the interrater reliability of .32 obviously is lower. However, the meta-analyses included studies using attributes or job-/task-lists, whereas in our study, CI were used as items. Regarding the demands of the task for the participants, CI are more complex to rate because they include a complete description of a situation and behavior. This discrepancy underlines the importance to differentiate between studies using simplified task-lists to more complex CI ratings. As the CI are used for further developments on the basis of the job-analysis, e.g. for assessment centre tasks or interview guides, the observed coefficients are particularly relevant.

For the classification of the behavior into non-predefined categories, low to moderate concordance (24 % to 56 %) resulting in slight Kappa coefficients (.09 - .14) was observed. These are noticeably lower values compared to previous studies using predefined categories with an average interrater reliability of .83 for workshops of two (Andersson & Nielsson, 1964) and a 79 % concordance between three independent raters (Ronan & Latham, 1974). Taking into account that a free behavior-classification in which the categories are developed during the classification process is to be favored in order to use all advantages of the CIT (Anderson & Wilson, 1997), for the first time the observed coefficients actually reflect the quality of the CIT under real conditions.

Perceived cost-benefit ratio and practicability

The cost-benefit-ratio and the practicability of the CIT-procedure were rated positive by the participants of the workshops. A limitation of these results is the number and character of participants that took part in the workshops. They were presumably highly motivated and took part in all the steps of the study. Thus, the generalizability of the ratings of this group remains to be determined and the results have to be replicated. Furthermore, in addition to the subjective ratings, further studies should investigate utility aspects to supplement our results with economic benchmarks like money or effectiveness (e.g. Klehe, 2004; Macan & Foster, 2004).

Nevertheless, for the first time the *participants'* view of the CIT process was evaluated and yielded positive results. As described above, in studies dealing with selection procedures (Klingner & Schuler, 2004; Schuler et al., 2007), cost-benefit ratio and practicability as indicators for acceptance are important aspects to facilitate the application of task analysis methods as well. There is a whole research domain examining utility perceptions and determinants of adopting human resource practices in organizations from the managers' view (e.g., Klehe, 2004; Macan & Foster, 2004; Subramony, 2006). Although we feel that the managers' view is an important determinant for adopting scientifically recommended instruments, recent studies showed that the participants' view influenced managers' decisions as well (Klingner & Schuler, 2004; Schuler et al., 2007). In our study, the resulting requirement

profile was used to develop an interview guide and development centre tasks for assessing teachers and trainers in occupational health. One can imagine that employees participating in the basic task analysis are likely to accept the selection procedures developed from the task analysis data.

Taken together, due to the study design, for the first time the observed coefficients actually reflect the quality of the CIT under real conditions. So, though the results are somehow disappointing concerning the interrater reliability, our study provides valuable information for the use and implementation of the CIT in practice. The interrater reliability for the key steps of the procedure is widely low and at most in a middle range when performing the CIT in its most meaningful and fruitful way. At this point it should be thought about facilitations using the CIT without setting aside the advantages of the procedure. The way the CIT is described in academic research literature is hard to understand without an appropriate education (Terpstra & Rozell, 1997). Users of the CIT need specific skills and experience, as no manuals exist for its application. Gremler (2004) presented first guidelines for the use of the CIT in service research. We are currently constructing tools for each step of the CIT. Each tool will provide different options for achieving the task analysis according to the specific situation within the organization. These tools should further enhance the acceptance of the CIT in practice and support the positive initial findings to economical and practical aspects of the procedure.

Acknowledgements

The authors would like to thank all the participants from the German Institutions for Statutory Accidents Insurance and Prevention who made this study possible. We are also indebted to Sabine Herbst, Rena Osternack, Reinhard Göbel, Peter Wende, Manfred Walter for their support, and to Ulrike Bollmann, Bodo Pfeiffer and Eva Höhne for their help. Furthermore, we want to thank Alexander Strobel for providing valuable feedback on earlier drafts of this manuscript. We also thank two anonymous reviewers for helpful comments that improved the quality of this paper.

References

Anderson, L., & Wilson, S. (1997). Critical incident technique. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement methods in industrial psychology* (pp. 89-112). Palo Alto, CA: Davis-Black.

Andersson, B.-E., & Nielsson, S.-G. (1964). Studies in the reliability and validity of the critical incident technique. *Journal of Applied Psychology, 48*, 398-403.

Carson, K. P., Becker, J. S., & Henderson, J. A. (1998). Is utility really futile? A failure replicate and an extension. *Journal of Applied Psychology, 83,* 84-96.

Chell, E. (1998). Critical incident technique. In G. Symon & C. Cassell (Eds.), *Qualitative methods and analysis in organisational research. A practical guide* (pp. 51-72). London: SAGE Publications.

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

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51,* 327-358.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76,* 378-382.

Gremler, D. D. (2004). The critical incident technique in service research. *Journal of Service Research, 7,* 65-89.

Heider-Friedel, C., Strobel, A., & Westhoff, K. (2006). Anforderungsprofile zukunftsorientiert und systematisch entwickeln – Ein Bericht aus der Unternehmenspraxis zur Kombination des Bottom-up- und Top-down-Vorgehens bei der Anforderungsanalyse [Future-oriented and systematic development of requirement profiles – a practice-report on the combination of bottom-up and top-down task analysis]. *Wirtschaftspsychologie, 1,* 23-31.

Janz, T., Hellervik, L., & Gilmore, D. C. (1986). *Behavior description interviewing.* Boston: Allyn & Bacon.

Jonassen, D. H., Hannum, W. H., & Tessmer, M. (1989). *Handbook of task analysis procedures.* New York: Praeger Publishers.

Klehe, U.-C. (2004). Choosing how to choose: institutional pressures affecting the adoption of personnel selection procedures. *International Journal of Selection and Assessment, 12*, 327-342.

Klingner, Y., & Schuler, H. (2004). Improving participants' evaluations while maintaining validity by a work sample-intelligence test hybrid. *International Journal of Selection and Assessment, 12*, 120-134.

Krause, D. E., & Gebert, D. (2003). A comparison of assessment center practices in organizations in German-speaking regions and the United States. *International Journal of Selection and Assessment, 4*, 297-312.

Landis, J., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159-174.

Landis, R.S., Fogli, L., & Goldberg, E. (1998). Future-oriented job analysis: a description of the process and its organizational implications. *International Journal of Selection and Assessment, 3,* 192-197.

Latham, G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). The situational interview. *Journal of Applied Psychology, 17,* 422-427.

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.

Macan, T. H., & Foster, J. (2004). Managers' reactions to utility analysis and perceptions of what influences their decisions. *Journal of Business Psychology, 19*, 241-253.

Morgeson, F. P., Delaney-Klinger, K., 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.

Rindermann, H. & Hentschel, C. (2003, September). *Problematik und Möglichkeiten von Effektstärkemaßen* [Problems and chances of effect sizes]. Vortrag auf der 6. Tagung Methoden und Evaluation, Wien, Österreich.

Ronan, W. W., & Latham, G. P. (1974). The reliability and validity of the critical incident technique: A closer look. *Studies in Personnel Psychology, 1,* 53-64.

Schmelzer, R. V., Schmelzer, C. D., Figler, R. A., & Brozo, W. G. (1987). Using the critical incident technique to determine reasons for success and failure of university students. *Journal of College Student Personnel, 28,* 261-266.

Schuler, H. (1993). Social validity of selection situations: a concept and some empirical results. In H. Schuler, J.L. Farr, & M. Smith (Eds.), *Personnel selection and assessment. Individual and organizational perspectives* (pp. 11-26). Hillsdale, NJ: Erlbaum.

Schuler, H. (2002). *Das Einstellungsinterview* [The job applicant interview]. Göttingen: Hogrefe.

Schuler, H. (2006). Arbeits- und Anforderungsanalyse [Job analysis]. In H. Schuler (Hrsg.), *Lehrbuch der Personalpsychologie* (S. 45-68). Göttingen: Hogrefe.

Schuler, H., Hell, B., Trapmann, S., Schaar, H., & Boramir, I. (2007). Die Nutzung psychologischer Verfahren der externen Personalauswahl in deutschen Unternehmen. Ein Vergleich über 20 Jahre [Use of personnel selection instruments in German organizations in the last 20 years]. *Zeitschrift für Personalpsychologie, 6*, 60-70.

Stephan, U., & Westhoff, K. (2002). Personalauswahlgespräche im Führungskräftebereich des deutschen Mittelstandes: Bestandsaufnahme und Einsparungspotenzial durch strukturierte Gespräche [Selection interviews for the placement of managerial personnel in German medium-sized enterprises: review of current practice and potential savings through structured interviews]. *Wirtschaftspsychologie, 3,* 3-17.

Subramony, M. (2006). Why organizations adopt some human resource management practices and reject others: An exploration of rationales. *Human Resource Management, 19,* 241-253.

Terpstra, D. E., & Rozell, E. J. (1997). Why some potentially effective staffing practices are seldom used. *Public Personnel Management, 26,* 483-495.

Voskuijl, O. F., & van Sliedregt, T. (2002). Determinants of interrater reliability of job analysis: A meta-analysis. *European Journal of Psychological Assessment, 18*, 52-62.

Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität* [Interrater agreement and interrater reliability]. Göttingen: Hogrefe.