Self and supervisor ratings of job-performance: Meta-analyses and a process model of rater convergence.

Heike Heidemeier
Self and supervisor ratings of job-performance:
Meta-analyses and a process model of rater convergence.

Heike Heidemeier (Dipl.-Psych.)
10. Mai 2005
Erstreferent:

Prof. Dr. Klaus Moser
Friedrich-Alexander-Universität Erlangen-Nürnberg
Lehrstuhl für Psychologie, insbesondere
Wirtschafts- und Sozialpsychologie
Lange Gasse 20, 90403 Nürnberg

Koreferent:

Prof. Dr. Ingo Klein
Friedrich-Alexander-Universität Erlangen-Nürnberg
Lehrstuhl für Statistik und Ökonometrie
Lange Gasse 20, 90403 Nürnberg
Self and supervisor ratings of job-performance: Meta-analyses and a process model of rater convergence.

Contents

1 Introduction
  1.1 The study’s purpose ........................................ 3
  1.2 Methodological comment: Validity, leniency, and self-ratings of job performance 5
  1.3 The importance of self-other congruence in performance ratings ................... 7

2 Previous research
  2.1 Previous meta-analytical results ........................................ 9
  2.2 Previous results regarding moderators of self-supervisor agreement ............ 10

3 Hypotheses
  3.1 Report format .................................................... 15
  3.2 Conditions of report .............................................. 22
  3.3 Sample composition .............................................. 27

4 Method
  4.1 Literature search procedures ..................................... 31
  4.2 Inclusion criteria and the coding of studies ..................................... 31
  4.3 Meta-analytical techniques ....................................... 33

5 Results
  5.1 Meta-analytical results for overall $r$ and $d$ ....................................... 40
  5.2 Moderator analyses .............................................. 41
    5.2.1 Report format .............................................. 44
    5.2.2 Conditions of report ...................................... 46
    5.2.3 Sample composition ...................................... 47

6 Discussion
  6.1 Correlations of self- and supervisor-ratings (the overall validity of self ratings) 52
  6.2 Leniency in self-ratings .......................................... 53
  6.3 Moderator analyses .............................................. 54
    6.3.1 Report format .............................................. 54
    6.3.2 Conditions of report ...................................... 59
    6.3.3 Sample composition ...................................... 63

7 A process model of performance appraisal ........................................ 66

8 Limitations and future research .................................................. 74
Appendix

A Coding criteria

B The classification of performance dimensions

C Overall results based on the mixed-model

D Moderator analyses based on the mixed-model

E Intercorrelation matrix of moderator variables

F The distributions of effect-size indexes $r$ and $d$

G Effect-sizes and moderator variables as they were extracted from primary research

Deutschsprachiger Anhang

A Deutschsprachige Zusammenfassung

A.1 Einleitung

A.2 Ziel der Arbeit

A.3 Hypothesen

A.4 Methode

A.5 Ergebnisse

A.6 Integration der Ergebnisse in einem Modell der Urteiler-Konvergenz

B Lebenslauf
List of tables

1  Moderator variables examined by previous meta-analyses .......................... 13
2  Studies in the review: publication statistics ............................................. 39
3  Overall results for both effect size estimates $r$ and $d$ ............................. 40
4  Moderator analyses for effect size index $r$ ............................................. 41
5  Moderator analyses for effect size index $d$ ............................................. 43
6  Moderator variables confirmed by the fixed effects analysis ......................... 49
7  Moderator variables confirmed by the mixed effects analysis ....................... 51
8  Coding criteria .................................................................................. 91
9  The classification of performance dimensions ........................................... 93
10 Overall results based on the mixed-effects model ...................................... 94
11 Moderator results based on the mixed-effects model .................................. 95
12 Intercorrelation matrix of moderator variables ........................................ 98

List of figures

1  Overview: The moderator variables examined .......................................... 15
3  Distributions of effect-size index $r$ and $d$ ............................................. 99
Abstract

This dissertation presents meta-analyses on the congruence of self- and supervisor-ratings for job performance (k=104 independent samples). It examines two indicators of rater convergence (self-supervisor correlations and mean difference scores) and reports results for an extensive set of moderator variables. Correlation coefficients are interpreted as a measure of self-rating validity whereas the higher mean level of self-ratings compared to supervisory ratings is considered to indicate leniency in self-ratings. Self-supervisor ratings yielded an overall correlation of r=.22 (k=96; n=22287). Job type (i.e. blue collar vs. white collar; managerial responsibility) and the use of non-judgmental performance dimensions were the main moderators. The analysis also confirmed the notion that self-ratings of performance are lenient (d=.33; k=70; n=29386). For Western samples, self-ratings were consistently lenient under all conditions, but the extent to which ratings were lenient depended on situational variables (e.g. purpose for ratings). The two measures of rater congruence varied independently and were moderated by different sets of variables. A process model of performance rating is presented that integrates the study’s findings.

1 Introduction

Performance appraisals are an inevitable activity in organizations (Viswesvaran & Ones, 2000). They serve a variety of purposes - such as making administrative decisions, satisfying legal requirements, managing and developing employees, or assessing training needs. The most frequently used measure of job performance continues to be performance ratings. Ratings are subjective evaluations that are reported in questionnaire-based feedback procedures. The sources of ratings may include various rater groups, with the most prominent source still being the incumbents’ immediate supervisors. But it is also not a rare case that other rater groups (peers, subordinates, self, and clients) contribute evaluations (Cascio, 1998).

The focus of this thesis is on performance ratings made by the target persons of ratings, i.e. on self-ratings. Trends in performance evaluation have emerged that led to a growing importance of self-assessment. For example, beginning in the 1950’s, management systems such as "Management by Objectives" started to include self-assessments into management practices to serve as a basis for negotiating goals and performance criteria for managers (Drucker, 1954). A second
trend gained impetus during the 70’s and has since stayed probably the most prominent reason for the use of self-appraisals in performance evaluation: the use of performance appraisals for purposes of employee development (Murphy & Cleveland, 1995). Multi-source or so-called 360-degree feedback procedures especially include self-appraisals for this reason (London & Smither, 1995). They typically aim at highlighting rating discrepancies between the self and other rating sources (e.g. supervisors, peers, subordinates or customers) and thus allow explicit comparisons of self-ratings to ratings from other sources. These comparisons are believed to be a potential source of insight to the target person (Kwan, John, Kenny, Bond, & Robins, 2004), and are expected to initiate behavior change and an increase in effectiveness on the job (Leslie & Fleenor, 1998). Changes in working conditions also make self-assessments an increasingly important tool. For instance, a growing number of highly specialized employees are working without direct supervision or perform tasks that supervisors find hard to evaluate. As self-management generally becomes an increasingly relevant demand (e.g. Frayne & Geringer, 2000), so does one of its core components, appropriate self-evaluation. Self-evaluation is an important part of self-regulatory capability, a basic personal capability in Bandura’s (1986) well-known social-cognitive theory (Cervone, 2004).

The growing importance of self-appraisals remains a paradox. Although, to the individual, the self is "the most available and trustworthy source of feedback" (Ashford, 1989, p. 135) and accurate self-appraisals have instrumental value to self-raters (Ashford, 1989; Ashford & Cummings, 1985), it is widely held that self-reports do not adequately reflect true performance due to deficiencies in self-ratings, such as low validity and leniency. In fact, both influential reviewers (e.g. Thornton, 1980; Campbell & Lee, 1988) and earlier meta-analyses have reported evidence that self-appraisals of performance tend to show rather low correlations with ratings made by others (Harris & Schaubroeck, 1988; Mabe & West, 1982). Correlations of self-ratings with ratings from other sources are typically lower than correlations between other rating sources, such
as supervisors and peers (Conway & Huffcutt, 1997). In addition, self-appraisals are frequently lenient, i.e. individuals tend to rate their job performance at higher levels of favorability than others do (Harris & Schaubroeck, 1988). For reasons like these, Campbell and Lee (1988) even stated that it was "unreasonable" to expect self-assessment to serve as an evaluation device.

It seems that the growing relevance of self-evaluation is in contradiction to its repeated critique. Both practitioners and researchers may have contributed to this situation. On the one hand, the practice of performance management has not really collected much information from research, on the other hand, research has produced contradictory findings that are difficult to integrate. If it is possible to show that the outcomes of self-appraisals depend on contextual variables, a contribution to resolving this contradiction can be made. The psychometric problems of self-ratings may be specific to certain circumstances. For this reason, and as earlier reviews of self-appraisals may be out of date, a systematic review of contextual factors that moderate the outcomes of self-appraisals is warranted.

1.1 The study’s purpose

The primary purpose of this study is to examine the convergence of self- and supervisory ratings in performance appraisals. The present study tries to identify contextual factors that determine the extent to which self-assessments agree with assessments provided by the target person’s supervisor. Three sets of such factors will be examined, (a) features of the appraisal system, including properties of rating scales and scale anchors, (b) factors that pertain to conditions under which appraisal procedures are conducted, such as the purpose of performance ratings or their confidentiality, and (c) characteristics of the samples studied, such as job type and educational level. The review is primarily "design-oriented" in that it asks under what conditions self-ratings reach lower or higher levels of congruence; that is, we ask: Does congruence depend on rating method, rating context, and type of rating target? I do not provide a single, unifying body of theory which motivates the examination of all the hypotheses. The main reason is that the mech-
anisms underlying the influence of each moderator are both complex and sometimes specific to single moderators. Therefore, a short rational will be presented for the examination of each of the moderator variables separately, along with an interpretation of its expected effect and key findings of previous research.

The technique used to assess the influence of these variables is that of meta-analysis. Therein this dissertation strives to integrate existing research findings to reach an overview of the empirical support that various hypotheses have achieved regarding the determinants of self-supervisor agreement in performance ratings. This review is directed at both specialized researchers in the field of performance appraisals and a broader audience, including researchers who are interested in self-appraisals or agreement in ratings more generally, as well as practitioners who wish to gain an overview of situational factors that are relevant to the outcomes of performance appraisals - if the self is included as a rating source.

The presentation of the analysis is sequenced as follows. First, the hypotheses that will be examined are presented. That is, in accordance with its principal research question, this thesis initially focuses on the identification of contextual factors that affect interrater congruence. I am aware that several theoretical models of the appraisal process have been suggested (e.g. Ilgen, Barness-Farrell & McKelling, 1993), but found that none of them could entirely guide the analysis. Thus I provide a rationale for each of the hypotheses based on previous research findings without referral to an integrating model of performance rating. Only in the discussion, a process model of performance appraisal will be presented to integrate findings and locate the variables studied in the larger appraisal system. In this part, I also discuss major theoretical approaches that explain self-supervisor convergence in performance ratings.

Contributions of this study come in two areas. Compared to earlier meta-analyses (Mabe & West, 1982; Harris & Schaubroeck, 1988; Conway & Huffcutt, 1997), a more comprehensive set
of moderator variables is examined, some of which have not been included in meta-analytical research before. The present study contributes to the existing literature in that it assesses how contextual factors impact self-other agreement in ratings. This should also provide information about the level of agreement that can be expected under various conditions. Second, to my knowledge this study is the first meta-analysis that assesses moderator effects for both indicators of convergence in ratings: correlational agreement (index $r$) and differences in the mean levels of ratings, i.e. leniency (index $d$). Two previous meta-analyses on self-other agreement in performance ratings did not examine leniency (Conway & Huffcutt, 1997; Mabe & West, 1982). The only meta-analysis that did, could not examine moderator variables due to a rather small database of samples (Harris & Schaubroeck, 1988). The two indicators of convergence, effect size indexes $r$ and $d$, have been reported to be empirically independent from each other (Warr & Bourne, 1999) and may also be moderated by different sets of variables. In fact, I expect some moderator variables to influence mean difference scores more strongly than correlation coefficients.

1.2 Methodological comment: Validity, leniency, and self-ratings of job performance

The validity of self-ratings. To assess the validity of self-appraisals, this study links self-ratings to supervisor ratings. Four reasons led to the decision to use supervisor ratings as a criterion measure to assess the validity of self-ratings. First, cumulative evidence suggests that supervisors are the most reliable source of ratings (Conway & Huffcutt, 1997; Viswesvaran, Ones, & Schmidt, 1996). Second, supervisory judgments are likely to be particularly important to job incumbents. For example supervisory appraisals are usually most clearly associated with organizational reward systems. Third, research has accumulated some evidence that supervisor ratings are more highly related to performance as measured by external criteria than are ratings
from other sources (e.g. Atkins & Wood, 2002; Becker & Klimonski, 1989; Beehr, Ivanitskaya, Hansen, Erofeev, & Gudanowski, 2001). Finally, correlations of self and supervisor ratings have been reported more frequently in primary research than correlations of self-ratings with other sources. Thus, choosing supervisor ratings as criterion for self-rating validity promised a large database as well as the use of an important criterion. Of course, I am aware that relationships with other criteria (e.g. other rating sources) may be of equal interest. However, the main purpose of the present paper is not to draw conclusions regarding the construct validity of self-appraisals. The study aims at examining various factors that moderate the validity of self-ratings referring to a given single criterion. Thus, the analysis is confined to self-supervisor correlations. Throughout the text, the term validity of self-ratings will be used to refer to this single indicator of criterion-based validity.

Leniency in performance ratings. Leniency bias refers to a tendency of raters to report undeservedly favorable scores. Whether ratings are undeservedly favorable is often concluded from a comparison of the empirical distribution of performance ratings and a hypothetical distribution of performance, which is typically that of a normal distribution centered around the scale midpoint (Guilford, 1954). Based on the assumption that the scale midpoint represents the true mean level of performance, a shift of average ratings away from the scale midpoint is interpreted as rating error. In fact, it is not uncommon that more than 80% of the ratees are evaluated as "above average" in job performance (Thomas & Bretz, 1994). Numerous studies have demonstrated that self-evaluations are likely to show particularly high levels of leniency bias as compared to supervisory or peer evaluations (e.g. Fox, Caspy, & Reisler, 1994; Fox & Dinur, 1988; Somers & Birnbaum, 1991). An early illustrative example of overly optimistic self-ratings was reported by Meyer (1980) who found that about forty percent of respondents placed themselves into the top ten percent category with regard to their job performance. For the purposes of the current study, leniency in self-ratings is assessed by analyzing the mean difference between self and supervisor
ratings. Self-ratings are considered lenient to the degree that self-ratings are more favorable than supervisory ratings. Of course, ratings made by raters other than the self can also be lenient. Supervisors, for instance, might inflate ratings of their subordinates for reasons such as guaranteeing resources and rewards, or maintaining good relations (Murphy & Cleveland, 1991). The fact that supervisors may also inflate ratings will be left out of consideration. The study’s research question is whether self-ratings reach higher levels of favorability, even though leniency might be a general rater tendency in performance evaluation.

1.3 The importance of self-other congruence in performance ratings

Both correlation coefficients and mean difference scores measure the convergence of ratings made by different raters. These measures of convergence can be studied for various reasons. I suggest that there exist two main issues, namely the validity of performance ratings in general, and the consequences self-other discrepancies have for the management of employees’ job performance.

Outside of laboratory settings, true performance scores are usually unavailable. Thus, agreement among different sources of ratings (supervisors, peers, self, etc.) has been used as a proxy for the validity of ratings (e.g. Fox & Dinur, 1988; Tsui & Ohlott, 1988). Underlying this view is the notion that an (unknown) "true performance" score exists for each ratee, and as raters fail to agree on the performance for individual ratees, raters’ judgments lack validity. As will be reviewed in the next section, research has repeatedly reported rather low levels of between-source rater agreement in performance appraisals. From this, Tsui and Ohlott (1988) concluded that what they dubbed the "low agreement phenomenon" challenged the validity of performance ratings, and implied at least that ratings from different sources could not substitute for each other. Ratings reflect informational and motivational distortions made by all raters from various perspectives - even though the same performance constructs are being assessed (Facteau
& Craig, 2001). However, it is also possible that ratings from multiple perspectives contribute unique (valid) information (e.g. Murphy & Cleveland, 1995) and that discrepancies reflect real differences in a person’s behavior on the job towards different constituencies. Taken together, a combination of ratings from various sources may then be a more valid measure of performance as the resulting measure covers the criterion domain more completely (Atkins & Wood, 2002). The present analysis will not be able to reach a decision as to which proposition is the more appropriate one. But investigating the amount of agreement found between rating sources as well as the effects of various moderators can contribute to our knowledge about the causes of the "low agreement phenomenon".

Beyond general implications that between-source rater agreement has for the validity of performance appraisals, self-other agreement may warrant specific consideration due to the effects it has on the targets’ attitudes and behavior. Some recent research has focused on effects of self-other convergence in ratings. Bass and Yammarino (1991) established the hypothesis that the degree of self-other congruence in ratings may be directly related to job performance. Their hypothesis has been confirmed by a number of authors (Atwater, Ostroff, Yammarino & Fleenor, 1998; Atwater & Yammarino, 1992; Furnham & Stringfield, 1994) and triggered a debate whether self-other congruence itself should be used as an assessment measure in predicting effectiveness on the job (e.g. Fletcher & Baldry, 2000; Randall, Ferguson & Patterson, 2000). A related idea is that making a person aware of discrepancies between his or her self-view and the view of others can help the targets increase their effectiveness on the job. This is in fact a basic assumption underlying most multi-source feedback procedures (Leslie & Fleenor, 1998).²

Paralleling these findings, research has accumulated evidence which suggests that discrepant

²Outside of research in I/O-psychology, the amount of congruence has also been related to psychological adjustment (Kwan, John, Kenny, Bond, & Robins, 2004). Within an experimental setting, Kwan et al. (2004) found that self-enhancement (e.g. leniency) was positively correlated with self-esteem but negatively related with task performance.
feedback may lead to negative outcomes. "Discrepant feedback" is used here to refer to feedback that results if others’ ratings are less favorable than self-ratings, or if the pattern of high and low ratings diverges from self-reports. Discrepant feedback can lead to unwanted consequences including negative beliefs about the accuracy and usefulness of ratings as well as generally negative affective reactions (Brett & Atwater, 2001), less satisfaction with the appraisal, lower acceptance of evaluation procedures (Brett & Atwater, 2001; Farh, Werbel & Bedeian, 1988; Halperin, Snyder, Shekel & Houston, 1976), or reduced willingness to participate in career planning (Wohlers, Hall & London, 1993).

2 Previous research

Two meta-analyses have previously studied the convergence of self and supervisor ratings for performance (Conway & Huffcutt, 1997; Harris & Schaubroeck, 1988). A third meta-analysis examined the agreement of self-evaluations with a variety of performance criteria (Mabe & West, 1982). Note, that the latter study is not easily comparable to that of the others mentioned, as well as to the present one. Mabe and West (1982) included both various criteria for performance - other than ratings - and highly diverse settings in which performance was assessed. But as this study has received much attention in the literature on self-ratings of job performance (e.g. Murphy & Cleveland, 1991; Viswesvaran & Ones, 2000), these authors’ findings will be discussed as well.

2.1 Previous meta-analytical results

Validity coefficients. Meta-analytical results have confirmed the notion that correlations of self-ratings and ratings made by others are rather small in magnitude. This becomes more apparent if self-other correlations are compared to coefficients obtained for ratings made by two groups of observers. Conway and Huffcutt (1997) reported a mean (corrected) correlation between self and
supervisor ratings of .22 (.31), and a similar correlation between self and peer ratings .19 (.31). However, supervisor and peer ratings yielded a correlation of .34 (.79). An earlier meta-analysis conducted by Harris and Schaubroeck (1988) reported self-supervisor and self-peer correlations of .22 (.35) and .24 (.36), respectively. But coefficients for supervisor and peer ratings reached the substantially higher magnitude of .48 (.62).

*Leniency.* The only meta-analysis that has examined leniency in performance ratings reported a mean weighted difference between self- and supervisor-ratings of $d = .70$ (Harris & Schaubroeck, 1988). The authors interpreted this difference score as a measure of leniency in self ratings and thus contributed to the general notion that self ratings are lenient. In fact, Harris and Schaubroeck (1988) found self ratings to be more than half a standard deviation higher than supervisory ratings.

Moreover, given the small number of samples (18 independent studies) included in this study, meta-analytical evidence for leniency in self-ratings is comparatively weak. The current study provides a meta-analytical estimate of index $d$ based on a larger set of studies.

### 2.2 Previous results regarding moderators of self-supervisor agreement

The following section presents an overview of variables that have been suggested as moderators of self-supervisor agreement in previous meta-analytical research. The present meta-analysis builds upon these authors’ thinking especially in choosing moderator variables. But it will also be pointed out how the present analysis differs from the approaches taken by previous authors.

*Harris and Schaubroeck (1988).* The meta-analysis conducted by Harris and Schaubroeck (1988) found a (corrected) correlation between self and supervisory ratings of job performance of $r = 22 (.35)$. These authors examined three moderator variables of rater agreement including *job-type*, and two variables pertaining to the format of ratings: *dimensional* as opposed to *global*
performance dimensions, and *behavioral* as compared to *trait-like* scale labels (see Table 1). From these three variables, only job-type proved to have effects as a moderator of (correlational) agreement among raters. Correlations were lower for managerial and professional than for blue-collar and low-skilled service jobs.

Harris and Schaubroeck (1988) were the first to report meta-analytical results for mean difference scores calculated between self and supervisor ratings, interpreting this difference score as a measure of leniency. In fact, Harris and Schaubroeck (1988) found self ratings to be more than half a standard deviation higher than supervisory ratings ($d=.70$).

The present meta-analysis goes beyond the work of Harris and Schaubroeck (1988) in that it not only provides an overall estimate of leniency in self ratings, but examines a set of hypothesized moderators. Harris and Schaubroeck (1988) could not perform moderator analyses on leniency in self ratings as too few primary studies could be located that provided information on difference scores between self and supervisor ratings. As the database of primary research has grown substantially since 1988, an examination of factors that moderate leniency in self-ratings should now be possible.

Conway and Huffcutt (1997). A more recent study of multi-source performance ratings reported estimates of rater convergence among self, supervisor, peer, and subordinate appraisals (Conway & Huffcutt, 1997). These authors analyzed both within-source and between-source correlations. Thus they provide estimates of (within-source) rater reliabilities as well as estimates of (between-source) validity for ratings made from different perspectives. In line with Harris and Schaubroeck (1988), Conway and Huffcutt found self-other correlations to be rather low, ranging from .14 to .22. Concerning self and supervisor ratings, the authors also reported results for three moderator variables. As in earlier research, job-type moderated self-supervisor agreement: Correlations were lower for managerial and jobs of higher complexity.
Mabe and West (1982). The earliest meta-analysis of self-evaluation in performance appraisal was published in 1982. Mabe and West examined the validity of self-assessments analyzing the correlation between self-evaluations and criterion measures of performance. Criterion measures included objective tests, grades, peer- or supervisor-ratings, as well as job turnover. This way, the authors obtained an estimate of validity for self-evaluations of $r=.31$.

Moreover, Mabe and West (1982) presented a theoretical rationale for testing a rather comprehensive set of factors that may moderate the relationship between self-evaluation and relevant criteria. They grouped these factors into two broad categories: person variables and measurement conditions (see Table 1 for an overview). Altogether, Mabe and West (1982) examined twelve potential moderators of self-evaluation validity. Four of these accounted for most of the variance in the validities of self-evaluations: the expectation that self-assessments will be validated, social comparison instructions, experience with self-evaluation, and instructions of anonymity. The current analysis examines all four of these moderators again. But compared to the meta-analysis conducted by Mabe and West (1982), the current study focuses on a more narrowly defined body of primary research.

To this day, the meta-analysis by Mabe and West (1982) is cited in many textbooks (e.g. Cascio, 1998) and in most of the research articles in the field of performance evaluation, whenever self-assessments are discussed. Re-analyzing their hypotheses may be worthwhile for two reasons. First, the database of primary research has grown since 1982. Second, the set of studies that Mabe and West analyzed makes it questionable whether their results are applicable to the area of job performance. Mabe and West integrated findings for various subject populations (including college students (81%), managerial samples, clerical workers, and psychiatric patients), as well as various performance categories (ranging from scholastic ability, technical or physical skill, to intelligence or job interview performance). In addition, criterion measures comprised objective tests, grades, and ratings.
Table 1: Moderator variables examined by previous meta-analyses

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject populations</td>
<td>various (student, managerial, and blue-collar samples)</td>
<td>job incumbents</td>
<td>job incumbents</td>
</tr>
<tr>
<td>Criterion measures</td>
<td>various (tests, grades, ratings)</td>
<td>ratings (self, supervisor, peer)</td>
<td>ratings (self, supervisor, peer, subordinate)</td>
</tr>
<tr>
<td>Person variables</td>
<td>High intelligence</td>
<td>High achievement status</td>
<td>Internal locus of control</td>
</tr>
<tr>
<td></td>
<td>Job type *</td>
<td>managerial/professional vs. blue collar/service</td>
<td></td>
</tr>
<tr>
<td>Conditions of measurement</td>
<td>Match of criteria</td>
<td>Dimension vs. global ratings</td>
<td>Dimension type</td>
</tr>
<tr>
<td></td>
<td>Performance vs. ability</td>
<td>Trait-labels vs. behavioral scales</td>
<td>Interpersonal vs. cognitive dimensions</td>
</tr>
<tr>
<td></td>
<td>Past vs. future performance</td>
<td>Social comparison *</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Social comparison *</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluation motivation</td>
<td>Anonymity instructions *</td>
<td>Expectation of validation *</td>
<td>Evaluation experience *</td>
</tr>
<tr>
<td>Overall</td>
<td>( \tau(r_p) )</td>
<td>.31</td>
<td>.22 (.35)</td>
</tr>
<tr>
<td></td>
<td>( \bar{d} )</td>
<td>–</td>
<td>.70 *</td>
</tr>
</tbody>
</table>

Note. \( \tau = \) overall sample size weighted mean correlation of self and supervisor ratings; \( \tau(r_p) \) = correlation coefficient corrected for attenuation; \( \bar{d} \) = overall sample size weighted mean difference score; * = authors reported significant effect.

* Harris & Schaubroeck (1988) did not clearly state whether the effect size \( d \) they reported had been corrected for attenuation.

The present meta-analysis is confined to samples in which both self-evaluation and criterion measures consist of ratings for performance. It is limited to a specific context, namely questionnaire-based ratings of performance on an actual job. The present study includes moderators that have been examined before, but also contributes to the current body of research by examining an extended set of variables. A summary of all moderator variables that have been included in previous meta-analyses appears in Table 1, together with the overall estimates of self-supervisor agreement reported in these articles.


3 Hypotheses

The following section presents an outline of all the hypotheses that are examined. In accordance with the study’s primary purpose, the hypotheses regard variables that potentially moderate the amount of agreement found in self and supervisor ratings. Compared to earlier meta-analytic research (Mabe & West, 1982; Harris & Schaubroeck, 1988) the current moderator analyses are based on a considerably larger set of samples. Drawing on this database, moderator variables are re-examined that have been suggested by previous meta-analysts. Moreover, the current review introduces additional moderator variables that were not included in previous meta-analytic research.

The moderator variables studied can be grouped into three sets. The first group of variables pertains to characteristics of the report format and the kind of performance constructs that ratings are made for. The second group refers to contextual conditions under which ratings are made. Finally, a third set of variables is related to the composition of the respondents under study (see figure 1).

The mechanisms underlying these moderator effects are complex. Thus, I will not make an effort to present an extended theoretical background to derive each of the hypotheses on moderator variables, nor to discuss previous empirical findings in detail. But I present a short rational for the examination of each of the moderator variables along with an interpretation of its effect. Note, that moderator variables were specified before the analysis was conducted to avoid capitalization on chance concerning the testing for moderator effects. In a later section, a process model of performance appraisal will be presented to integrate findings and discuss the theoretical mechanisms underlying rater convergence.

Of course, overall estimates of effect sizes will be reported. In addition to estimating the overall correlation of self and supervisor ratings, the average standardized difference in the level
of ratings will be calculated (if self-ratings are more favorable than supervisory ratings they are considered to be "lenient"). Based on the results reported by Harris and Schaubroeck (1988), self-ratings are predicted to show significantly higher levels of favorability than supervisory assessments.

### 3.1 Report format

The hypotheses discussed in the following section regard properties of the scales (item aggregation, behavioral scale definitions, social comparison scale labels, non-judgmental performance indicators) and the performance constructs, as they are used in rating instruments (global vs. dimensional performance constructs, task- vs. contextual performance, and the use of traits as scale labels).
Single-item vs. aggregated measures. Psychometric theory predicts that instrument length, i.e. the number of items integrated into a final measure, should be positively related to the reliability, and thus the validity of ratings. If raters provide repeated evaluations by rating greater numbers of items, random response error should decrease. In addition, integrating more items into an aggregate measure may reduce conceptual disagreement between raters. Testing this prediction is of interest as some researchers have suggested that it does not necessarily hold for ratings of job performance (Viswesvaran et al., 1996). Thus, this study attempts to provide a meta-analytical estimate to answer the question whether instrument length positively influences (correlational) rater-agreement for self-supervisor ratings. Reduced mean difference scores for aggregate measures can be expected due to statistical regression. As this is a purely probabilistic phenomenon, it will not provide much insight into self-evaluation and will not be discussed further. The following hypothesis will be tested for correlation coefficients:

Hyp. 1: Aggregate ratings show higher validity coefficients than single-item measures.

Broad vs. behavioral scale definitions. One distinction of scale labels that previous research has examined is that of trait-like and behavioral labels (Harris & Schaubroeck, 1988). I prefer to use different terminology to make a similar distinction. I distinguish broad competency labels from behaviorally defined scales. (The term "trait" will later be used to refer to traits in the sense of personality characteristics). Broad labels comprise broad or somewhat ambiguous constructs such as "leadership ability" or "quality of work". In contrast, behavioral scale definitions include items that provide concrete descriptions of behavior (e.g. "I get subordinates’ ideas before I make major decisions"), as well as behaviorally anchored rating scales (see Cascio, 1998, p. 71-73, for an example). The rationale behind testing this distinction is as follows: If scale definitions consist of a broad label rather than descriptions of behavior, raters have to rely more heavily on their idiosyncratic beliefs about relevant behavior as clues for high or low performance. Thus, these ratings should involve higher levels of inference and yield lower agreement between self
and supervisor ratings:

Hyp. 2a: Behaviorally defined rating scales are associated with higher validities than are scales that have respondents rate broad constructs.

I also expect the use of broad labels as compared to behaviorally defined items to affect leniency in self-ratings. To the extent that items do not provide concrete explanations of the targeted performance construct, raters are free to apply their own interpretations of the scales’ content. Assuming that self-raters are motivated to maintain a positive view of the self, self-ratings of non-well-defined constructs may yield even higher favorability in self-ratings (Farh & Dobbins, 1989). Moreover, differences in the frame of reference between self and supervisor ratings may decrease if behaviorally defined items are used. This should lead to decreased leniency in self-ratings for behaviorally defined scales.

Hyp. 2b: Self-ratings on behaviorally defined scales show reduced leniency compared to self-ratings of broad constructs.

Social comparison. Another approach to understanding the validity of self-reports draws on the idea that self-raters actually lack a basis for making accurate judgments. More specifically, self-raters may not have information about relevant comparison standards, or choose a comparison group other than the one intended by the designers of the performance appraisal system. As Mabe and West (1982) argued, performance evaluation involves a relative measurement process rather than an absolute one. Individuals have considerable freedom to choose or construct a reference target or reference group, though. For example, when asked to rate their job performance, individuals may select such a frame of reference that helps them to control the psychological implications that social comparisons have. Such appraisals may be based on comparisons between occupational communities rather than with other members of the same occupational community. It is not necessarily colleagues and coworkers that job incumbents compare themselves with.
when evaluating their performance. Moreover, job incumbents can even avoid social comparisons. When asked to rate their job performance, job incumbents may report how they feel about their achievements in general. Thus, asking subjects to consider their standing relative to a concrete comparison group (e.g. coworkers) is likely to increase agreement between self-raters and others. Mabe and West (1982) suggested two response mode properties that should induce subjects to take a social comparison view: 1) wording performance items in relative terms ("below average", "average", "above average" ...) as opposed to absolute terms ("poor", "satisfactory", "excellent" ...), and 2) an explicit instruction to think of a particular comparison group when making performance evaluations.

**Hyp. 3a:** Validities are higher if scales are defined in relative terms or provide an explicit social comparison instruction.

**Hyp. 3b:** Leniency in self-ratings decreases if scales are defined in relative terms or provide an explicit social comparison instruction.

**Global / overall vs. dimensional ratings.** Asking respondents to report an evaluation of the targets’ overall performance rather than performance in specific domains has been suggested as a moderator of conceptual disagreement between raters. In fact, Harris and Schaubroeck (1988) included this distinction in their meta-analysis on performance ratings. However, they could not confirm it as a moderator. In a similar vein, using estimates of coefficient alpha to assess reliability, Viswesvaran, Ones, and Schmidt (1996) found that supervisors and peers rated overall job performance more reliably than any other performance construct. The authors suggested that global performance constructs might generally be more reliably rated than narrower constructs (i.e. specific dimensions). Following these authors, I also examine the distinction of global and dimensional ratings. However, the present study’s approach differs in that it tries to control for the effect of form length. Many of the "overall performance" ratings that are reported in the literature are aggregate measures, i.e. sums of items across different job-performance dimensions.
But there are also single-item ratings of "overall performance". I seek to disentangle the effects of aggregation and of the performance construct rated. Thus, single-item ratings of overall performance are compared with single-item measures for specific dimensions of job-performance. Therein, I predict ratings of global performance to be associated with higher agreement in self-supervisor ratings. The underlying assumption is that ratings of global performance may implicitly eliminate part of the conceptual disagreement between rating sources. Even if raters disagree on an individual’s performance in specific domains, they may still have a similar impression of the target’s overall performance.

Hyp. 4a: Ratings of global / overall performance as compared to dimensional ratings yield higher validity coefficients.

The distinction of ratings of overall job performance as opposed to performance in specific performance domains will also be examined as a potential moderator of leniency in self-appraisals. Given that individuals are generally motivated to maintain a positive view of the self (Jones, 1990), an unfavorable evaluation of overall performance might be more threatening to the self than lower ratings in specific performance dimensions. As a result, individuals might overrate their overall performance more than performance in specific domains:

Hyp. 4b: Ratings of global / overall performance as opposed to dimensional ratings yield higher leniency in self-appraisals.

Task- vs. contextual performance vs. ratings for traits. Mainly referring to the work of Borman and Motowidlo (1993) as well as Motowidlo and van Scotter (1994), contextual performance will be distinguished from task-performance and examined as a moderator of self-supervisor agreement. The distinction contrasts activities that are prescribed by organizational role requirements with activities that are also important to the organization but are usually discretionary. Task performance includes meeting standards of task proficiency for which explicit expectations exist. Performance in this domain is directly related to the "technical core" of an organization. In
contrast, contextual performance includes behavior that implies going beyond explicitly defined role requirements by performing such actions as helping and cooperating with others, making suggestions for improvement, or presenting the organization favorably to outsiders. Behavior that is related to contextual performance supports the social environment in which regular task-performance takes place. Motowidlo and van Scotter (1994) demonstrated that task performance and contextual performance contribute independently to judgments of an employee’s overall contribution to an organization.

The moderator hypothesis that is suggested here is based on the following assumptions. Task performance includes explicit expectations that are held toward a job incumbent. This may imply that job incumbents receive more feedback concerning their task-proficiency. Consequently, job incumbents have opportunity to develop a comparatively elaborated understanding of their own task performance, but may lack a similarly differentiated understanding of their behaviors in the domain of contextual performance. Supervisors, too, may assess contextual performance of their subordinates less reliably. For example, Conway (1999) reported that supervisors tended to pay less attention to contextual performance in targets than colleagues did, who gave equal weight to both performance dimensions. One reason for this may be that task performance is simply more relevant to supervisors. Besides, supervisors might have less opportunity to observe behavior in their subordinates that is related to contextual performance.

Hyp. 5a: Correlations of self and supervisory ratings are higher for task performance than for contextual performance.

The distinction of task- and contextual performance may also moderate leniency in self-evaluations. The rationale behind this assumption is as follows: research has provided some evidence that the perceived importance of rating dimensions is related to leniency in self-reports (e.g. Fox, Caspy & Reisler, 1994 ). In work settings, task performance is generally likely to be of higher importance than contextual performance – to employees as well as to supervisors (Con-
way, 1999). This difference in perceived importance should cause ratings of task-performance to correspond with higher levels of leniency.

Hyp. 5b: Self-ratings of task-performance show higher leniency compared to ratings of contextual performance.

Another relevant distinction regarding the kind of performance construct rated is the evaluation of traits, i.e. of performance constructs that refer to qualities of personality. Examples for such performance dimensions include "outgoingness", "creativity", "self-confidence" and "action orientation". These scale labels represent traits as they refer to characteristics of a person that are rather stable over time (although they are not measured in performance appraisals as scientific psychology would do to define traits), and are also likely to be thought of as "traits" by laymen. The reason for distinguishing traits as scale labels from other performance constructs is at least twofold: ratings for traits involve rather high levels of inference (Harris & Schaubroeck, 1988), and their relative stability over time renders ratings for traits a special case. The greater stability of traits is especially relevant to self-ratings, as it makes trait ratings prone to self-deceptive responding. Selective recall and weighting of information that is in accord with the self-raters’ self-concepts is likely to be pronounced if the qualities rated are stable over time and not specific to certain situations (such as the work context). Thus, ratings for traits should yield rather low self-supervisor correlations, and, unlike contextual performance, high levels of leniency.

Hyp. 5c: Self-ratings of traits (personality characteristics) yield lower correlations with supervisor ratings as compared to ratings of contextual and task performance.

Hyp. 5d: Self-ratings of traits (personality characteristics) show higher leniency compared to ratings of task- and contextual performance.

(Non-)judgmental performance indicators. Indicators of performance can be grouped into two categories: judgmental (verbal, more or less abstract descriptions of behavior, or performance
constructs such as "leadership"), and non-judgmental (e.g. time to complete a task, financial indicators, production output). Although all ratings are judgments I will call them "non-judgmental" if they refer to an "objective output" criterion. Researchers have reported ratings for both judgmental and non-judgmental performance indicators. In a study by Farh, Werbel and Bedaian (1988), for instance, research faculty provided self-ratings for seven areas of performance, one of which read "journal publications" on a scale ranging from "poor" to "outstanding". Schrader and Steiner (1996) reported ratings for dimensions such as "monthly transaction rate", "total sales", or "hourly productivity" on a nine-point scale ranging from "very poor" to "very good". In cases like these, ratings are made for performance dimensions that are based on explicit indicators of performance. Their use may yield higher rater congruence for at least two reasons. First, this procedure reduces conceptual disagreement among raters. Supplied with the information what indicators of performance are to be rated, raters do not have to identify indicators of performance themselves. Thus ratings depend less on idiosyncratic beliefs about what constitutes good performance. Second, the existence of objective criteria as a basis for ratings renders ratings potentially verifiable. At least, these ratings may appear more verifiable to raters. Thus, respondents should be more reluctant to distort ratings as this involves the risk of losing face through a potential (or imagined) validation of ratings.

Hyp. 6a: If ratings use non-judgmental performance indicators, convergence of self-supervisory ratings is higher.

Hyp. 6b: If ratings use non-judgmental performance indicators, leniency in self-ratings decreases.

3.2 Conditions of report

The rating instrument is not the only factor that influences rater behavior. Raters in organizations are very likely to consider the context that ratings are made in as they take into account the
consequences their ratings may have. Thus, the section below studies features regarding the conditions under which reports are made. Among those examined are the confidentiality of ratings, respondents expecting a validation of their evaluations, and rating purpose.

Confidentiality of ratings. A situation of public self-presentation exists in performance appraisals if individuals who rate their own performance know that these ratings may be viewed by others. Assuring confidentiality to the targets is a clear means of reducing self-presentational concerns in raters (the very reason for which confidentiality is often assured). Confidential or even anonymous ratings should provide fewer incentives to intentionally manipulate self reports in order to generate favorable impressions in others. In short, confidentiality should reduce disagreement in self-supervisory ratings:

Hyp. 7a: Assuring confidentiality to respondents will lead to higher self-supervisory correlations.

Hyp. 7b: Assuring confidentiality to respondents will lead to reduced leniency in self-ratings.

Validation expectation. Mabe and West (1982) found the expectation that self-reports may be validated to be among the four variables that best predicted correlational agreement between self-appraisal of performance and other criteria, such as tests or performance ratings. Regarding leniency in self-ratings, empirical evidence exists that shows how a possible disconfirmation of self-descriptions decreases the favorability of self-ratings. As an example, Aitkenhead (1984) exposed participants to a number of subsequent trials on an alleged intelligence test, and found that subjects were more modest in rating their expectations for their future performance when they expected further tests to follow than when they knew that no further tests were to follow. The author concluded that individuals were careful to avoid disconfirmation of over-optimistic self-assessments in the face of anticipated further feedback. In line with this argument, subjects
in Aitkenhead’s study enhanced their self-presentations when there was no possibility that ratings could be proved to be immodest.

The current study assesses whether expecting a validation has an effect on correlation coefficients, and whether a leniency-effect emerges due to the expectation that self-ratings will be validated. In the context of performance appraisal, ratings may be validated either by comparison to other data (such as tests or outcome measures) or socially; for example via a feedback meeting in which supervisor and subordinate discuss their appraisals.

Hyp. 8a: If self-raters expect their evaluations to be validated, self-supervisor correlations increase.

Hyp. 8b: If self-raters expect their evaluations to be validated, leniency in self-ratings decreases.

**Rating purpose.** Among the contextual factors that should influence self-supervisory agreement in performance appraisals, the most prominent is the purpose for which ratings are collected. The present study distinguishes three rating purposes or settings in which performance ratings are made: performance appraisal for administrative use, developmental feedback, and data being gathered for research purposes. I suggest that the three appraisal conditions are differently prone to cause rater bias.

The oldest and still most predominant use of performance appraisal is as a basis for administrative decisions in organizations, such as salary increases and promotions, merit-pay, or termination. Unfortunately, data from performance ratings for administrative reasons are often not accessible to researchers. Although it is still likely that only a small number of published studies exist which report data gathered for administrative reasons, I will try to compare administrative-based performance ratings to those gathered in other settings. Administrative-based ratings most likely motivate raters to intentionally distort their ratings (Murphy & Cleveland, 1995), which is likely to have consequences for both indicators of rater agreement: deliberately manipulating
reports should lower correlations between self and supervisor ratings and leniency in self-ratings is likely to increase.

Hyp. 9a: Validity coefficients are lower if ratings are made for administrative as compared to developmental or research purposes.

Hyp. 9b: Self-assessments are more lenient if they are made for administrative as compared to developmental or research purposes.

During the second half of the last century, a new trend in performance appraisal emerged (Murphy & Cleveland, 1995). Especially throughout the 60s and 70s, practitioners in organizations began to introduce the use of performance evaluation for developmental and feedback purposes (Murphy & Cleveland, 1995). If appraisals are part of developmental feedback sessions, low congruence between self-assessments and those of others can be perceived as undesirable outcomes for several reasons. Thinking of feedback sessions as evaluative social situations, low congruence in ratings may be perceived as a potential loss of face. A possible reaction then is that individuals present themselves in rather modest ways, or even try to guess others’ views of themselves. In the more positive case, job incumbents may make highly serious efforts to report beliefs they truly hold about themselves as accurately as possible. "Accurate" self-reports may appear as a prerequisite for developmental feedback procedures to become a learning opportunity. Either of the two strategies should result in higher levels of congruence between self and supervisory ratings - at least as compared to ratings made for administrative purposes.

Finally, quite a number of studies are conducted for research purposes only. I coded studies to be research based if the authors stated that ratings were not determined for any other organizational use. If results were reported back to the organization at all, then in aggregate formats only (i.e. at the level of group statistics). Collecting ratings for "research purposes only" is often considered to reduce rater bias, as appraisal outcomes are hardly of any consequence for raters.
But the raters’ accountability for their ratings decreases and may in turn lead to less accurate ratings (London, Smither, & Adsit, 1997). In sum, whereas I clearly expect that administrative ratings will be the least accurate (highest levels of leniency, lowest correlational agreement), predicting differences between developmental and research-based ratings is difficult. Whereas a few studies have compared administrative and research-based ratings (e.g. Harris, Smith, & Chamgagne, 1995), to my knowledge, ratings for developmental purposes have not been included in such comparisons. Thus, the examination of developmental purpose as a setting for performance appraisals is rather exploratory, and the following hypothesis is rather unspecific.

Hyp. 9c: Developmental performance-ratings yield similar psychometric properties as do research-based ratings.

360-Degree-Feedback. The term 360-degree-feedback refers to appraisal procedures that include feedback from the full circle of an employee’s points of contact, typically the supervisor, colleagues, and subordinates. Moreover, the inclusion of self-ratings is also a core feature of 360-degree feedback instruments (Fletcher, Baldry, & Cunningham-Snell, 1998; Fletcher & Perry, 2001). The underlying rationale is to use performance feedback as a means for increasing managerial effectiveness through self-insight in developmental needs. Leslie and Fleenor (1998) have reviewed 24 multi-source management-assessment instruments and report that 22 of these explicitly emphasize self-other discrepancies in ratings which are documented and fed back to the targets.

In a rather exploratory way, I will compare 360-degree feedback procedures to appraisals that included fewer groups of raters. The underlying rationale is that 360-degree procedures might create situations that are conducive to increased accuracy in self-assessments. 360-degree feedback procedures are characterized by a common general goal as well as procedural similarities that explicitly try to serve the purpose of employee development and self-insight among the target persons (Leslie & Fleenor, 1998). With reference to theories of self-awareness and self-
knowledge, it has been suggested that these procedures have the potential to increase inter-rater congruence (London & Smither, 1995). London and Smither (1995) argued that 360-degree-feedback procedures may be especially suited to introduce normative ideas about leadership and management behavior. In addition, rating performance constructs repeatedly and applying them to others as well as to the self directs attention to relevant behavior. To test the hypothesis that these mechanisms are really able to promote rater congruence, I will compare 360-degree-feedback results to those from other appraisal settings. As 360-degree feedback is mostly applied to evaluate the behavior of managers, its effects will be analyzed within managerial samples to ensure the comparability of results.

Hyp. 10a: 360-degree-feedback procedures are associated with higher validities as compared to "traditional" appraisal settings (for managerial samples).

Hyp. 10b: 360-degree-feedback procedures are associated with reduced leniency as compared to "traditional" appraisal settings (for managerial samples).

3.3 Sample composition

The following paragraphs discuss characteristics of the respondents that potentially influence self-supervisor congruence. These include characteristics of the job or position that respondents occupy, the respondents’ gender, and their cultural background. Again, as the mechanisms underlying these effects are complex, an extended theoretical background cannot be presented.

*Job-type.* Harris and Schaubroeck (1988) reported higher correlations among self- and others’ ratings for blue-collar and service jobs than for managerial and professional jobs. Conway and Huffcutt (1997) distinguished managerial from non-managerial samples. In addition, these authors categorized samples according to job complexity. Both variables moderated correlational agreement in ratings: managerial samples yielded lower validities than non-managerial samples,
and higher complexity, too, was associated with substantially reduced interrater agreement. I will again scrutinize whether both the distinction of blue-collar versus white-collar jobs and that of managerial versus non-managerial jobs moderate self-supervisory correlations. An important underlying assumption is that the unequivocality and the number of available cues for good performance are likely to differ between blue-collar and white-collar jobs. This can readily explain differential levels of rater agreement between blue- and white-collar jobs. The same might be true for managerial vs. non-managerial jobs, although this distinction is a very general one.

I will also analyze the moderating effect of the job-incumbents’ educational background, assuming that educational level is a proxy variable for job complexity. In sum, I will distinguish 1) blue-collar versus white-collar samples, 2) high level (more than 80% of respondents hold graduate degrees) versus lower level of education in the samples, and 3) managerial versus non-managerial samples. This makes it possible to disentangle the effects of managerial position and educational level. A hierarchical analysis can determine whether the distinction of managerial and non-managerial samples still yields a moderator effect if only samples of comparable educational level are analyzed. I expect both an effect of educational level and of managerial samples. In addition, I hypothesize that self-other convergence is specifically low for managerial samples.

In an exploratory way, I also test whether these distinctions in job-type moderate the amount of leniency in self-ratings.

Hyp. 11: If respondents are sampled from blue-collar jobs, validities of self-reports are higher than if respondents are sampled from white-collar jobs.

Hyp. 12: If respondents have higher levels of education, validities of self-reports decrease.

Hyp. 13: Validities are lower for managerial samples as compared to non-managerial samples even when level of education is controlled.

**Gender composition.** Research suggests that females both overrate their performance less and reach higher levels of profile agreement with others’ ratings, than is the case in men. Regarding
index $d$ of agreement, Wohlers & London (1989) reported that female managers’ self-ratings tended to be lower than those of their male colleagues. Moreover, this sample of female managers rated themselves lower than they were rated by their immediate supervisors. (As the subgroup sample sizes were small, the study’s results for men and women were not reliably different). Experimental research on the self-evaluation of ability has provided some evidence supporting the hypothesis that females’ self-assessments are often lower than men’s (e.g. Pallier, 2003).

In addition, research suggests that profile agreement (index $r$) in self-other ratings is higher for females than for males. For example, London and Wohlers (1991) found that the correlation between self-ratings and the average subordinates’ ratings was higher for female than for male managers. As London and Wohlers (1991) argue, women tend to be more concerned about interpersonal relationships and strive for attachment as well as for achievement in their careers (Gallos, 1989). Social psychology has more generally promoted the notion that women’s self-construal differs from men’s. In particular, interpersonal relations are likely to be of greater importance to women’s self-construal than to men’s (Avsec, 2003; Kashima, Yamaguchi, Kim, Choi et al., 1995). Females pay more attention to social cues and other’s views of themselves and, as Fletcher (1999) suggests, are more likely to seek feedback and to accept it than is the case in men. If this is true, women’s self-perceptions should reach higher levels of agreement with the views of others.

Hyp. 14a: Samples including high percentages of female self-raters show higher validities as compared to predominantly male samples.

Hyp. 14b: Samples including high percentages of female self-raters show reduced leniency as compared to predominantly male samples.

Culture. In societies that emphasize collectivism, differentiation of rewards are considered detrimental to norms of group harmony. In contrast, equity as distribution rule allocates rewards according to individual contributions within a group, a rule that is typically well accepted in in-
individualist societies. Countries identified as high in individualism by Hofstede (1983) are more likely to use an equity rule in allocating rewards. For example, Kim, Park, and Suzuki (1990) concluded that the equity norm for three countries which differ in their emphasis of individualism (USA, Japan, Korea) also differed in their use of an equity norm in allocating rewards within groups. Collectivism is also negatively correlated with preferences for merit-based promotion systems (Ramamoorthy & Carroll, 1998). As Aycan and Kanungo (2001) state, "in individualistic cultures, self-serving bias in performance evaluations occurs more frequently than 'self-effacement' or 'modesty bias', while the reverse holds true for collectivist cultures" (p. 399). This should imply that differences in individual contribution are deemphasized rather than emphasized. In the following, a broad distinction of Western and Asian/Eastern cultures is made to capture the influence of collectivism and individualism, respectively. (A more differentiated discussion of different cultures is beyond the scope of this thesis. Besides, the number of samples that stem from single countries - other than the US - is likely to be rather small). With respect to performance ratings, Farh et al. (1991) were the first who presented empirical evidence that a modesty-effect in self-appraisals may be observed in Asian societies:

Hyp. 15: Samples from Asian countries show modesty instead of leniency in self-ratings.
4 Method

4.1 Literature search procedures

Several literature search procedures were used to retrieve both published and unpublished primary studies. The first strategy involved examining the reference sections of previous research reviews (Conway & Huffcutt, 1997; Harris & Schaubroeck, 1988; Hoyt & Kerns, 1999; Mabe & West, 1982; Viswesvaran et al., 1996). Second, a computer search of the PsychINFO database was conducted. Keywords such as performance appraisal and job performance were combined with the terms self-appraisal, self-evaluation and ratings to identify potentially relevant studies. Third, the Educational Resources Information Center database (ERIC) and the Dissertation Abstracts Database (UMI ProQuest) were searched using the same keywords as those used for the PsychLit search. Fourth, a manual search of the following journals was completed: Personnel Psychology, Academy of Management Journal, Journal of Applied Psychology, Journal of Occupational and Organizational Psychology, Organizational Behavior and Human Decision Processes, International Journal of Selection and Assessment, and Human Relations. Finally, the reference sections of relevant studies identified in previous searches were examined for additional references. The literature search was concluded in December 2003, and research articles that appeared later than this were not considered.

4.2 Inclusion criteria and the coding of studies

Inclusion and exclusion criteria. To be included in the current analysis, research reports had to meet the following criteria: (1) Research reports contained a quantitative measure of agreement for self and supervisory ratings - either a correlation coefficient, or mean levels of ratings together with standard deviations. (2) Ratings were made in field settings, i.e. laboratory research studies were not included. (3) Appraisals were made for job performance. Data gathered in educational
contexts (e.g. students rating their classroom performance) or in the context of assessment centers for selection purposes were not included.

**Coding of studies.** Effect-sizes and moderator variables were coded in a way that allowed for computer-based data analysis. If a research article did not provide information about the presence of a specific moderator, the value for this variable was coded as "missing". This approach resulted in varying numbers of samples in each of the subsets of studies used to examine moderator-effects. All coding was completed by the author. To obtain an estimate of inter-coder reliability, ten randomly chosen research articles (40 coefficients) were selected for coding by a second person who held a PhD degree in organizational psychology. Based on these 40 coefficients, inter-coder reliabilities were calculated for all variables whose coding involved subjective judgment. These included the distinction of blue-collar and white-collar jobs (agreement percentage: 92%, Kappa: .75), the classification of performance constructs as representing overall, task or contextual performance (84%, .76), judgmental vs. non-judgmental ratings (91%, .75), and the distinction of personality traits from other performance constructs (44% agreement). Only the distinction of traits from other performance constructs was not reliably made. Therefore, the two raters discussed each scale label (for all studies) to select only those for which both raters agreed that they represented qualities of personality. If the two raters did not reach consensus, the respective scale was excluded from the moderator analysis testing for effects of performance constructs. For the coding of factual information, reliability estimates were not calculated (effect size estimates, sample sizes, demographic data, the stated purpose of ratings, behavioral scale definitions\(^2\), the number of scales and items aggregated in each measure, and the use of social comparison terminology to define scale anchors). The coding of this information required extracting explicitly stated information from research texts. Coding criteria had been specified in a coding manual, and both raters coded three articles for training purposes to detect potential

\(^{2}\)As many articles did not provide examples of how items were worded, scale labels were coded to be "behavioral" if researchers explicitly stated that they used "behavioral" scale descriptions.
sources of disagreement before the author started coding the entire set of articles (see appendix A for an overview of the coding criteria).

4.3 Meta-analytical techniques

*Effect size estimates.* Two indexes of effect size were analyzed, correlation coefficients (*r*-index), and mean difference scores (*d*-index). Leniency, i.e. index *d*, was measured by calculating the difference between the mean levels of supervisory and self-ratings. The difference score was weighted by the pooled standard deviations from both rater groups. Difference scores were calculated by subtracting supervisor-ratings from self-ratings. A positive effect size *d* thus indicates that self-ratings yielded higher averages than supervisory evaluations, i.e. leniency in self-ratings. (Note that the *r* and *d* statistics as used in this study are two empirical, non-interchangeable measures. No arithmetical transformations between the two indexes of effect size were made.)

Missing information was substituted by estimates if available. For the *r*-index, such substitutions were only made for a few articles that failed to report the exact number of supervisors who provided ratings. In these cases, the sample sizes of self-raters and supervisors were assumed to be equal. For the *d*-index, missing information was substituted by estimates in six studies which provided means for self- and supervisory ratings but did not report any measure of dispersion. In these cases, standard deviations were estimated by averaging the values reported in all the other studies that had used scales with identical numbers of points.

Most studies yielded multiple effect sizes. To deal with the potential problems of non-independent data, the contribution of any single study was limited to a single effect size for each index (*r* and *d*). To achieve this, coefficients were averaged within samples whenever individual studies reported several coefficients. For moderator analyses, a single study could contribute more than one coefficient - but only one to each effect size estimate. Cooper (1998) described this approach as "shifting unit of analysis", which retains as much information as possible in the dataset without putting a threat to the independence of data points (Viswesvaran, Schmidt, &
Ones, 2002).

*The meta-analytical model.* In choosing the analytical approach two goals were pursued, namely ensuring robustness of results to model assumptions and allowing for comparisons with earlier meta-analyses on rater agreement. As earlier meta-analyses have applied the meta-analytical techniques outlined in Hunter and Schmidt (1990), I report such analyses for index $r$ as an indicator of validity. For index $d$ as a measure of leniency in self-ratings, I referred to the equations described in Hedges and Olkin (1985). To assess overall and moderator effects, I used the SPSS macros described by Lipsey and Wilson (2001), using a random effects model to estimate overall effects and the fixed-effects model for moderator analyses. This approach ensured that the present results can be compared with that of earlier meta-analyses. In addition, I repeated the same analyses with a mixed-effects model (using MlwiN 2.0, see Rashbash, 2003). The advantages of this approach include that estimating the effects of moderator variables while controlling for other variables and estimating the amount of variance explained by the hypothesized moderators is possible (Hox & Leeuw, 2003). By combining these two analytical approaches, it is possible to test the robustness of moderator effects to model assumptions. To reveal the extent to which the results of moderator-analyses depend upon the model chosen for the analysis, the results of the fixed-effects analysis could be compared to the results obtained from a mixed effects model used to analyze the same set of moderator variables. As it is generally difficult to

---

3For the mixed-effects analysis, a two-level model was set up in MlwiN (Rashbash, 2003). Effect size estimates were declared as the response variable; and two predictors, the constant and the estimated standard error of the effect sizes, were included. The constant had a fixed coefficient on the first level and a coefficient that varied randomly on the second level. The standard error had a random coefficient that varied at the first level only, with a coefficient fixed at one (Lambert & Abrams, 2002). Explanatory variables were entered as fixed effects. To deal with the problem of missing values, these were recoded for each covariate, and dummy variables that indicated missing values were entered along with the covariates. Thus it was also possible to test whether missing cases carried information. Detailed results of these analyses are displayed in appendix D.
decide which models’ statistical and sampling assumptions apply to a given dataset, a pragmatic approach is to compare the results of both fixed and mixed effects models. On the one hand, a significant moderator effect is more likely to indicate a true moderator variable if the finding results from a mixed model. On the other hand, a non-significant effect is more likely to indicate a truly non-significant effect if it stems from a fixed-effects model (Overton, 1998). Besides, whereas the fixed-effects model allows for a generalization of results across the sampled studies, the mixed model provides information about the extent to which results generalize beyond the given set of studies.

**Correction for study artifacts.** Hunter and Schmidt (1990) list potentially correctable artifacts that alter the value of outcome measures. The general goal of artifact correction is twofold: reaching an improved estimate of the true value of the outcome measure and testing whether there is variation in study results that can be attributed to moderator variables. I made corrections for two of these artifacts, sampling error and error of measurement. Accordingly, the sample size weighted means of effect size indexes were calculated and then corrected for unreliability.

There has been debate over which measure of reliability should be used to correct for measurement error in the context of performance ratings (Murphy & De Shon, 2000; Schmidt, Viswesvaran, & Ones, 2000). In agreement with authors such as Viswesvaran et al. (2002) and Schmidt et al. (2000), interrater reliabilities were used to correct supervisory ratings. Interrater reliability indicates the extent to which different raters agree on the performance of a group of individuals. It comprises the effects of all four sources of measurement error that are relevant in supervisor ratings: differences in leniency, rater-by-ratee interactions, random response error, and transient error (Schmidt et al., 2000). As information on interrater-reliability was only sporadically available in the present sample of studies, I employed a meta-analytical reliability estimate provided by earlier research. Viswesvaran (1996) reported inter-rater reliabilities as well as intra-rater reliabilities (rate-rerate, coefficient alpha) for performance ratings. These authors
reported a mean interrater reliability for supervisory ratings of .52. Conway and Huffcutt (1997) obtained a value of .50 for interrater reliability in supervisory ratings. For the current study, the mean of these two estimates (.51) was employed to correct for unreliability in supervisory ratings.

The case of self-evaluations has not been explicitly discussed by the authors named above, nor in the related literature (Moser, 1999). In fact, the meta-analysis of Viswesvaran et al. (1996) did not include self-ratings and I am not aware that any meta-analytically obtained measure of reliability has been reported for self-ratings. I suggest that the appropriate measure to correct for unreliability in self-ratings is rate-rerate reliability. Rate-rerate reliability assigns transient error as a source of variance to measurement error, assuming that the level of performance remains constant throughout the period between test and retest. Transient error includes random variance from differences in mood, mental state, and other factors in the raters which vary over the relatively short periods of time used in retest-reliability studies. An estimate of self-rating reliability was derived by calculating the sample-size-weighted average of all the rate-rerate reliabilities that were reported in the current set of primary research articles for self-ratings. The resulting estimate was .74.

As mentioned before, self-ratings have often been found to be lenient. This may also imply that self-ratings show restriction in range. Giving lenient evaluations, self-raters tend not to use the full range of the scales presented to them. Although this restriction of range does mathematically reduce the validities calculated for self-reports, I did not consider this effect to represent an artifact that effect size estimates should be corrected for. After all, leniency in self-ratings does not represent an artifact of study design, but reflects behavior that is likely to truly characterize self-raters. Thus, measurement error can be considered the only relevant artifact that corrections should be made for, beyond sampling error. The present analysis followed this rationale, and no corrections were made for range restriction.
Checking for publication bias. Studies that find statistically significant results may have a higher probability of being published - and thus of being retrieved by meta-analysts. This problem is referred to as publication bias. Publication bias can be assessed by examining the influence of sample size on study outcomes. The underlying rationale is as follows: Given that all studies estimate the same population parameter, the findings of smaller studies should be more variable. They yield both the largest under- and the largest over-estimations of the underlying population parameter. If studies then get published depending on the statistical significance of results, this should lead to the smaller studies reporting larger effect-sizes. If it can be assumed that all effect-sizes estimate the same population parameter, the "funnel plot" is a prominent means of assessing publication bias. A funnel plot is a simple plot of each study’s effect-size against the study’s precision (i.e. sample size). If no publication bias is present, the plot should look like an inverted funnel: studies with small sample sizes show greater variation in effect-sizes without over-representing large outcome measures.

For the present analysis, effect-sizes are expected to depend on a number of covariates, though. Thus, a different approach than interpreting a single funnel plot is warranted. Covariate effects need to be controlled before a potential effect of sample size can be examined. This can be done by using a two-level model, as described in Hox & De Leeuw (2003). Publication bias is assessed by including sample size along with relevant covariates as explanatory variable. This way, a formal statistical test can be conducted to decide whether publication bias is likely to be an issue.

Examining multiple moderators. To deal with the fact that moderator variables are likely to be correlated with each other, I used an approach described by Cooper (1998, p.152): Homogeneity statistics were generated for each variable separately to assess the main effects of each moderator. Prior to an interpretation of the results, an inter-correlation matrix was examined to identify confounded variables and to assess which conclusions can be drawn from the data. Whenever
possible, attempts were made to control for confounds.

5 Results

A total of 82 research articles was located that reported information on 104 independent samples. For a total of 96 independent samples, self-supervisory correlations (417 coefficients), and for 70 samples mean difference scores (324 coefficients) between self and supervisory ratings were obtained. Correlational studies included a total of 22,287 respondents, and 29,386 respondents had provided ratings that mean difference scores were calculated for. A majority (70, 67%) of the studies retrieved were conducted in the United States. 34 articles reported information on samples from other countries than the US, including Great Britain (13), Republic of China (9), Germany (3), Israel (2), Australia (2), China (1), Nigeria (1), the Netherlands (1), Finland (1), and a mixed sample with respondents from two nations, the UK and Hong Kong. All these studies were published between 1955 and 2003. For 64 samples, information on the gender composition of the sample was reported. The percentage of women in the samples averaged 40%, with a minimum of 0% and a maximum of 100%.

The mean age of respondents in the samples was 36 years, based on 54 samples for which this information was provided. The majority of studies used convenience samples, and collected data from an average of 221 respondents. The most frequent setting of research studies was the private sector industry (68, 65%), followed by public service and government agencies (20, 21%), and the military (5, 5%). Additional publication statistics appear in table 2.
Table 2 Studies in the review: publication statistics

<table>
<thead>
<tr>
<th>Publication source</th>
<th>Number of samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personnel Psychology</td>
<td>32</td>
</tr>
<tr>
<td>Journal of Occupational and Organizational Psychology</td>
<td>16</td>
</tr>
<tr>
<td>Journal of Applied Psychology</td>
<td>14</td>
</tr>
<tr>
<td>Academy of Management Journal</td>
<td>5</td>
</tr>
<tr>
<td>International Journal of Selection and Assessment</td>
<td>3</td>
</tr>
<tr>
<td>Journal of Educational Psychology</td>
<td>3</td>
</tr>
<tr>
<td>Environment and behavior</td>
<td>3</td>
</tr>
<tr>
<td>Human Relations</td>
<td>2</td>
</tr>
<tr>
<td>Human Performance</td>
<td>2</td>
</tr>
<tr>
<td>Journal of Applied Social Psychology</td>
<td>2</td>
</tr>
<tr>
<td>IEEE Transactions on EM</td>
<td>2</td>
</tr>
<tr>
<td>Unpublished Document</td>
<td>2</td>
</tr>
<tr>
<td>Public Personnel Management</td>
<td>2</td>
</tr>
<tr>
<td>Zeitschrift fuer Personalspsychologie</td>
<td>2</td>
</tr>
<tr>
<td>Organizational Behavior and Human Decision Processes</td>
<td>1</td>
</tr>
<tr>
<td>Zeitschrift fuer experimentelle und angewandte Psychologie</td>
<td>1</td>
</tr>
<tr>
<td>Sociology of Education</td>
<td>1</td>
</tr>
<tr>
<td>Dataset published in a book</td>
<td>1</td>
</tr>
<tr>
<td>Evaluation and Program Planning</td>
<td>1</td>
</tr>
<tr>
<td>Journal of Management</td>
<td>1</td>
</tr>
<tr>
<td>Journal of Organizational Behavior</td>
<td>1</td>
</tr>
<tr>
<td>Journal of Management Development</td>
<td>1</td>
</tr>
<tr>
<td>ERIC Document Reproduction Service</td>
<td>1</td>
</tr>
<tr>
<td>Dissertation</td>
<td>1</td>
</tr>
<tr>
<td>Applied Psychology: An International Review</td>
<td>1</td>
</tr>
<tr>
<td>Journal of Industrial Psychology</td>
<td>1</td>
</tr>
<tr>
<td>European Journal of Work and Organizational Psychology</td>
<td>1</td>
</tr>
<tr>
<td>Journal of Social Behavior and Personality</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>104</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year of publication</th>
<th>Number of samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1955 - 1965</td>
<td>7</td>
</tr>
<tr>
<td>1966 - 1975</td>
<td>11</td>
</tr>
<tr>
<td>1976 - 1985</td>
<td>18</td>
</tr>
<tr>
<td>1986 - 1995</td>
<td>39</td>
</tr>
<tr>
<td>1996 - 2003</td>
<td>29</td>
</tr>
</tbody>
</table>

As described in the method part, the sample of research articles that could be retrieved was tested for publication bias. Sample-size did not show any effect on study outcomes, for neither indicator of rater agreement. Thus, it could be concluded that publication bias was not an issue in the given database of samples.
5.1 Meta-analytical results for overall $r$ and $d$

To examine the validity of self-ratings the overall weighted correlation of self- and supervisor-ratings was calculated for the entire dataset. On the basis of 96 independent samples (417 coefficients) an average correlation of $r = .22$ was obtained when $r$ was corrected for sampling error only. After correction for unreliability, an average correlation of $.36$ resulted. A total of 70 samples (324 coefficients) included mean difference scores. A meta-analytical estimate of $d = .33$ resulted. As the confidence interval for the sample size weighted mean did not include zero, it can be assumed that self-ratings were significantly higher than supervisory ratings. Correcting for unreliability yielded a mean effect size estimate of $d = .41$. Table 3 presents an overview of the overall results for both effect size estimates, $r$ and $d$.

Table 3 Overall results for both effect size estimates $r$ and $d$

<table>
<thead>
<tr>
<th>Effect size index</th>
<th>k</th>
<th>n</th>
<th>$ES_{wt}$</th>
<th>$SE_{wt}$</th>
<th>95% CI</th>
<th>$ES_{pop}$</th>
<th>% Varart</th>
<th>QW</th>
<th>95% Cred</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect size $r$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all samples</td>
<td>96</td>
<td>22287</td>
<td>.22**</td>
<td>.136</td>
<td>.19 - .26</td>
<td>.36</td>
<td>20%</td>
<td>409.37**</td>
<td>-.04 - .74</td>
</tr>
<tr>
<td>Effect size $d$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all samples</td>
<td>70</td>
<td>29386</td>
<td>.33**</td>
<td>.046</td>
<td>.24 - .42</td>
<td>.41</td>
<td>21%</td>
<td>1449.56**</td>
<td>.30 - .53</td>
</tr>
</tbody>
</table>

Note. $k = \text{number of coefficients}; n = \text{total number of respondents}; E_{\text{size}_{\text{wt}}} = \text{sample size weighted mean effect size (significance of effect sizes was assessed by the normal z distribution)}; SE_{\text{size}_{\text{wt}}} = \text{standard error of the sample size weighted mean effect size}; CI = \text{confidence interval (random effects model)}; ES_{\text{pop}} = \text{estimate of population effect size}; \% \text{Varart} = \text{percent of variance accounted for by artifacts}; Q_{W} = \text{within class test of homogeneity – a non-significant Q reflects homogeneity within effect sizes of all the samples included (df = k-1, k = number of samples)}; C_{\text{red}} = \text{credibility interval (random effects model)}. Effect size $r$ was estimated following the procedures outlined in Schmidt & Hunter, 1990; index $d$ was estimated following the procedures suggested by Hedges & Olkin, 1985. Analysis with z-transformation; random effects model. $** = p \leq .01, * = p \leq .05$.

Omnibus tests of homogeneity revealed significant heterogeneity for both $r$ ($Q = 409$, df = 95, $p = .000$) and $d$ ($Q = 1449$, df = 69, $p = .000$). As the variation among the effect sizes obtained from independent samples can not be attributed solely to artifacts, the next step was to test for effects of the hypothesized moderator variables.
5.2 Moderator analyses

Moderator effects were tested by conducting a meta-analysis on each of the subsets of samples that resulted from breaking samples up according to the categories of each moderator variable. To determine whether correlations or mean differences varied between the subsets, between-class homogeneity tests were calculated as described by Hedges and Olkin (1985). In addition, a random-effects model variance component was calculated to quantify the extent of between study variability.

Table 4 Moderator analyses for effect size index r

<table>
<thead>
<tr>
<th>Set</th>
<th>k</th>
<th>( r_{wt} )</th>
<th>SE</th>
<th>95%CI</th>
<th>( r_p )</th>
<th>( Q_W )</th>
<th>( Q_B )</th>
<th>( p_{Q_B} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale format</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>single-item</td>
<td>38</td>
<td>.23</td>
<td>.014</td>
<td>.201-.255</td>
<td>.45</td>
<td>76.90**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>aggregated measure</td>
<td>67</td>
<td>.21</td>
<td>.007</td>
<td>.197-.226</td>
<td>.33</td>
<td>349.52**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>105</td>
<td>.21</td>
<td>.006</td>
<td>.202-.228</td>
<td>.38</td>
<td>427.56**</td>
<td>1.14</td>
<td>.2851</td>
</tr>
<tr>
<td>Aggregation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>homogeneous</td>
<td>34</td>
<td>.18</td>
<td>.011</td>
<td>.155-.199</td>
<td>.34</td>
<td>59.94**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>heterogeneous</td>
<td>35</td>
<td>.24</td>
<td>.010</td>
<td>.221-.259</td>
<td>.37</td>
<td>275.21**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>69</td>
<td>.21</td>
<td>.007</td>
<td>.199-.227</td>
<td>.35</td>
<td>352.62**</td>
<td>17.47</td>
<td>.0000**</td>
</tr>
<tr>
<td>Scale definition</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>broad</td>
<td>21</td>
<td>.24</td>
<td>.016</td>
<td>.206-.270</td>
<td>.29</td>
<td>40.53**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>behavioral</td>
<td>9</td>
<td>.25</td>
<td>.036</td>
<td>.180-.324</td>
<td>.39</td>
<td>5.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>30</td>
<td>.24</td>
<td>.015</td>
<td>.211-.270</td>
<td>.35</td>
<td>46.50**</td>
<td>0.12</td>
<td>.7268</td>
</tr>
<tr>
<td>Social comparison</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>not given</td>
<td>72</td>
<td>.22</td>
<td>.008</td>
<td>.205-.236</td>
<td>.39</td>
<td>348.51**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>given</td>
<td>15</td>
<td>.23</td>
<td>.017</td>
<td>.201-.269</td>
<td>.41</td>
<td>48.49**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>87</td>
<td>.22</td>
<td>.007</td>
<td>.209-.237</td>
<td>.39</td>
<td>397.56**</td>
<td>0.56</td>
<td>.4527</td>
</tr>
<tr>
<td>Performance construct</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>global</td>
<td>23</td>
<td>.22</td>
<td>.018</td>
<td>.185-.255</td>
<td>.36</td>
<td>73.87**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>specific</td>
<td>27</td>
<td>.23</td>
<td>.015</td>
<td>.201-.261</td>
<td>.38</td>
<td>39.94**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>50</td>
<td>.23</td>
<td>.012</td>
<td>.204-.249</td>
<td>.36</td>
<td>114.05**</td>
<td>0.24</td>
<td>.6257</td>
</tr>
<tr>
<td>Performance construct</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>task-performance</td>
<td>58</td>
<td>.21</td>
<td>.009</td>
<td>.190-.224</td>
<td>.34</td>
<td>135.16**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>contextual perf.</td>
<td>46</td>
<td>.19</td>
<td>.010</td>
<td>.166-.206</td>
<td>.30</td>
<td>114.54**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>trait</td>
<td>13</td>
<td>.22</td>
<td>.020</td>
<td>.184-.260</td>
<td>.36</td>
<td>8.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>117</td>
<td>.20</td>
<td>.007</td>
<td>.185-.211</td>
<td>.37</td>
<td>261.50**</td>
<td>3.60</td>
<td>.1652</td>
</tr>
<tr>
<td>Performance indicators</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>judgmental</td>
<td>94</td>
<td>.21</td>
<td>.007</td>
<td>.198-.225</td>
<td>.34</td>
<td>388.06**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>non-judgmental</td>
<td>5</td>
<td>.46</td>
<td>.041</td>
<td>.383-.543</td>
<td>.75</td>
<td>8.45**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>99</td>
<td>.22</td>
<td>.007</td>
<td>.205-.231</td>
<td>.35</td>
<td>433.31**</td>
<td>36.81</td>
<td>.0000**</td>
</tr>
<tr>
<td>Confidentiality</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>non-confidential</td>
<td>17</td>
<td>.26</td>
<td>.022</td>
<td>.221-.306</td>
<td>.43</td>
<td>13.56**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>confidential</td>
<td>51</td>
<td>.25</td>
<td>.010</td>
<td>.233-.274</td>
<td>.41</td>
<td>283.09**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>68</td>
<td>.26</td>
<td>.009</td>
<td>.237-.274</td>
<td>.42</td>
<td>296.83**</td>
<td>0.17</td>
<td>.6806</td>
</tr>
<tr>
<td>Validation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>not expected</td>
<td>78</td>
<td>.22</td>
<td>.007</td>
<td>.209-.239</td>
<td>.36</td>
<td>342.92**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>expected</td>
<td>13</td>
<td>.27</td>
<td>.023</td>
<td>.225-.314</td>
<td>.44</td>
<td>22.05**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>91</td>
<td>.23</td>
<td>.007</td>
<td>.215-.243</td>
<td>.37</td>
<td>368.58**</td>
<td>3.60</td>
<td>.0577</td>
</tr>
</tbody>
</table>
Table 4, continued

<table>
<thead>
<tr>
<th>Set</th>
<th>k</th>
<th>r_{wt}</th>
<th>SE</th>
<th>95% CI</th>
<th>r_ρ</th>
<th>Q_W</th>
<th>Q_B</th>
<th>p_{Q_B}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purpose</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>administration</td>
<td>5</td>
<td>.29</td>
<td>.042</td>
<td>.212-.376</td>
<td>.48</td>
<td>20.45**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>development</td>
<td>18</td>
<td>.16</td>
<td>.010</td>
<td>.141-.181</td>
<td>.26</td>
<td>41.02**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>research</td>
<td>70</td>
<td>.25</td>
<td>.009</td>
<td>.236-.272</td>
<td>.41</td>
<td>298.22**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>88</td>
<td>.21</td>
<td>.007</td>
<td>.201-.228</td>
<td>.35</td>
<td>408.27**</td>
<td>48.58</td>
<td>.0000**</td>
</tr>
<tr>
<td>Managerial samples</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>development</td>
<td>15</td>
<td>.18</td>
<td>.012</td>
<td>.156-.204</td>
<td>.29</td>
<td>24.61**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>research</td>
<td>12</td>
<td>.21</td>
<td>.019</td>
<td>.173-.248</td>
<td>.34</td>
<td>18.79**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>27</td>
<td>.19</td>
<td>.010</td>
<td>.169-.209</td>
<td>.31</td>
<td>45.25**</td>
<td>1.84</td>
<td>.1744</td>
</tr>
<tr>
<td>360-degree fewer sources</td>
<td>18</td>
<td>.19</td>
<td>.016</td>
<td>.162-.226</td>
<td>.32</td>
<td>32.66**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>360-degree combined</td>
<td>12</td>
<td>.18</td>
<td>.013</td>
<td>.157-.208</td>
<td>.30</td>
<td>14.58**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>30</td>
<td>.19</td>
<td>.010</td>
<td>.167-.207</td>
<td>.30</td>
<td>47.57**</td>
<td>0.33</td>
<td>.5665</td>
</tr>
<tr>
<td>Jobtype</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>blue-collar</td>
<td>7</td>
<td>.33</td>
<td>.028</td>
<td>.278-.387</td>
<td>.54</td>
<td>46.62**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>white-collar</td>
<td>85</td>
<td>.22</td>
<td>.007</td>
<td>.203-.232</td>
<td>.35</td>
<td>305.16**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>92</td>
<td>.22</td>
<td>.007</td>
<td>.211-.239</td>
<td>.37</td>
<td>367.71**</td>
<td>15.92</td>
<td>.0001**</td>
</tr>
<tr>
<td>White-collar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>high education &lt;80%</td>
<td>19</td>
<td>.33</td>
<td>.019</td>
<td>.290-.363</td>
<td>.53</td>
<td>137.90**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>high education &gt;80%</td>
<td>13</td>
<td>.19</td>
<td>.015</td>
<td>.157-.217</td>
<td>.30</td>
<td>39.40**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>32</td>
<td>.24</td>
<td>.012</td>
<td>.220-.266</td>
<td>.40</td>
<td>211.22**</td>
<td>33.91</td>
<td>.0000**</td>
</tr>
<tr>
<td>White-collar non-managerial</td>
<td>30</td>
<td>.21</td>
<td>.017</td>
<td>.175-.243</td>
<td>.34</td>
<td>81.57**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>managerial</td>
<td>30</td>
<td>.19</td>
<td>.010</td>
<td>.167-.207</td>
<td>.30</td>
<td>47.57**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>60</td>
<td>.19</td>
<td>.009</td>
<td>.175-.210</td>
<td>.31</td>
<td>130.38**</td>
<td>1.24</td>
<td>.2653</td>
</tr>
<tr>
<td>High education</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>non-managerial</td>
<td>6</td>
<td>.22</td>
<td>.037</td>
<td>.148-.295</td>
<td>.30</td>
<td>37.31**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>managerial</td>
<td>6</td>
<td>.17</td>
<td>.019</td>
<td>.138-.212</td>
<td>.36</td>
<td>.77</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>12</td>
<td>.19</td>
<td>.017</td>
<td>.152-.218</td>
<td>.29</td>
<td>39.31**</td>
<td>1.23</td>
<td>.2681</td>
</tr>
<tr>
<td>% females</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; median</td>
<td>8</td>
<td>.24</td>
<td>.031</td>
<td>.202-.286</td>
<td>.37</td>
<td>14.96**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt; median</td>
<td>14</td>
<td>.23</td>
<td>.022</td>
<td>.184-.281</td>
<td>.40</td>
<td>79.71**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>22</td>
<td>.24</td>
<td>.018</td>
<td>.203-.272</td>
<td>.37</td>
<td>94.93**</td>
<td>0.25</td>
<td>.6147</td>
</tr>
<tr>
<td>Culture</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Western</td>
<td>85</td>
<td>.21</td>
<td>.007</td>
<td>.199-.226</td>
<td>.35</td>
<td>334.09**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>11</td>
<td>.23</td>
<td>.025</td>
<td>.184-.281</td>
<td>.38</td>
<td>74.66**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>96</td>
<td>.21</td>
<td>.007</td>
<td>.201-.227</td>
<td>.35</td>
<td>409.37**</td>
<td>0.63</td>
<td>.4278</td>
</tr>
</tbody>
</table>

Note. k = number of coefficients included in the meta-analysis; r_{wt} = sample size weighted mean correlation; SE = standard error of the mean effect size; CI = confidence interval; r_ρ = population correlation, corrected for artifacts; Q_W = within class test of homogeneity – a non-significant Q reflects homogeneity within category; Q_B = between class test of homogeneity – a significant Q_B indicates that classes differ significantly (df = m-1, m = number of categories); p_{Q_B} = probability of Q_B; ** = p≤.01, * = p≤.05. Fixed-effects model; analysis with z-transformation.
<table>
<thead>
<tr>
<th>Set</th>
<th>k</th>
<th>$d_{w}$</th>
<th>SE</th>
<th>95% CI</th>
<th>$d_{5}$</th>
<th>$Q_W$</th>
<th>$Q_B$</th>
<th>$p_{Q_B}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale format</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>single-item</td>
<td>28</td>
<td>.38</td>
<td>.024</td>
<td>.330-.425</td>
<td>.48</td>
<td>328.60**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>aggregated</td>
<td>49</td>
<td>.37</td>
<td>.013</td>
<td>.346-.398</td>
<td>.47</td>
<td>761.58**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>77</td>
<td>.37</td>
<td>.012</td>
<td>.350-.396</td>
<td>.47</td>
<td>109.22**</td>
<td>0.04</td>
<td>.8451</td>
</tr>
<tr>
<td>Scale definition</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>broad</td>
<td>14</td>
<td>.34</td>
<td>.032</td>
<td>.280-.407</td>
<td>.44</td>
<td>218.29**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>behavioral</td>
<td>8</td>
<td>.20</td>
<td>.052</td>
<td>.098-.303</td>
<td>.25</td>
<td>34.09**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>22</td>
<td>.30</td>
<td>.028</td>
<td>.250-.358</td>
<td>.38</td>
<td>257.75**</td>
<td>5.37</td>
<td>.0204*</td>
</tr>
<tr>
<td>Social comparison</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>not given</td>
<td>61</td>
<td>.40</td>
<td>.013</td>
<td>.370-.421</td>
<td>.50</td>
<td>871.55**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>given</td>
<td>6</td>
<td>.35</td>
<td>.035</td>
<td>.287-.422</td>
<td>.45</td>
<td>40.57**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>67</td>
<td>.39</td>
<td>.012</td>
<td>.366-.414</td>
<td>.49</td>
<td>913.32**</td>
<td>1.20</td>
<td>.2738</td>
</tr>
<tr>
<td>Performance construct</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>global</td>
<td>17</td>
<td>.37</td>
<td>.033</td>
<td>.304-.434</td>
<td>.47</td>
<td>200.52**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>specific</td>
<td>19</td>
<td>.30</td>
<td>.029</td>
<td>.241-.353</td>
<td>.37</td>
<td>222.45**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>36</td>
<td>.33</td>
<td>.022</td>
<td>.285-.370</td>
<td>.41</td>
<td>425.71**</td>
<td>2.75</td>
<td>.0974</td>
</tr>
<tr>
<td>Performance construct</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>contextual perf.</td>
<td>41</td>
<td>.33</td>
<td>.016</td>
<td>.306-.370</td>
<td>.43</td>
<td>540.70**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>trait</td>
<td>49</td>
<td>.37</td>
<td>.015</td>
<td>.342-.400</td>
<td>.47</td>
<td>715.78**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>11</td>
<td>.72</td>
<td>.035</td>
<td>.647-.784</td>
<td>.90</td>
<td>263.30**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Validation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>not expected</td>
<td>68</td>
<td>.40</td>
<td>.012</td>
<td>.372-.420</td>
<td>.50</td>
<td>912.86**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>expected</td>
<td>4</td>
<td>.32</td>
<td>.064</td>
<td>.199-.450</td>
<td>.41</td>
<td>20.20**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>72</td>
<td>.39</td>
<td>.012</td>
<td>.370-.417</td>
<td>.50</td>
<td>943.26**</td>
<td>1.21</td>
<td>.2718</td>
</tr>
<tr>
<td>Confidentiality</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>non-confid.</td>
<td>12</td>
<td>.42</td>
<td>.038</td>
<td>.345-.495</td>
<td>.53</td>
<td>75.84**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>confidential</td>
<td>35</td>
<td>.35</td>
<td>.019</td>
<td>.313-.388</td>
<td>.44</td>
<td>569.30**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>47</td>
<td>.36</td>
<td>.017</td>
<td>.331-.397</td>
<td>.46</td>
<td>647.78**</td>
<td>2.64</td>
<td>.1045</td>
</tr>
<tr>
<td>Purpose</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>administration</td>
<td>5</td>
<td>.48</td>
<td>.061</td>
<td>.366-.607</td>
<td>.61</td>
<td>76.28**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>development</td>
<td>15</td>
<td>.39</td>
<td>.019</td>
<td>.353-.427</td>
<td>.49</td>
<td>164.20**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>research</td>
<td>48</td>
<td>.37</td>
<td>.017</td>
<td>.334-.400</td>
<td>.46</td>
<td>648.75**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>68</td>
<td>.38</td>
<td>.012</td>
<td>.358-.405</td>
<td>.38</td>
<td>803.13**</td>
<td>3.91</td>
<td>.1417</td>
</tr>
<tr>
<td>Managerial samples</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>development</td>
<td>13</td>
<td>.33</td>
<td>.022</td>
<td>.289-.378</td>
<td>.42</td>
<td>141.21**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>research</td>
<td>10</td>
<td>.45</td>
<td>.036</td>
<td>.381-.522</td>
<td>.57</td>
<td>169.21**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>23</td>
<td>.37</td>
<td>.019</td>
<td>.329-.404</td>
<td>.46</td>
<td>318.15**</td>
<td>7.73</td>
<td>.0054**</td>
</tr>
<tr>
<td>360-degree</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fewer sources</td>
<td>15</td>
<td>.45</td>
<td>.029</td>
<td>.394-.507</td>
<td>.57</td>
<td>170.39**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>11</td>
<td>.33</td>
<td>.024</td>
<td>.284-.377</td>
<td>.42</td>
<td>150.97**</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>.38</td>
<td>.018</td>
<td>.343-.415</td>
<td>.48</td>
<td>340.61**</td>
<td>10.25</td>
<td>.0014**</td>
</tr>
</tbody>
</table>
Table 5, continued

<table>
<thead>
<tr>
<th>Jobtype</th>
<th>k</th>
<th>(d_{wt})</th>
<th>SE</th>
<th>95%CI</th>
<th>(d_{δ})</th>
<th>(Q_W)</th>
<th>(Q_B)</th>
<th>(p_{Q_B})</th>
</tr>
</thead>
<tbody>
<tr>
<td>blue-collar</td>
<td>5</td>
<td>.59</td>
<td>.052</td>
<td>.489-.693</td>
<td>.74</td>
<td>45.35**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>white-collar</td>
<td>61</td>
<td>.36</td>
<td>.014</td>
<td>.333-.387</td>
<td>.45</td>
<td>753.48**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>66</td>
<td>.37</td>
<td>.013</td>
<td>.349-.401</td>
<td>.47</td>
<td>817.24**</td>
<td>18.41</td>
<td>.0000**</td>
</tr>
<tr>
<td>White-collar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>high education &lt;80%</td>
<td>11</td>
<td>.54</td>
<td>.045</td>
<td>.447-.624</td>
<td>.67</td>
<td>101.38**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>high education &gt;80%</td>
<td>7</td>
<td>.29</td>
<td>.039</td>
<td>.212-.364</td>
<td>.36</td>
<td>53.29**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>18</td>
<td>.39</td>
<td>.029</td>
<td>.335-.451</td>
<td>.49</td>
<td>172.03**</td>
<td>17.36</td>
<td>.0000**</td>
</tr>
<tr>
<td>White-collar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>non-managerial</td>
<td>16</td>
<td>.45</td>
<td>.034</td>
<td>.381-.516</td>
<td>.56</td>
<td>170.58**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>managerial</td>
<td>26</td>
<td>.38</td>
<td>.018</td>
<td>.343-.415</td>
<td>.48</td>
<td>340.61**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>42</td>
<td>.39</td>
<td>.016</td>
<td>.363-.426</td>
<td>.50</td>
<td>514.33**</td>
<td>3.14</td>
<td>.0766</td>
</tr>
<tr>
<td>% females</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; median</td>
<td>3</td>
<td>.99</td>
<td>.077</td>
<td>.839-1.142</td>
<td>1.25</td>
<td>20.10**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt; median</td>
<td>12</td>
<td>.37</td>
<td>.036</td>
<td>.301-.443</td>
<td>.47</td>
<td>94.89**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>15</td>
<td>.48</td>
<td>.033</td>
<td>.419-.547</td>
<td>.61</td>
<td>167.34**</td>
<td>52.35</td>
<td>.0000**</td>
</tr>
<tr>
<td>Culture</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Western</td>
<td>60</td>
<td>.44</td>
<td>.013</td>
<td>.412-.462</td>
<td>.55</td>
<td>709.46**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>10</td>
<td>-.07</td>
<td>.039</td>
<td>-.147-.006</td>
<td>-.09</td>
<td>50.15**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combined</td>
<td>70</td>
<td>.39</td>
<td>.012</td>
<td>.364-.412</td>
<td>.49</td>
<td>913.23**</td>
<td>153.61</td>
<td>.0000**</td>
</tr>
</tbody>
</table>

Note. \(k\) = number of coefficients included in the meta-analysis; \(d_{wt}\) = sample size weighted mean difference; \(SE\) = standard error of the mean effect size; \(CI\) = confidence interval; \(d_{δ}\) = estimate of population effect size corrected for artifacts; \(Q_W\) = within class test of homogeneity – a non-significant \(Q\) reflects homogeneity within category; \(Q_B\) = between class test of homogeneity – a significant \(Q_B\) indicates that classes differ significantly (df = m-1, m = number of categories); \(p_{Q_B}\) = probability of \(Q_B\); ** = \(p\leq.01\), * = \(p\leq.05\). Fixed-effects model; weighted integration method (Hedges & Olkin, 1985); analysis with z-transformation.

5.2.1 Report format

Scale format. Contrary to expectations, self-supervisor agreement as measured by correlation coefficients did not significantly differ between single-item and aggregated measures, i.e. measures that are the arithmetic mean of several items. For correlation coefficients, the mean effect size estimate for single-item measures (.23) was even slightly higher than that obtained for aggregated measures (.21), though this difference was not significant (\(Q=1.14, df=1, p=.2851\)). Within aggregated measures, composite scores which integrated a number of heterogeneous items showed significantly higher self-supervisor correlations (\(r=.24\)) than aggregated scores that were based on a set of homogeneous items (\(r=.18; p=.0000\)).

\(^4\)The aggregate ratings that Viswesvaran et al. (1996) reported the highest alpha estimates for were mostly sums of ratings across several performance dimensions. To further explore possible causes of low validities for aggregate measures, in a post-hoc analysis the samples that reported validities for aggregate ratings were broken up into two sets. Composite scores that integrated several scales (which are then typically interpreted as measures of overall performance) were compared to aggregate measures of performance in specific dimensions, i.e. to measures that aggregated items composing a single scale. Moderator analysis showed that composite scores which aggregated
The use of social comparison terminology moderated neither the size of self-supervisory correlations nor the size of mean difference scores. The distinction of broad as opposed to behavioral scale definitions did not influence correlational agreement but mean difference scores: broad scale labels (d=.34) yielded higher discrepancies between self- and supervisor-ratings than behavioral item descriptions (d=.20).

*Performance dimensions.* Recall the expectation that global ratings of performance should lead to higher interrater correlations than ratings for specific performance constructs. To ensure that the effects of this variable were not confounded with instrument length, only single-item measures were included in the analysis. Contrary to the expectations, correlations reported for single-item measures of global performance (r=.22) were not different from single-item ratings for specific performance dimensions (r=.23; Q=.24, df=1, p=.6257). Also contrary to predictions, global ratings (d=.37) were not associated with significantly higher levels of leniency than ratings for specific performance dimensions (d=.30), although there was a trend in the predicted direction (Q=2.75, df=1, p=.097).

In accordance with the hypotheses, sorting scale labels into three categories of performance constructs yielded significantly different estimates of leniency for contextual performance (d=.33), task performance (d=.37), and trait labels (d=.72) as scale anchors. However this same distinction did not influence the magnitude of correlation coefficients (see table 4 for details).

The extent to which performance ratings are based on non-judgmental criteria was related to correlational self-supervisory-agreement. The correlation increased from r=.21 to r=.46 for non-judgmental indicators of performance. Note that the effect size estimates for non-judgmental performance measures is based on a small number of samples (n=5), so that these results should be interpreted cautiously. Finally, non-judgmental performance ratings were not associated with ratings for various performance areas into a final score, tended to reach higher validities (r=.24) than scores that aggregated items belonging to a single scale (r=.18) (Q=3.75, df=1, p=.053). Composite scores based on several scales had integrated an average of 18.6 items, whereas dimensional aggregates were on average based on 5.6 items. Thus, composite scores may have a reliability advantage. Besides, combining several scales may lead to increased variability in composite scores.
higher leniency.

5.2.2 Conditions of report

Confidentiality of ratings. The majority of primary studies guaranteed confidentiality to respondents. (No study reported that ratings were strictly anonymous.) Samples that were studied in a developmental or administrative context are excluded from the subsequent analysis to avoid a confounding of confidentiality instructions with rating purpose. From the remaining 68 research-based studies 17 studies did not mention confidentiality to respondents. Assuring confidentiality to respondents did not prove to be associated with higher agreement in self and supervisor ratings as measured by both correlation coefficients and leniency estimates ($r=.25 / .26, p=.1694; d=.35 / .42, p=.1045$).

Expecting a validation of ratings tended to correspond with higher correlations of self and supervisory ratings ($r=.27$ vs. $r=.22$). But the between-categories homogeneity test showed this association to be only marginally significant ($Q=3.60, df=1, p=.058$). However, respondents’ expectation of a validation proved to be a strong correlate of reduced leniency in self-ratings. The mean difference between self and supervisor ratings decreased from $d=.40$ in the case that a validation was not expected to $d=.23$ when it was expected ($Q=22.01, df=1, p=.0000$).

Homogeneity testing revealed that the purpose of ratings moderated self-supervisory correlations ($Q=48.58, df=2, p=0.0000$), but did not influence leniency in self-ratings ($Q=3.91, df=2, p=0.1417$), if the analysis was based on the entire database of samples. Since "developmental samples" consisted mostly of managers as respondents, the effect of sample composition was controlled for by including only managerial samples in the moderator analyses comparing developmental and research-based ratings. No differential effect size estimates resulted for correlation coefficients, but for leniency as measure of agreement, a significant moderator effect showed. An average mean difference of $d=.45$ was obtained for ratings in research settings, whereas developmental ratings obtained a mean weighted $d$ of $.33$. That is, developmental ratings were
substantially less lenient than research-based ratings. The subset of studies with administrative purpose included only four samples, so these results should be interpreted only very carefully.

For twelve samples, ratings were based on 360-degree feedback procedures. All of these samples comprised managers. Therefore, 360-degree feedback ratings were compared to those for other managerial samples in moderator analyses. Contrary to the predictions, correlation coefficients for 360-degree feedback procedures \( (r=0.18) \) did not differ from the ones obtained for other managerial samples \( (r=0.19) \). Leniency in self-ratings was reduced in the subset of 360-degree feedback. 360-degree feedback procedures yielded an effect size estimate \( (d=0.33) \), which was significantly lower than the estimate for managerial samples for which ratings had been reported by fewer sources than at least four different perspectives \( (d=0.45) \). As 360-degree feedback is mostly used for developmental purposes, the moderator analyses for "purpose" and "360-degree ratings" resulted in a comparison of almost identical subsets of samples. Thus, no clear decision about which of the two variables caused the moderator effect is possible.

### 5.2.3 Sample composition

In line with previous meta-analyses, job type was found to moderate correlations between self and supervisor ratings. As expected, coefficients were higher when job performance was rated for incumbents of blue-collar jobs \( (r=0.33) \) rather than white-collar jobs \( (r=0.22; Q=15.93, df=1, p=0.0001) \). A test of job-type as a moderator of leniency revealed that ratings for blue-collar samples were considerably more lenient \( (d=0.59) \) than self-ratings obtained from white-collar samples \( (d=0.36; Q=18.41, df=1, p=.0000) \).

Within white-collar samples, effects of higher levels of education and of occupying managerial positions were explored. First, samples with professional-level education were compared to other white-collar samples. White-collar samples in which more than 80% of all respondents held graduate college degrees were compared to samples that had been classified as white-collar, but sampled lower percentages of college graduates with Masters’ or PhD degrees. Substantial
effect size differences appeared: samples with higher levels of education yielded lower inter-rater correlations ($r=.19$) and lower levels of leniency ($d=.29$) than non-professional white-collar samples ($r=.33; d=.54$). Second, managerial samples were compared to other white-collar samples. Managerial samples yielded correlations that were similar to those of high-education samples ($r=.19$), but higher levels of leniency than these ($d=.38$ vs. $d=.29$). Within samples with high averages of educational level, correlations were not significantly lower for managerial samples ($d=.17$) than for non-managerial samples ($d=.22$).

**Percentage of women.** Effect size estimates differed substantially between classes for index $d$ as measure of interrater agreement: as hypothesized, predominantly female samples showed reduced leniency ($Q=52.35$, df=1, $p=.0000$). In contrast, interrater agreement as measured by correlation coefficients was not related to the percentage of women sampled ($Q=.25$, df=1, $p=.6147$).

The examination of gender effects was not without methodological problems, though. As research articles hardly ever reported interrater agreement for female and male self-raters separately, the only way to examine gender effects was to use the percentage of women in the samples as a proxy for the moderator variable intended. Thus, all the samples which reported the gender of their respondents were broken up into two subsets using a median-split on the percentage of women.

As correlational analyses revealed, the percentage of women in the samples was related to the type of job that respondents occupied. Samples with high percentages of women were likely to have lower levels of education and no managerial responsibility. Thus, job-type needs to be controlled for when gender effects were analyzed.

For the fixed-effects analysis controlling for confounds implied the analysis of subsets of samples. Gender composition was examined within white-collar non-managerial samples (blue-collar samples were so predominantly female and managerial samples so predominantly male that breaking up these samples according to their gender composition left too few cases in the
subcategories). In the mixed-effects analysis the two moderators (job-type and gender composition) were entered simultaneously and their interaction was included. Both approaches of analysis supported the same conclusion.

The effect size estimate for index $d$ as measure of interrater agreement was significant, but was based on only three samples that were coded as "non-managerial" and as below the median split regarding the percentage of women in the samples. Thus, this result should hardly be interpreted as a substantial finding and is not discussed as such in the following sections.

Finally, recall the expectation that leniency in self-ratings depended on the samples’ cultural background. As hypothesized, leniency in self-reports disappeared within the subset of Asian samples ($d= -.07$). For non-Asian (Western) samples, the effect size estimate for leniency in self-ratings increased from $d=.35$ to $d=.44$, if the subset of Western samples was analyzed separately. Thus, an estimate of $d=.44$ may describe leniency for Western self-ratings more accurately than the reported overall estimate that was based on all samples in the database.

### Table 6 Moderator variables confirmed by the fixed effects analysis

<table>
<thead>
<tr>
<th>VALIDITY – index $r$</th>
<th>LENIENCE – index $d$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scale format properties</strong></td>
<td>Use of non-judgmental perf. indicators</td>
</tr>
<tr>
<td></td>
<td>Aggregation of heterogeneous scales</td>
</tr>
<tr>
<td><strong>Conditions of report</strong></td>
<td>–</td>
</tr>
<tr>
<td><strong>Characteristics of the respondents sampled</strong></td>
<td>Job-type</td>
</tr>
<tr>
<td></td>
<td>- white-collar vs. blue-collar</td>
</tr>
<tr>
<td></td>
<td>- non-professional vs. professional</td>
</tr>
<tr>
<td></td>
<td>Job-type</td>
</tr>
<tr>
<td></td>
<td>- managerial / professional vs.</td>
</tr>
<tr>
<td></td>
<td>- non-managerial / non-professional vs.</td>
</tr>
<tr>
<td></td>
<td>Cultural background</td>
</tr>
</tbody>
</table>

Are validity coefficients independent of leniency in performance ratings? Throughout this review, leniency in self-ratings and congruence in self-supervisory ratings as measured by
correlation coefficients were investigated independently from each other. The question might arise whether this is really appropriate. For example, inflated ratings might reduce correlation coefficients due to range restriction. To empirically test the independence of leniency and the magnitude of correlations, the two coefficients were correlated for all samples that reported both measures of rater agreement. The resulting correlation confirmed that the indexes vary independently (r = .045, p = .412; n = 333).

*Testing for the effects of model assumptions and explained variance* To reveal the extent to which the results of moderator-analyses depended upon the statistical model chosen for the analysis, I compared fixed-effects results to that of the mixed effects analysis (Overton, 1998). As can be expected, the number of significant moderator effects was reduced for both indicators of rater-agreement when the mixed model was used. One major variable still moderated self-supervisor correlations significantly: the use of non-judgmental performance indicators. Using multi-level modeling software (MLwiN 2.0), the amount of variance explained by each variable was estimated as follows: the use of non-judgmental performance indicators explained about 11% of the variance, the distinction of managerial and non-managerial samples accounted for another 4.6%, that of professional and non-professional for 4.5%, that of white-collar and blue-collar jobs for 1.9%, and the aggregation of heterogeneous items for 1.2% of the between sample variance. Entered simultaneously, these variables explained 19.4% of the between study variance. Two variables remained significant under the assumptions of a mixed model for leniency as measure of rater agreement: the kind of performance construct rated (traits vs. other constructs), and the cultural background of the respondents sampled. The latter two variables, the kind of performance construct and cultural background, explained about 16.5% of the total variance between studies. Among these, cultural background accounted for 15.6% of the total variance explained and effects of the performance constructs rated accounted for the remaining 0.9%.
Table 7 Moderator variables confirmed by the mixed effects analysis

<table>
<thead>
<tr>
<th></th>
<th>VALIDITY – index $r$</th>
<th>LENIENCE – index $d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale format properties</td>
<td>Use of non-judgmental perf. indicators</td>
<td>The performance construct rated:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>traits vs. task- and contextual performance</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Use of non-judgmental performance indicators</td>
</tr>
<tr>
<td>Conditions of report</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Characteristics of</td>
<td>–</td>
<td>Cultural background</td>
</tr>
<tr>
<td>the respondents</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
6 Discussion

The present meta-analysis provided meta-analytical estimates of the validity and leniency of self-ratings for job performance. The study’s main purpose was to investigate relevant features of the context of performance ratings to identify factors that promote (dis)agreement in ratings. Contributions of this meta-analysis come in two areas. First, it examined a rather large number of moderator variables, some of which have not been included in meta-analytical research before. Second, in addition to studying moderators of self-rating validity, this study assessed how moderators affect leniency in self-ratings. The latter is important, as moderators of leniency have not been studied by previous meta-analysts.

6.1 Correlations of self- and supervisor-ratings (the overall validity of self ratings)

The current overall estimate of the sample size weighted correlation of self-supervisory ratings of $r=.22$ is identical to that obtained by earlier meta-analyses, i.e. the studies of Conway and Huffcutt (1997) and Harris and Schaubroeck (1988). Correcting the sample size weighted correlation for attenuation resulted in an estimated population correlation of $r=.33$, a result that is also highly consistent with earlier findings. This is encouraging in that it can be assumed that no unexpected changes have occurred regarding the magnitude of effect sizes. Thus the present results for moderator analyses can well be compared to previous results.

Inspecting the confidence interval for the sample-size weighted mean correlation supported the conclusion that the correlational agreement between self and supervisory appraisals differs significantly from zero. Still, the current results also confirm that self-supervisor correlations are likely to be of rather small magnitude. One possible way of assessing the magnitude of this correlation is to compare it to correlation coefficients that have been reported elsewhere for other sources of performance ratings, such as correlations between supervisor and peer ratings: The
current as well as previous overall estimates of self-supervisor agreement are considerably lower than correlations that have been found between other rating sources (e.g. Harris and Schaubroeck, 1988; Viswesvaran et al., 1996; Conway and Huffcutt, 1997).

6.2 Leniency in self-ratings

The present results confirm the general notion that self-ratings are lenient in comparison with supervisory performance appraisals. An overall sample size weighted mean difference score between self- and supervisor-ratings of $d=.33$ was calculated, based on 70 independent samples. An estimate of $d=.41$ resulted after corrections for attenuation were made. The confidence interval around the mean effect size estimate did not include zero.

The first step in interpreting this result can again be to compare it to those of earlier meta-analytical research. Harris and Schaubroeck (1988) reported a mean standardized difference between self- and supervisor-ratings of $d=.70$, an estimate that is considerably higher than the one found in the present analysis. To explore whether this considerable difference reflects true changes in the database rather than calculational differences, I conducted a separate meta-analysis that included only research articles that had been published before 1988. This analysis in fact produced a higher overall sample size weighted mean difference of $d=.54$, and a corrected estimate of $d=.66$. Thus, the earlier estimate may differ from the present one due to changes in the database that have occurred since Harris and Schaubroeck (1988) presented their results.

Homogeneity analysis showed significant variation of effect sizes across samples. Thus, the next step was to see whether this variation could be explained by moderator variables. Generally, the difference in favorability between self and supervisor ratings remained significant in all the subsets that were examined in moderator-analyses (the only exception were a number of Taiwanese samples examined by Farh, 1991). Thus, at least for Western samples, leniency in self-evaluation seems to be a consistent phenomenon. The extent to which self-ratings were lenient depended on contextual variables, though, as will be discussed below.
6.3 Moderator analyses

Three of the hypothesized moderator variables showed substantial relationship to rater agreement as measured by correlation coefficients: job type, the distinction of judgmental and non-judgmental performance indicators, and the aggregation of heterogeneous items into a single score. Besides, the current review presented meta-analytical evidence that a number of contextual factors influence leniency in self-ratings. These included the kind of performance construct rated (broad vs. behavioral item labels; traits vs. other constructs; contextual vs. task-performance), the purpose of ratings (developmental / 360-degree feedback vs. research-based ratings), respondents expecting a validation of ratings, the type of job sampled, and the cultural background of respondents. In the following section I discuss both confirmed and disconfirmed moderator effects. The discussion is based on the results of fixed-effects moderator analyses to ensure comparability to the results of earlier meta-analyses.

6.3.1 Report format

Scale length. Longer forms were not associated with higher self-supervisory correlations: correlational agreement was the same for single-item as for aggregated measures (not considering the homogeneity/heterogeneity of the aggregate measures). In spite of the fact that aggregate ratings were on average based on 12.1 items, the resulting estimates did not reach higher levels of inter-rater agreement as compared to ratings that were based on single items only. Two explanations are relevant to an interpretation of this finding. First, deficiencies in scale construction could prevent the longer forms from showing up their potential reliability advantage. Adding extra items to a scale should increase the reliability of the resulting measure through eliminating item specific variance and random response error, but only under the condition that the items are designed to measure one common construct. 34 of the articles reviewed reported coefficient alpha for their scales, which ranged from .32 to .95 with a mean of .75 for self-ratings, and from .51
to .96 with a mean of .83 for supervisory ratings. Considering the large range of internal consistencies, aggregate ratings are probably not consistently associated with high levels of intra-rater reliability. In fact, the use of ad-hoc designed scales is not uncommon in the field of research on performance ratings, so that the psychometric properties of the scales are often unknown and possibly of rather low quality. Aggregating several scales or items with low intra-rater consistency can reduce the variability of the scores, so that the total variance of aggregated scores then turns out to be lower than that of scores based on fewer or even single items, resulting in lower correlation coefficients (Hoyt & Kerns, 1999).

It is possible, too, that the reliability advantage of longer scales fails to exert substantial influence on the congruence of self- and supervisory ratings, given it even existed. Certain supporting evidence can be found in a meta-analysis by Viswesvaran et al. (1996) on reliability estimates of performance ratings. These authors found that longer forms were associated with higher intra-rater reliabilities (coefficient alpha). But in this study inter-rater reliabilities did not reflect the advantage that longer forms had in terms of intrarater reliability, i.e. the length of the rating form did not improve interrater reliability. Why might reliability advantages fail to affect interrater correlations positively? An intriguing answer can be found in Wherry and Bartlett’s (1982) approach on the process of performance rating. These authors suggested that three error components arise in performance ratings: measurement error, areal bias, and overall bias. Measurement error, or random response error, can be reduced by adding extra items to a scale. The other two error components describe tendencies in raters to evaluate separate aspects of the target’s performance similarly, being guided by general impressions of a target person. "Areal bias", i.e. bias held against an individual as a performer in a specific area of behavior, can be reduced by integrating items that assess performance in various areas. "Overall bias", i.e. a general bias-effect against an individual regardless of performance domain, can be reduced by integrating ratings made by various raters. Increasing rating accuracy requires that true performance components are increased, and all error components reduced - including measurement error, areal and over-
all bias. Form length may reduce random response error but is likely to leave the other error components unchanged.

In a post hoc analysis, I further scrutinized the finding that scale length (single-item vs. aggregate measures) did not moderate correlational rater agreement. I found that integrating heterogeneous items (from several scales), rather than homogeneous items (from a single scale), influenced rater agreement positively. Heterogeneous aggregated measures performed better than single-items, whereas homogeneous aggregated measures yielded correlations that tended to be even lower than those based on single-items. This result is consistent with the interpretation provided above: using the terminology of Wherry and Bartlett (1982), areal bias is reduced in addition to measurement error if several dimensions of performance are integrated.

Scale definition. I could not find evidence that any of the scale properties that were examined influenced correlational agreement between self- and supervisor-ratings. The present results suggest that social-comparison instructions as well as behavioral scale definitions fail to influence self-supervisory correlations positively. Contrary to these results, Mabe and West (1982) found that social comparison instructions increased self-supervisor agreement (index \( r \)). One possible explanation is that the datasets of primary studies differ between the present analysis and that of Mabe and West (see section "previous research"). Of course, this is a vague proposition that can not be tested here.

In general, job-performance is a complex construct that has many facets. It is possible that this complexity allows raters to avoid negative comparisons with others very successfully, throughout their careers. Unlike in academic contexts, better outcomes of others can to a great extent be attributed to situational advantages rather than superior performance (thinking of performance as a person-quality similar to ability). Remember that subjective evaluations are used mostly in situations where objective criteria are hard to obtain and where it is hard to discount situational influences. Thus, job incumbents may develop quite positive views of their own per-
formance, which they are generally reluctant to question.

The use of behaviorally defined items did not have advantages over using broad scale labels if rater-agreement was measured by correlation coefficients. This result is consistent with earlier findings. Harris and Schaubroeck (1988) could also not confirm that their distinction of trait and behavioral scale labels moderated self-other congruence in ratings, as measured by index \( r \). In contrast, leniency in self-ratings was moderated by the distinction of behavioral and broad scale labels. In accordance with my expectation, less-well-defined scales were associated with higher levels of leniency. Probably the two most important explanations for this finding refer to biased information processing and to impression management. Biased or self-deceptive information processing may be pronounced for broad performance dimensions as these facilitate a selective recall and weighting of information that is in accord with the self-raters’ self-concepts. As an example, Farh and Dobbins (1989) showed that the ambiguity of the performance dimension moderated the relationship between self-esteem and leniency bias in self-ratings. Other research has demonstrated that targets over-rate their own performance less if observers are likely to know about true levels of performance (Aitkenhead, 1984; Baumeister & Jones, 1978), which is more likely to apply to low ambiguity performance dimensions. The discussion of a process model of performance rating (next section) will pick up the question, again, why technical scale properties fail to influence interrater convergence.

**Performance dimensions.** I examined three distinctions between performance constructs that raters evaluated. First, I compared ratings of global performance to ratings of performance in specific domains. Second, I classified the performance dimensions into one of three categories: task-performance, contextual performance, and ratings of traits (comparatively stable characteristics of personality). And finally, indicators of performance were classified as either judgmental or non-judgmental.
The distinction of overall from specific performance did not prove to be associated with differential effects regarding rater agreement. This finding is consistent with that of Harris and Schaubroeck (1988). These authors could not confirm their hypothesis that ratings for overall (global) performance should yield higher correlations than dimensional ratings. Contrary to predictions, ratings of overall performance also did not show higher leniency than ratings of performance in specific domains. The prediction made for this variable was mainly based on the assumption that ratings of overall performance may be associated with higher levels of bias, as low overall ratings are likely to be more threatening to job incumbents than low ratings for specific performance dimensions. No evidence could be found to support this assumption.

Partitioning samples into groups according to a classification of performance constructs as either contextual performance or task performance did not lead to differences in effect size for correlations, but affected leniency in self-ratings. As expected, self ratings of task performance were associated with higher levels of leniency than ratings of contextual performance. Raters can be expected to ascribe higher importance to task performance than to contextual performance (Conway, 1999). Yammarino and Waldman (1993) found that this effect is especially strong for self raters, i.e. ratees’ importance ratings are more closely related to self-ratings of performance. This may explain why a tendency in employees to overrate their own task performance more than their contextual performance was found.

Ratings for traits did not differ from ratings of other performance constructs in terms of correlational agreement between raters. But as predicted, markedly increased levels of leniency resulted if ratings were made for traits. Self-ratings for traits showed even higher levels of leniency than ratings of task performance. One explanation for this particularly strong leniency effect can be that trait ratings do not only address past behaviors but also raters’ expectations on current and future behavioral tendencies. That is, whereas ratings of (previous) task perfor-
mance are directed toward past events, trait ratings are also directed toward the present and the future. Wilson and Ross (2001) recently showed that people appraise their current selves more positively than their past selves which could also be a partial explanation for increased leniency in trait ratings.

The use of non-judgmental performance indicators was related to higher correlations between self and supervisory ratings (but not to leniency in self-ratings). In fact, the resulting effect size estimate for correlational agreement was higher than in any other of all the subsets of samples. This result should be interpreted with caution, as the effect size estimate is based on only five samples. However, the present finding is in accordance with other research. For instance, Hoyt and Kerns (1999) found in a meta-analysis of observer ratings that the use of "explicit" rating criteria reduced the variance attributable to rater bias to up to 5%. In contrast, measures requiring more rater inference yielded ratings with nearly half the variance attributable to bias.

Only five studies reported measures of leniency for ratings that were made for "non-judgmental" performance indicators. The difference of $d=.32$ for judgmental as compared to $d=.40$ for non-judgmental criteria was not significant. Future meta-analyses may be able to reach a more definitive conclusion about the importance of non-judgmental performance criteria as a moderator of leniency on the bases of additional primary studies.

6.3.2 Conditions of report

Confidentiality. Assuring confidentiality to respondents is very much a standard procedure in the context of performance appraisals, but it failed to influence either correlation coefficients or leniency. As mentioned in the hypothesis section, it cannot be ruled out that the reports researchers provide regarding procedures designed to ensure confidentiality were incomplete. This makes the present finding hard to interpret. But as confidentiality is a prominent variable in the context of rater agreement, I decided to not simply discard this variable from the analysis.
After all, it is possible that reports were complete and confidentiality truly fails to influence rater agreement positively.

Note again, that the analysis compared research-based samples only, to avoid a confounding with rating purpose. But it is also true that confidentiality procedures are less common in settings that are not research-based. In administrative or developmental settings, the possibilities of assuring confidentiality may simply be limited by the intended use of ratings. But it is actually these contexts, in which positive effects of confidentiality or anonymity might be of great interest.

Validation expectation. Under the condition that respondents expected their ratings to be validated, mean correlations of self and supervisor ratings tended to reach greater magnitudes. Contrary to Mabe and West (1982) who found that expecting a validation of self-reports was an important moderator of self-criterion agreement, the current study found only a marginally significant moderator effect. Again, this is possibly due to differences in the sample of studies reviewed. Besides, Mabe and West (1982) had included other criteria than performance ratings into their analysis, whereas the current study exclusively examined supervisory ratings as criterion measures. I coded the expectation of a validation to be present, if independent information on the targets’ performance existed (test results; sales commission systems etc.) or if a social validation was to follow appraisals, i.e. if self-raters and supervisors were to meet after appraisals to discuss their evaluations. Consider the latter case: Beyond the fact that expecting a validation may motivate raters to avoid a loss of face and thus to more accurately rate themselves, the size of correlation coefficients depends on additional factors, such as those that Kenny (1991) referred to as a "shared meaning system". The lack of a shared meaning system may explain why expecting a validation of ratings through a feedback meeting fails to influence correlations positively. For job-incumbents, establishing a shared meaning system strongly depends on supervisory feedback. But the amount and quality of feedback is often limited, in organizations
It is possible that expecting a validation can increase self-supervisory-correlations, but only if self-raters have the opportunity to receive sufficient amounts of feedback from their supervisors prior to the appraisals (Zempel & Moser, in press).

The current results indicate a difference in leniency, depending on whether respondents expected their ratings to be validated. As predicted, expecting a validation is associated with more modest self-appraisals. This finding is consistent with the notion that expecting ratings to be validated changes impression-management concerns: Job incumbents fear a loss of face given the possibility that overly positive self-reports could be disconfirmed. Again, this result suggests that raters in performance appraisals consider which implications contextual factors may have for the impressions their self-ratings cause in others.

**Rating purpose.** Reporting evaluations of one’s job performance involves the disclosure of sensitive information. Job incumbents are likely to carefully consider possible consequences their reports could have. The nature of these concerns is likely to depend on appraisal purpose. Accordingly, rating purpose has been examined as a moderator of self-other agreement in performance ratings (Jawahar & Williams, 1997).

Regarding the influence of rating purpose on correlation coefficients, the current results suggest that developmental performance ratings yield correlations comparable to those found in research settings. The effect size estimate for developmental ratings was lower than the overall estimate for samples that were examined for research purposes, but this comparison ignores the fact that developmental ratings were exclusively made for managerial samples. In fact, the difference in effect size disappears if the results for developmental ratings are compared to those for managerial samples that were studied for research purposes. Whether the validity of administrative-based ratings differs from research-based or developmental ratings could not be examined on the basis of the current dataset. Only four studies provided effect size estimates,
and even these four were not homogeneous.

With regard to leniency effects of rating purpose, the present analysis was again limited to a comparison of research-based and developmental ratings (there were only four studies that reported ratings made for administrative purposes). Results showed development-based self-ratings to be less lenient than research-based ratings. Developmental seminars typically use performance ratings to foster self-insight and personal development by highlighting discrepancies between self-ratings and ratings made by others. In this situation, overly optimistic reports might be perceived as undesirable outcomes. Respondents in developmental appraisal settings may take the time and effort needed to report "accurate" self-ratings, as they expect to gain information and support from the appraisals. If job incumbents are interested in learning about their supervisors’ judgments, modest self-ratings can be a means to invite supervisors to provide feedback without concerns for subsequent conflict. Besides, incumbents may expect to receive support as a result of modest self-ratings. For example, they could receive additional support from their supervisors, reduced workloads, or training sessions. In either case, developmental settings may reward less lenient self-ratings.

Research-based appraisals do not provoke inflation in ratings, as might administrative appraisals, but they also do not offer incentives for modest ratings. It could even be that respondents who participate in performance appraisals for research purposes never receive feedback regarding the study’s results. Dispensing individual feedback is often part of a research strategy to increase trust among respondents that their evaluations will remain confidential. Unfortunately, this procedure also renders performance ratings an effort of little consequence, possibly leading raters to not take the time needed to make self-ratings with due consideration (London et al., 1997).
6.3.3 Sample composition

Job type. Correlations between self- and supervisory ratings were higher for blue-collar and less-educated white-collar samples than for managerial samples and samples with higher educational attainment. As predicted, the lowest correlations were observed for managerial samples with high levels of education. These results are in accordance with the findings of earlier meta-analyses which studied how "sampling persons in managerial positions" and "job-complexity levels" affected interrater agreement (Harris & Schraubroeck, 1988; Conway & Huffcutt, 1997). I argue that these effects can be attributed to conceptual disagreement among raters or to rating difficulty. Wherry and Bartlett (1982) derived a complex equation to represent the response which performance ratings involve. This equation described a complete theory of rating by providing a definition of the various factors that determine the accuracy of ratings. As ratee performance is a combination of true aptitude for a job and environmental influences, Wherry and Bartlett (1982) concluded that the nature of the work situation should be an important determinant of rating accuracy. According to these authors, factors of the work situation that should increase rating accuracy include a) working conditions that are constant from worker to worker, b) the ease with which relevant stimuli can be observed, and c) the use of rating items that refer to frequently performed tasks rather than to rare and infrequent behaviors which may also involve long time intervals. All these factors are more typically found in blue-collar work settings.

The current meta-analysis also tested whether job-type moderated leniency in self-ratings. Ratings for employees in blue-collar jobs and white-collar, but non-professional jobs showed comparable levels of leniency. In contrast, leniency in highly educated samples was considerably lower. Leniency also decreased within white-collar samples if managerial rather than non-managerial samples were rated. Cleveland, Morrison and Bjerke (1986) offer an explanation for these findings: Raters are reluctant to evaluate job-incumbents as average or below average for jobs with higher educational requirements (as cited in Murphy and Cleveland, 1995, p.69).
When rating employees with higher educational attainment, raters may consider the professional standing of the group rather than compare group members with each other. Thus others’ ratings reach similar levels of favorability as do self-evaluations. As "leniency" is measured by the mean difference between self- and supervisor-ratings, this measure decreases when others’ ratings are also lenient.

**Percentage of women.** Overall, the samples’ gender composition corresponded with both correlation coefficients and leniency as measures of rater agreement. But these main effects should not be interpreted until important confounds are controlled for. The percentage of women in the samples was clearly correlated with other sample characteristics, such as educational level and managerial responsibility. The percentage of women in the sample did not have an effect on correlation coefficients when educational level was held constant. Predominantly female samples showed reduced leniency even if educational level and managerial responsibility were held constant. But these effects were based on so small numbers of cases that they should hardly be interpreted (only one out of twelve samples that were classified as above the median with respect to the percentage of females was managerial; within non-managerial samples, only three were coded as above the median-split regarding the percentage of women sampled). Thus, no clear conclusion could be reached regarding the effects gender composition has as a potential moderator of leniency. Further research is needed so meta-analytical evidence for gender effects can be generated based on a larger set of primary research articles.

**Culture.** A total of 11 samples from Eastern nations were located for which results of self-supervisor performance ratings were reported. Nine of these sampled workers from Taiwan (Farh et al., 1991), one reported data that was collected in Mainland China (Yu & Murphy, 1993), and another collected data in Hong Kong (Furnham & Stringfield, 1994). All the Taiwanese samples were studied by the same researchers (Farh et al., 1991) who replicated what they called a
modesty bias in nine independent samples. The overall difference of self and supervisory performance appraisals for the nine samples was $d = -0.22$. Yu and Murphy (1993) did not replicate this modesty bias for workers from Mainland China\(^5\), but reported leniency in self-ratings as it is typically found in Western nations. These authors concluded from their findings that more specific differences in national work values should be considered to explain leniency or modesty in self-ratings beyond broad East-West differences in culture. They cite evidence that the modesty bias reported for Taiwanese samples may in fact reflect work values present in the Taiwanese culture, which may be more conducive to modest self-ratings than are work values in other Eastern nations such as Mainland China. Thus, the current investigation of the literature could not illuminate the (non-) generalizability of a modesty bias within Eastern countries. However, the estimate of index $d$ for Western samples changed substantially when all the Asian samples were excluded. A separate analysis for Western samples resulted in a weighted mean difference of $d = -0.44$, which is considerably higher than the overall estimate of leniency that was obtained for the entire dataset ($d = -0.33$). Thus, the estimate of $d = -0.44$ (which results in a corrected estimate of $d = -0.55$), probably represents a better estimate of the overall amount of leniency that characterizes self-ratings in Western nations.

\(^5\)Modesty bias in self-appraisal is often expected due to collectivism in Asian cultures. However, it is not clear whether collectivism is the only relevant variable, as another important culture dimension (power distance) is correlated with collectivism. For example, according to Aycan and Kanungo (2001), one of the functions of performance appraisals is to reinforce the authority structure in high power distance cultures. Lenient self-ratings can thus be considered to undermine the supervisor’s authority, and modest self-ratings help to avoid deviations from social norms (Blanton & Christie, 2003). In addition, self-raters in high power distance cultures are likely to expect their ratings to have little impact (Clive Fletcher & Perry, 2002).
7 A process model of performance appraisal

In the following section, I draw on a model that was presented by Landy and Farr (1980) to organize the variables I examined in a broader framework (see figure 2). The model provides a taxonomy of relevant factors in the appraisal system and visualizes their interrelations. As a core component, it includes the rater's cognitive processes ("rating process") that lead to the actual provision of performance ratings. Factors that influence the outcomes of performance ratings can only exert their influence through changes in these cognitive processes on the part of the raters. With regard to the cognitive stages of the rating process I extend the model suggested by Landy and Farr (1980). In addition, I make propositions concerning where rater bias comes from and at what stages of the rating process it occurs.

Figure 2 Process model of performance rating. A modification and extension of the model suggested by Landy & Farr, (1980)
The model’s core is the raters’ cognitive process, which is embedded in the organization’s administrative processes as well as the broader context of national culture. As it is the cognitive process that determines the outcomes of appraisals I give the most emphasis to discussing this part of the model. I suggest that the rating process comprises of three stages: observation, retrieval and judgment, and communication (Brandstaetter, 1970). Note that the communication component was not included in the Landy and Farr model. I argue that each of these three stages not only describes the rating process but contributes to an understanding of rater congruence. For each of the three stages of the rating process, I specify critical determinants of rater agreement, as well as theoretical approaches that may account for their effects. Theories that specifically apply to self-ratings are incorporated, and the existing meta-analytical evidence is summarized.

**Observation.** Typically, ratings are not provided following immediate observation, but are made in retrospect for longer periods of time, such as a year. Nevertheless, I conceive of observation as the first component of the rating process as the observation of behavior is a prerequisite for appraisals. Observations are what later judgments are based on. Incongruence between two raters’ judgments at this stage is mostly determined by variables that affect the "observability" of performance. A critical determinant of how well and how reliably performance can be observed is the type of position the target occupies (e.g. routine and physical work vs. mental or non-routine work). Other factors include the rater’s opportunity to observe, the rater’s motivation to observe, how representative observed behavior is, or different raters observing different behavior. I suggest that these factors contribute to lowering agreement between raters, mainly because they influence rating difficulty and conceptual disagreement among raters.

In the process model I link position characteristics to outcomes at the stage of behavior observation. Other factors that should mainly affect the "observation-stage", such as the ones mentioned above, have not been included into field research of performance ratings, probably because they are difficult to obtain. The most prominent variables among the group of position charac-
teristics that have been included in meta-analytical research are job-complexity and managerial responsibility (Conway & Huffcutt, 1997; Harris & Schraubroeck, 1988). In accordance with earlier research, I could confirm that educational level as a proxy for job-complexity moderated rater agreement. Position type is a comprehensive category that determines various characteristics of an appraisal situation simultaneously, such as the kind of performance dimensions included in the ratings, idiosyncrasy, or the extent to which evaluations involve inference. As the observability of performance exerts its influence on rating outcomes mostly through its effects on rating difficulty and conceptual disagreement, it should mainly be correlational coefficients that reflect the observability of performance. In fact, the current as well as earlier meta-analyses found job type to moderate self-supervisory correlations.

Retrieval and judgment. Behavior observation lays the foundations for later judgment. At the second stage, the rater has to reach a judgment based on earlier observation of the target. Again, the level of inference and conceptual disagreement involved in the ratings are important determinants of rater congruence. But now, scale development and the rating instrument are major intervening variables. The rating instrument defines what dimensions of performance are assessed and how these are depicted. Independent from the observations that raters made, the stimuli and clues provided by the rating instrument influence rater behavior: the extent to which raters lack a common understanding of the behavior they are to evaluate, do not share common standards of evaluation, or use idiosyncratic or implicit theories, may be moderated by properties of the rating instrument.

In the present analysis, the use of non-judgmental performance indicators contributed to a substantial reduction in (conceptual) disagreement among raters (see also Cheung, 1999; Hoyt & Kerns, 1999). It is worth noting that none of the technical scale format properties (e.g. social comparison scales, behavioral item descriptions) substantially influenced correlational rater
agreement. These format properties are all means to structure the observation of job performance. Why should they still fail to influence (correlational) rater agreement positively? I assume that these factors can indeed influence rater agreement but only provided that raters really use performance scales to guide their observations. However, this is not common practice in performance appraisals as they are conducted in organizations. In practice, raters typically rely on their recollections of past observations and base the evaluation process on set impressions of the target.

Unlike correlation coefficients, mean difference scores (leniency) were influenced by the rating instrument, i.e. by scale properties. I suggest that this is due to the fact that leniency reflects non-deliberate distortion in ratings. These may include general rater tendencies or stereotypical judgment (for instance regarding age, gender, or race). Cognitive factors may also apply. As Campbell and Lee (1988) argued, appraisals are often simplified as raters lack the ability or the motivation to provide highly detailed ratings. Instead, raters often use general impressions or schemas in their evaluations.

One such effect was observed in the present meta-analysis, namely a leniency effect of job-type (i.e. educational level). Comparing mean supervisor-ratings\(^6\) showed that supervisors rated employees with higher levels of education at higher levels of favorability, even though their task was not to differentiate between groups of employees in different positions but to assess employees in rather similar positions. As self-ratings are generally lenient, this behavior can explain to a large part why managerial and other more highly educated respondents yielded lower leniency scores in self-ratings.

Another case of non-deliberate distortion in ratings is specific to self-ratings: self-deception that is motivated by self-protective or self-enhancement needs. To some extent, it can be expected that self-raters reach rather positive evaluations regardless of their "true" performance as it may have been observed by others. Self-raters can unconsciously inflate self-ratings in order

---

\(^6\)In a post hoc analysis, the average level of supervisor ratings depending on the respondents’ educational level were calculated.
to maintain self-esteem and to protect their individual’s self-concepts. Two of the present results can be interpreted from the perspective of self-deceptive reporting: using traits as scale labels and providing raters with broad item labels rather than behavioral descriptions. Both variables were associated with increased leniency in self-ratings. Especially negative evaluations of personality characteristics may be threatening to a person’s self-concept, and self-deceptive responding is a comparatively likely reaction.

In sum, mainly due to the operation of conceptual disagreement and non-deliberate distortion, I suggest that both diminished correlations and leniency effects result at the second stage of the appraisal process.

*Communication.* At the third stage of the rating process, raters have to communicate their judgments by means of a final rating. This communication is likely to be goal-directed. Raters can have various goals, such as being accurate, preventing trouble, or gaining rewards. Thus, it can be expected that, at the communicational level, raters may and often will systematically distort their ratings. These deliberate distortions will mostly concern the mean level of the rating made, i.e. leniency effects. I suggest that "conditions of report" play a central role in motivating raters to distort the favorability of ratings. This assumption received some support in the present analysis: conditions of report had effects on leniency but not on validity (correlation coefficients).

The most prominent variable among the report conditions is rating purpose. But other moderating effects may be interpreted in similar ways, including effects of the performance dimensions rated (e.g. task- vs. contextual performance), or the expectation that ratings might be validated. Rating purpose explicitly appears in the process model as it is probably the most important determinant of intentional distortions in ratings. Besides, purpose is likely to be associated with other "conditions of report" (e.g. ratings for administrative use can hardly be confidential).
The impression-management view is of high relevance for the communication component of performance rating. The impression-management approach suggests that employees manage their self-presentations to serve personal goals and to avoid negative outcomes that may be related to performance ratings. Impression-management behavior in performance appraisal has mostly been assessed through its hypothesized influence on the favorability of ratings (e.g. Aitkenhead, 1984). The present study confirms that effects that are likely to be attributable to impression-management influence leniency in ratings. These include the effects of such variables as rating purpose, expecting a validation of ratings, the overrating of task-performance, and culture. Rating purpose is a prominent moderator that can be interpreted from an impression-management view. Research has primarily examined the hypothesis that ratings for administrative purposes should show increased levels of leniency in ratings (e.g. Harris et al., 1995; Jawahar & Williams, 1997). Harris et al. (1995) had raters evaluate subjects for varying purposes and confirmed the expected leniency effect. Jawahar and Williams (1997) present a meta-analysis of experimental and field research ratings of "paper people" and video-taped subjects and also found that ratings made for administrative purposes were more lenient. Generally, more favorable performance ratings can be expected to be desirable in organizations as far as appraisals are usually associated with organizational rewards. This interpretation receives further support if different situational incentives (e.g. for modest ratings) can be shown to decrease leniency in self-ratings. In fact, I found self-ratings to be substantially less lenient in developmental settings as compared to research-based ratings. Another leniency effect that can very well be attributed to impression-management behavior regards effects of the relevance of the performance dimension rated. If respondents consider possible consequences their ratings may have, the perceived importance of rating dimensions should increase leniency in ratings (given that favorable ratings are more desirable). Fox, Caspy and Reisler (1994), for example, varied the relevance that performance dimensions had to the assessment setting, and found that relevance moderated the amount of leniency in self-appraisals. Similarly, the present study found that task-performance was rated with
higher levels of leniency by self-raters than was contextual performance. This may be attributed to the fact that task-performance is likely to be of higher relevance to self-raters.

*Rater characteristics.* The third of the context components in the appraisal system, rater and ratee characteristics, likely has a complex pattern of possible influences. For instance, a rater’s intelligence should affect the entire cognitive process, from observation to communication. Response sets that raters bring to the rating task are often related to rater characteristics (e.g. age, gender, race, or work-related personality or behavior styles). Effects that rater characteristics have on rating outcomes may show as both, direct effects and as interactions between rater and ratee characteristics (e.g. a male subordinate rating a female boss). Other rater characteristics are specifically relevant to self-ratings. These include self-esteem (Fahr & Dobbins, 1989), self-awareness (Fletcher, & Baldry, 2000; Nasby, 1989), or narcissism (Campbell, Reeder, Sedikides, & Elliot, 2000).

As mentioned above, the model presented here suggests that rater characteristics exert their influence mostly during judgment formation. Their effects should show as non-deliberate rater bias. But based on the present set of samples, little evidence could be generated for the effects of rater characteristics. The reason is that these are rarely reported in field research studies.

*Culture.* Finally, culture determines to a great extent the context in which appraisal systems function. It has been suggested that in Asian countries modesty applies to self-ratings rather than leniency, which is so characteristic of "Western" performance ratings (Farh, Dobbins, & Cheng, 1991). Recent work in social psychology has emphasized theories of self-construal to understand cultural differences in cognition, emotion, and motivation (Markus, & Kitayama, 1991). The "self" is constructed from cultural contexts: self-views originate from engaging in these contexts, their social practices and institutions (Heine, Lehman, Markus, & Kitayama, 1999). As Heine, Lehman, Markus, & Kitayama (1999) suggested, the need for positive self-regard is
not necessarily universal, and self-serving bias may at least show significant variability across cultures (Amy, Abramson, Hyde, & Hankin, 2004). As management practices in organizations reflect such fundamental cultural differences, cultural background should be included in any model of performance appraisal.

*Data analysis and personnel action.* Whenever performance ratings are collected for organizational purposes, appraisal outcomes are likely to serve as a basis for personnel action. Personnel action includes for instance the promotion of employees, the determining of salaries, terminations, or the launch of training sessions. This way, position characteristics, organizational characteristics, and also employee performance, may eventually change due to the action taken in response to appraisal outcomes. Subsequent performance appraisals are then made under the resulting - differing - circumstances.
8 Limitations and future research

The moderators examined in the present meta-analysis could not explain all the variation in effect sizes between samples. Other variables that are relevant to an understanding of rater convergence in performance appraisals are likely to exist, which were not examined in the current investigation. For instance, the present study could not study many individual difference variables. In fact research in field settings has not accumulated a large database addressing the effects individual differences have on performance ratings. Besides, some individual difference variables (such as gender) are typically confounded with other variables (such as the type of job sampled). To allow for a meta-analytical study of these effects, field researchers had to control for confounds or needed to report effect sizes for subgroups of their samples separately, more often than has been the case. Another variable that has not often been examined in field settings is rater training. Effects of rater training have mostly been studied in experimental settings (Woehr & Huffcutt, 1994). In fact, none of the studies included in the current review reported that training sessions were held. It might be of great practical interest, though, to know whether training sessions actually have the intended effects. Finally, an important variable that deserves the attention of future research is "culture". It should be of great interest to accumulate a larger database of research articles that allows for an examination of the effects cultural background has on rater behavior. The current meta-analysis could only distinguish between Western and Asian/Eastern cultures. Moreover, within the category of Asian countries, the majority of samples stemmed from a single Asian country (Republic of China). The approach taken in the current study should not suggest that very broad distinctions of cultures suffice to guide the analysis of the role culture plays in performance evaluation or self-appraisals. Both within "Western" and "Asian" countries, cultures differ substantially. For instance, more research in diverse Asian countries is needed to decide on the generalizability of the "modesty-hypothesis" within Asian countries. Besides, theoretical explanations as to how culture influences rating behavior should be advanced. Recent
work in social psychology has focused on theories of self-construal (Heine, Lehman, Markus, & Kitayama, 1999; Markus, & Kitayama, 1991). The implications this body of theory has for rater agreement in various cultures (as well as for gender differences) should be of interest for future research.

The meta-analysis presented here analyzed field research and excluded experimental studies. Whereas this approach gives credibility to results with regard to their validity outside of research laboratories, it is also one of the study’s limitations. Analyzing experimental work could allow for the examination of a more diverse set of moderator variables. Testing more complex hypotheses regarding moderator effects, and more differentiated theory building might result from analyzing research conducted in laboratory settings.

Another limitation is related to the fact that a multitude of theoretical approaches exists that researchers can refer to in order to reach an understanding of rater agreement. I presented a model comprising propositions on the most relevant mechanisms that determine the outcomes of ratings at various stages in the rating process. Such a model cannot be comprehensive with respect to all potentially relevant moderators or the entire body of theory that has been applied to performance rating.

It is difficult to reach general recommendations for management practice, but a few suggestions may follow from the present study. First, practitioners in the area of performance appraisal should think about combining information from several performance dimensions into composite scores. Second, the use of personality related performance measures and broad item-labels should be avoided. Low ratings on these dimensions may unnecessarily offend the targets and are likely to possess rather low validity. Third, opting for non-judgmental performance indicators, if possible, may be a powerful way of increasing (self-other) convergence in performance ratings - whenever this is relevant. Non-judgmental or objective performance criteria are some-
times introduced to meet legal concerns. The present analysis provides some evidence that using non-judgmental performance indicators also has the potential to alter the psychometric qualities of performance ratings. Finally, and - most importantly - , in designing rating instruments, practitioners may want to consider important motivational consequences that performance ratings have, i.e. they may want to take on the perspective of self-raters.
References


Dalton, M. A. (1997). When the purpose of using multi-rater feedback is behavior change. In D. Bracken (Ed.), Should 360-degree feedback be used only for developmental purposes? (pp. 1-6). Greensboro, NC: Center for Creative Leadership.


feedback be used only for developmental purposes? (pp. 7-9). Greensboro, NC: Center for Creative Leadership.


Appendix
### A Coding criteria

#### Table 8 Coding criteria

<table>
<thead>
<tr>
<th>Hyp.</th>
<th>Moderator variable</th>
<th>Categories coded</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Report format</td>
<td></td>
</tr>
<tr>
<td>1)</td>
<td>Single-item vs. aggregated measure</td>
<td>1a) single-item measures</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1b) aggregated: several items/scales were integrated</td>
</tr>
<tr>
<td>2)</td>
<td>Broad vs. behavioral scale definition</td>
<td>2a) behavioral definition: behavior descriptions are provided (both Likert scales and BES / BARS)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2b) broad scale labels: unelaborated performance constructs</td>
</tr>
<tr>
<td>3)</td>
<td>Social comparison</td>
<td>3a) absolute scale anchors and none of the below</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3b) comparative scale anchors</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3b) raters rank entire group of ratees</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3b) social-comparison instruction</td>
</tr>
<tr>
<td>4)</td>
<td>Global vs. dimensional ratings</td>
<td>4a) ratings for global/overall job performance (e.g. &quot;overall performance&quot;, &quot;overall effectiveness&quot;)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4b) ratings for specific performance dimensions (e.g. &quot;oral presentation&quot;)</td>
</tr>
<tr>
<td>5)</td>
<td>Contextual- vs. task performance vs. ratings for traits</td>
<td>5a) task-performance (as defined in Borman and Motowidlo, 1993)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5b) contextual performance (as defined in Borman and Motowidlo, 1993)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5b) ratings for traits (qualities of personality)</td>
</tr>
<tr>
<td>6)</td>
<td>(Non)-judgmental performance indicators</td>
<td>6a) performance indicator is objective (e.g. financial indicators, hourly productivity rates, ...) rather than</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6b) judgmental in nature (&quot;leadership ability&quot;)</td>
</tr>
<tr>
<td>7)</td>
<td>Confidentiality of ratings</td>
<td>7a) via instruction: confidentiality is assured</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7a) via procedure: respondents send back questionnaires in sealed envelopes to researchers</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7b) ratings were not considered to be anonymous if data were forwarded to personnel departments or if feedback meetings were held</td>
</tr>
<tr>
<td></td>
<td>Validation expectation</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>------------------------</td>
<td>---</td>
</tr>
<tr>
<td>8)</td>
<td><strong>Validation expectation</strong></td>
<td>8a)</td>
</tr>
<tr>
<td>9)</td>
<td><strong>Rating purpose</strong></td>
<td>9a)</td>
</tr>
<tr>
<td>10)</td>
<td><strong>360-Degree-Feedback</strong></td>
<td>10a)</td>
</tr>
</tbody>
</table>

**SAMPLE COMPOSITION**

<table>
<thead>
<tr>
<th></th>
<th>Job type</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>11)</td>
<td>blue-collar</td>
<td>11a)</td>
<td>manual/little skilled service jobs</td>
<td>11b)</td>
<td>more highly skilled than 11a)</td>
<td></td>
</tr>
<tr>
<td>12)</td>
<td>white-collar; managerial position; non-managerial sample;</td>
<td>12a)</td>
<td>sample is explicitly described as consisting of managers</td>
<td>12b)</td>
<td>not 11c)</td>
<td></td>
</tr>
<tr>
<td>13)</td>
<td>high educational level; lower educational level</td>
<td>13a)</td>
<td>MA/PhD - degrees ≥ 80%</td>
<td>13b)</td>
<td>MA/PhD - degrees ≤ 80%</td>
<td></td>
</tr>
<tr>
<td>14)</td>
<td>Gender composition</td>
<td>14a)</td>
<td>percentage of females in the sample</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15)</td>
<td>Culture: Asian vs. Western samples</td>
<td>15a)</td>
<td>respondents’ country of origin is Asian</td>
<td>15b)</td>
<td>respondents’ country of origin is not Asian</td>
<td></td>
</tr>
</tbody>
</table>
## B The classification of performance dimensions

Table 9 The classification of performance dimensions

<table>
<thead>
<tr>
<th>Dimensions distinguished</th>
<th>Subdimensions</th>
<th>Examples from primary research</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A Contextual Performance</strong></td>
<td>1 Volunteering to carry out task activities that are not formally part of the job</td>
<td>1 initiative, process improvement, innovation, adopting new approaches, ...</td>
</tr>
<tr>
<td></td>
<td>2 Persisting with enthusiasm and extra effort as necessary to complete own task activities successfully</td>
<td>2 personal motivation, desire to work, resilience, effort, motivation and energy, conscientiousness, enthusiasm, ...</td>
</tr>
<tr>
<td></td>
<td>3 Helping and cooperating with others</td>
<td>3 maintaining good atmosphere, cooperation, helping, ...</td>
</tr>
<tr>
<td></td>
<td>4 Following organizational rules and procedures even when it is personally inconvenient</td>
<td>4 discipline, following work procedures, ...</td>
</tr>
<tr>
<td></td>
<td>5 Endorsing, supporting, and defending organizational objectives</td>
<td>5 representing, organizational commitment</td>
</tr>
<tr>
<td><strong>B Task-performance</strong></td>
<td>1 Technical-administrative task performance</td>
<td>1 business judgement, specialist knowledge, problem solving and analysis, written communication,...</td>
</tr>
<tr>
<td></td>
<td>2 Leadership task-performance</td>
<td>2 effectiveness in people management, leadership,</td>
</tr>
<tr>
<td><strong>C Overall performance</strong></td>
<td></td>
<td>1 quantity and quality of performance, overall effectiveness, summated performance scores</td>
</tr>
</tbody>
</table>

Performance dimensions were adapted from Borman and Motowidlo (1993);
Intercoder reliability: 81% agreement, Kappa=.71
C Overall results based on the mixed-model

Table 10 Overall results based on the mixed-effects model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>S.E.</th>
<th>Parameter</th>
<th>Estimate</th>
<th>S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed</td>
<td>β0</td>
<td>.225</td>
<td>.014</td>
<td>β0</td>
<td>.331</td>
</tr>
<tr>
<td>Level 2</td>
<td>σ²_u0</td>
<td>.013</td>
<td>.003</td>
<td>σ²_u0</td>
<td>.177</td>
</tr>
<tr>
<td>Level 1</td>
<td>var(e₁)</td>
<td>1</td>
<td>0</td>
<td>var(e₁)</td>
<td>1</td>
</tr>
</tbody>
</table>

Note. The analysis was set up as a simple random intercept model:

\[
outcome_{ij} \sim N(XB, \Omega)
\]

\[
outcome_{ij} = \beta_0 + u_{0j} + e_{1ij} \cdot stderr_{ij}
\]

\[
[u_{0j}] \sim N(0, \Omega_u) : \Omega_u = [\sigma_u^2]
\]

\[
[e_{1ij}] \sim N(0, \Omega_e) : \Omega_e = [\sigma_e^2]
\]

That is, the analysis was set up as a two-level model, with the effect-sizes at the first level, and the studies at the second. The predictor sampling error \( stderr_{ij} \) was included on level one, in the random part only, with a coefficient fixed at one (the lowest level variance \( var(e₁) \) is 1). The regression constant \( \beta_0 \) was included in both the fixed and the random part at level two.

The sampling variance for effect-size index \( r \) is:

\[
1/(n - 3), \text{ where } n \text{ is the number of cases.}
\]

The sampling variance for effect-size index \( d \) is:

\[
\frac{n_E + n_C}{n_E \cdot n_C} + \frac{\sigma^2}{2n_E + n_C}, \text{ where } n_E \text{ and } n_C \text{ are the number of cases in the experimental group and the control, respectively.}
\]

and \( d \) the effect-size estimate:

\[
d = \frac{\bar{Y}_E - \bar{Y}_C}{SD}, \text{ i.e. the standardized difference between the two group means } \bar{Y}_E \text{ and } \bar{Y}_C.
\]
### D Moderator analyses based on the mixed-model

Table 11 Moderator results based on the mixed-effects model

<table>
<thead>
<tr>
<th>Scale definition: Broad vs. behavioral item definition</th>
<th>Parameter</th>
<th>Estimate</th>
<th>S.E.</th>
<th>Parameter</th>
<th>Estimate</th>
<th>S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed β0 (cons)</td>
<td>.238</td>
<td>.021</td>
<td>β0 (cons)</td>
<td>.242</td>
<td>.069</td>
<td></td>
</tr>
<tr>
<td>Fixed β1 (broad)</td>
<td>.004</td>
<td>.032</td>
<td>β1 (broad)**</td>
<td>.032</td>
<td>.055</td>
<td></td>
</tr>
<tr>
<td>Level 2 σ²_u</td>
<td>.013</td>
<td>.002</td>
<td>σ²_u</td>
<td>.177</td>
<td>.032</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scale definition: Social comparison</th>
<th>Parameter</th>
<th>Estimate</th>
<th>S.E.</th>
<th>Parameter</th>
<th>Estimate</th>
<th>S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed β0 (cons)</td>
<td>.223</td>
<td>.015</td>
<td>β0 (cons)</td>
<td>.346</td>
<td>.062</td>
<td></td>
</tr>
<tr>
<td>Fixed β1 (social comp.)</td>
<td>.054</td>
<td>.036</td>
<td>β1 (social comp.)</td>
<td>-.168</td>
<td>.247</td>
<td></td>
</tr>
<tr>
<td>Level 2 σ²_u</td>
<td>.013</td>
<td>.002</td>
<td>σ²_u</td>
<td>.181</td>
<td>.033</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Performance construct: Global vs. specific + aggregated measure vs. single-item</th>
<th>Parameter</th>
<th>Estimate</th>
<th>S.E.</th>
<th>Parameter</th>
<th>Estimate</th>
<th>S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed β0 (cons)</td>
<td>.242</td>
<td>.024</td>
<td>β0 (cons)</td>
<td>.410</td>
<td>.068</td>
<td></td>
</tr>
<tr>
<td>Fixed β1 (aggregated)</td>
<td>-.048</td>
<td>.030</td>
<td>β1 (aggregated)**</td>
<td>-.197</td>
<td>.063</td>
<td></td>
</tr>
<tr>
<td>Fixed β2 (global)</td>
<td>.002</td>
<td>.023</td>
<td>β2 (global)</td>
<td>.064</td>
<td>.039</td>
<td></td>
</tr>
<tr>
<td>Level 2 σ²_u</td>
<td>.013</td>
<td>.002</td>
<td>σ²_u</td>
<td>.175</td>
<td>.031</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Performance construct: Task-performance contextual-performance trait</th>
<th>Parameter</th>
<th>Estimate</th>
<th>S.E.</th>
<th>Parameter</th>
<th>Estimate</th>
<th>S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed β0 (cons)</td>
<td>.226</td>
<td>.015</td>
<td>β0 (cons)</td>
<td>.320</td>
<td>.052</td>
<td></td>
</tr>
<tr>
<td>Fixed β1 (contextual)</td>
<td>.010</td>
<td>.010</td>
<td>β1 (contextual)</td>
<td>-.024</td>
<td>.015</td>
<td></td>
</tr>
<tr>
<td>Fixed β2 (trait)**</td>
<td>-.039</td>
<td>.015</td>
<td>β2 (trait)**</td>
<td>.151</td>
<td>.026</td>
<td></td>
</tr>
<tr>
<td>Level 2 σ²_u</td>
<td>.013</td>
<td>.002</td>
<td>σ²_u</td>
<td>.176</td>
<td>.032</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Performance construct: (non)-judgmental performance indicators</th>
<th>Parameter</th>
<th>Estimate</th>
<th>S.E.</th>
<th>Parameter</th>
<th>Estimate</th>
<th>S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed β0 (cons)</td>
<td>.220</td>
<td>.013</td>
<td>β0 (cons)</td>
<td>.354</td>
<td>.057</td>
<td></td>
</tr>
<tr>
<td>Fixed β1 (non-judgm.)**</td>
<td>.167</td>
<td>.039</td>
<td>β1 (non-judgm.)**</td>
<td>-.647</td>
<td>.074</td>
<td></td>
</tr>
<tr>
<td>Level 2 σ²_u</td>
<td>.012</td>
<td>.002</td>
<td>σ²_u</td>
<td>.182</td>
<td>.033</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Conditions of report: Validation expected</th>
<th>Parameter</th>
<th>Estimate</th>
<th>S.E.</th>
<th>Parameter</th>
<th>Estimate</th>
<th>S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed β0 (cons)</td>
<td>.223</td>
<td>.015</td>
<td>β0 (cons)</td>
<td>.325</td>
<td>.065</td>
<td></td>
</tr>
<tr>
<td>Fixed β1 (expected)</td>
<td>.059</td>
<td>.037</td>
<td>β1 (expected)</td>
<td>-.019</td>
<td>.177</td>
<td></td>
</tr>
<tr>
<td>Level 2 σ²_u</td>
<td>.013</td>
<td>.002</td>
<td>σ²_u</td>
<td>.181</td>
<td>.032</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Conditions of report: Confidentiality</th>
<th>Parameter</th>
<th>Estimate</th>
<th>S.E.</th>
<th>Parameter</th>
<th>Estimate</th>
<th>S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed β0 (cons)</td>
<td>.249</td>
<td>.033</td>
<td>β0 (cons)</td>
<td>.367</td>
<td>.139</td>
<td></td>
</tr>
<tr>
<td>Fixed β1 (confidential)</td>
<td>-.014</td>
<td>.038</td>
<td>β1 (confidential)</td>
<td>-.097</td>
<td>.158</td>
<td></td>
</tr>
<tr>
<td>Level 2 σ²_u</td>
<td>.012</td>
<td>.002</td>
<td>σ²_u</td>
<td>.179</td>
<td>.032</td>
<td></td>
</tr>
</tbody>
</table>

continued next page
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>S.E.</th>
<th>Parameter</th>
<th>Estimate</th>
<th>S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_0$ (cons)</td>
<td>.249</td>
<td>.035</td>
<td>$\beta_0$ (cons)</td>
<td>.336</td>
<td>.068</td>
</tr>
<tr>
<td>Fixed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$ (appraisal)</td>
<td>.151</td>
<td>.076</td>
<td>$\beta_1$ (appraisal)</td>
<td>.237</td>
<td>.272</td>
</tr>
<tr>
<td>Fixed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_2$ (development)</td>
<td>-.061</td>
<td>.034</td>
<td>$\beta_2$ (development)</td>
<td>-.118</td>
<td>.143</td>
</tr>
<tr>
<td>Level 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma_u^2$</td>
<td>.013</td>
<td>.002</td>
<td>$\sigma_u^2$</td>
<td>.178</td>
<td>.032</td>
</tr>
</tbody>
</table>

**Conditions of report: Appraisal, development, vs. research**

**Conditions of report: 360-degree + non-managerial vs. managerial samples**

**Sample characteristics: Job-type – white-collar vs. blue collar**

**Sample characteristics: Job-type – non-professional vs. professional**

**Sample characteristics: Job-type – non-managerial vs. managerial**

**Sample characteristics: Percentage of women (median split)**

**Sample characteristics: Culture – Asian vs. Western**

**Note.** The analysis was set up as a two-level model, with the effect-sizes at the first level, and the studies at the second. The predictor sampling error was included on level one, in the random part only, with a coefficient fixed at one. The regression constant $\beta_0$ was included in both the fixed and the random part at level two. Explanatory variables (with regression coefficients $\beta_1$, $\beta_2$, ...) were entered as fixed effects. $\sigma_u^2$ is the true variation between the studies. To deal with the problem of missing values, these were recoded into the means of each variable for each covariate, and dummy variables that indicated missingness were entered along with the covariates. Thus it was also possible to test whether missing cases carried information.
E  Intercorrelation matrix of moderator variables
### Table 12 Intercorrelation matrix of moderator variables

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 broad vs. behavioral</td>
<td>1.00</td>
<td>0.07</td>
<td>-0.32</td>
<td>-0.18</td>
<td>-0.03</td>
<td>0.25</td>
<td>-0.11</td>
<td>0.05</td>
<td>0.11</td>
<td>-0.10</td>
<td>0.26</td>
<td>0.11</td>
<td>0.22</td>
<td>-0.39</td>
<td>-0.12</td>
<td>-0.41</td>
</tr>
<tr>
<td>2 global vs. specific</td>
<td>1.00</td>
<td>.</td>
<td>.</td>
<td>0.10</td>
<td>0.08</td>
<td>0.01</td>
<td>0.25</td>
<td>0.31</td>
<td>-0.30</td>
<td>0.15</td>
<td>-0.01</td>
<td>-0.12</td>
<td>0.37</td>
<td>-0.01</td>
<td>-0.27</td>
<td></td>
</tr>
<tr>
<td>3 task vs. contextual</td>
<td>1.00</td>
<td>.</td>
<td>.</td>
<td>-0.09</td>
<td>-0.11</td>
<td>0.03</td>
<td>0.01</td>
<td>0.05</td>
<td>0.07</td>
<td>0.01</td>
<td>-0.06</td>
<td>-0.04</td>
<td>-0.10</td>
<td>0.06</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>4 trait</td>
<td>1.00</td>
<td>-0.05</td>
<td>-0.12</td>
<td>-0.16</td>
<td>-0.18</td>
<td>0.17</td>
<td>-0.06</td>
<td>-0.01</td>
<td>-0.06</td>
<td>-0.15</td>
<td>-0.30</td>
<td>-0.16</td>
<td>-0.21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 (non)-judgmental</td>
<td>1.00</td>
<td>0.07</td>
<td>0.02</td>
<td>-0.03</td>
<td>0.18</td>
<td>-0.19</td>
<td>-0.19</td>
<td>-0.10</td>
<td>0.07</td>
<td>0.05</td>
<td>-0.07</td>
<td>-0.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 nr. of dimensions</td>
<td>1.00</td>
<td>0.09</td>
<td>0.12</td>
<td>0.09</td>
<td>-0.28</td>
<td>0.23</td>
<td>0.43</td>
<td>0.12</td>
<td>0.21</td>
<td>-0.15</td>
<td>-0.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 social comparison</td>
<td>1.00</td>
<td>-0.08</td>
<td>-0.16</td>
<td>0.06</td>
<td>-0.06</td>
<td>-0.11</td>
<td>-0.07</td>
<td>0.19</td>
<td>-0.15</td>
<td>0.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 confidentiality</td>
<td>1.00</td>
<td>-0.09</td>
<td>.</td>
<td>0.12</td>
<td>-0.09</td>
<td>-0.10</td>
<td>0.37</td>
<td>0.24</td>
<td>-0.12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 validation expected</td>
<td>1.00</td>
<td>-0.58</td>
<td>0.21</td>
<td>0.27</td>
<td>0.12</td>
<td>0.06</td>
<td>-0.13</td>
<td>-0.31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 purpose</td>
<td>1.00</td>
<td>-0.35</td>
<td>-0.43</td>
<td>-0.09</td>
<td>-0.30</td>
<td>0.13</td>
<td>0.13</td>
<td>0.13</td>
<td>0.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11 managerial</td>
<td>1.00</td>
<td>0.48</td>
<td>0.33</td>
<td>0.46</td>
<td>0.01</td>
<td>-0.45</td>
<td>0.13</td>
<td>0.13</td>
<td>0.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12 360-degree</td>
<td>1.00</td>
<td>0.13</td>
<td>0.23</td>
<td>-0.16</td>
<td>-0.20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13 blue/white</td>
<td>1.00</td>
<td>0.14</td>
<td>-0.02</td>
<td>-0.11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14 professional</td>
<td>1.00</td>
<td>.</td>
<td>-0.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15 culture</td>
<td>1.00</td>
<td>0.01</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16 % women</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Correlations were calculated by using the SPSS factor analysis procedure (FACTOR [correlations]). Empty cells are due to small number of cases.
F The distributions of effect-size indexes $r$ and $d$

Figure 3 Distributions of effect-size index $r$ and $d$
G Effect-sizes and moderator variables as they were extracted from primary research
<table>
<thead>
<tr>
<th>Number</th>
<th>Author(s)</th>
<th>Year</th>
<th>Source</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yammarino, Dubinsky, &amp; Hartley</td>
<td>1987</td>
<td>Journal of Occupational and Organizational Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>2</td>
<td>Farh, Werbel, &amp; Bedeian</td>
<td>1988</td>
<td>Personnel Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>3</td>
<td>Becker &amp; Kimonis</td>
<td>1989</td>
<td>Personnel Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>4</td>
<td>Atwater &amp; Yammarino</td>
<td>1992</td>
<td>Personnel Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>5</td>
<td>Pritchard &amp; Sanders</td>
<td>1973</td>
<td>Journal of Applied Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>6</td>
<td>Lane &amp; Herriot</td>
<td>1990</td>
<td>Journal of Occupational and Organizational Psychology</td>
<td>UK</td>
</tr>
<tr>
<td>7</td>
<td>Gardner, Durham, Cummings, &amp; Pierce</td>
<td>1989</td>
<td>Journal of Occupational Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>8</td>
<td>McEnery &amp; McEnery</td>
<td>1987</td>
<td>Journal of Applied Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>9</td>
<td>Yammarino &amp; Waldman</td>
<td>1993</td>
<td>Journal of Occupational and Organizational Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>10</td>
<td>Brief, Aldag, &amp; Van Sell</td>
<td>1977</td>
<td>Journal of Occupational Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>11</td>
<td>Rossie &amp; Kraus</td>
<td>1983</td>
<td>Journal of Occupational Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>12</td>
<td>Warr, P.</td>
<td>1999</td>
<td>Journal of Occupational and Organizational Psychology</td>
<td>UK</td>
</tr>
<tr>
<td>13</td>
<td>Expo-Ufot</td>
<td>1979</td>
<td>Journal of Applied Psychology</td>
<td>Nigeria</td>
</tr>
<tr>
<td>14</td>
<td>Holzbach</td>
<td>1978</td>
<td>Journal of Applied Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>15</td>
<td>Levine, Flory, &amp; Ash</td>
<td>1977</td>
<td>Journal of Applied Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>16</td>
<td>Filimowski &amp; London</td>
<td>1974</td>
<td>Journal of Applied Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>17</td>
<td>Heneman</td>
<td>1974</td>
<td>Journal of Applied Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>18</td>
<td>Williams &amp; Seker</td>
<td>1973</td>
<td>Journal of Applied Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>19</td>
<td>Atwater, Ostroff, Yammarino, &amp; Fleener</td>
<td>1998</td>
<td>Personnel Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>20</td>
<td>Farh, Dobbins &amp; Cheng</td>
<td>1991</td>
<td>Personnel Psychology</td>
<td>Republic of China</td>
</tr>
<tr>
<td>21</td>
<td>Farh, Dobbins &amp; Cheng</td>
<td>1991</td>
<td>Personnel Psychology</td>
<td>Republic of China</td>
</tr>
<tr>
<td>22</td>
<td>Farh, Dobbins &amp; Cheng</td>
<td>1991</td>
<td>Personnel Psychology</td>
<td>Republic of China</td>
</tr>
<tr>
<td>23</td>
<td>Farh, Dobbins &amp; Cheng</td>
<td>1991</td>
<td>Personnel Psychology</td>
<td>Republic of China</td>
</tr>
<tr>
<td>24</td>
<td>Farh, Dobbins &amp; Cheng</td>
<td>1991</td>
<td>Personnel Psychology</td>
<td>Republic of China</td>
</tr>
<tr>
<td>26</td>
<td>Farh, Dobbins &amp; Cheng</td>
<td>1991</td>
<td>Personnel Psychology</td>
<td>Republic of China</td>
</tr>
<tr>
<td>27</td>
<td>Farh, Dobbins &amp; Cheng</td>
<td>1991</td>
<td>Personnel Psychology</td>
<td>Republic of China</td>
</tr>
<tr>
<td>28</td>
<td>Farh, Dobbins &amp; Cheng</td>
<td>1991</td>
<td>Personnel Psychology</td>
<td>Republic of China</td>
</tr>
<tr>
<td>29</td>
<td>Fletcher, Baldry, &amp; Cunningham-Snell</td>
<td>1998</td>
<td>International Journal of Selection and Assessment</td>
<td>UK</td>
</tr>
<tr>
<td>30</td>
<td>Fox &amp; Dirar</td>
<td>1988</td>
<td>Personnel Psychology</td>
<td>Israel</td>
</tr>
<tr>
<td>31</td>
<td>Mount, Judge, Scullen, Sysma, &amp; Heutte</td>
<td>1998</td>
<td>Personnel Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>32</td>
<td>Scullen, Mount &amp; Goff</td>
<td>2000</td>
<td>Journal of Applied Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>33</td>
<td>Wohlers</td>
<td>1989</td>
<td>Personnel Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>34</td>
<td>Arnold &amp; Davey</td>
<td>1992</td>
<td>Journal of Occupational and Organizational Psychology</td>
<td>UK</td>
</tr>
<tr>
<td>35</td>
<td>Fletcher &amp; Baldry</td>
<td>2000</td>
<td>Journal of Occupational and Organizational Psychology</td>
<td>UK</td>
</tr>
<tr>
<td>36</td>
<td>Furnham &amp; Stringfield</td>
<td>1994</td>
<td>Journal of Occupational and Organizational Psychology</td>
<td>UK / Hong Kong</td>
</tr>
<tr>
<td>37</td>
<td>Motowidlo</td>
<td>1992</td>
<td>Journal of Applied Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>38</td>
<td>Lawler</td>
<td>1988</td>
<td>Journal of Applied Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>39</td>
<td>Lawler</td>
<td>1987</td>
<td>Journal of Applied Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>40</td>
<td>Baird</td>
<td>1977</td>
<td>Academy of Management Journal</td>
<td>USA</td>
</tr>
<tr>
<td>41</td>
<td>Morgan</td>
<td>1993</td>
<td>Academy of Management Journal</td>
<td>USA</td>
</tr>
<tr>
<td>42</td>
<td>Church, Rogelberg, &amp; Waclawowski</td>
<td>2000</td>
<td>Personnel Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>43</td>
<td>Ostroff, C.</td>
<td>1991</td>
<td>Personnel Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>44</td>
<td>Hoffman, Nathan, &amp; Holden</td>
<td>1991</td>
<td>Personnel Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>45</td>
<td>Yu, &amp; Murphy</td>
<td>1993</td>
<td>Personnel Psychology</td>
<td>China</td>
</tr>
<tr>
<td>46</td>
<td>Blank, Weitzel, &amp; Green</td>
<td>1990</td>
<td>Personnel Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>47</td>
<td>Steel &amp; Ovall</td>
<td>1984</td>
<td>Personnel Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>48</td>
<td>Steel &amp; Ovall</td>
<td>1984</td>
<td>Personnel Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>49</td>
<td>Steel &amp; Ovall</td>
<td>1984</td>
<td>Personnel Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>50</td>
<td>Mount</td>
<td>1984</td>
<td>Personnel Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>51</td>
<td>Waldman, Yammarino, &amp; Avolio</td>
<td>1990</td>
<td>Personnel Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>52</td>
<td>Ferris, Yates, Gilmore, &amp; Rowland</td>
<td>1985</td>
<td>Personnel Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>53</td>
<td>Cleveland &amp; Shore</td>
<td>1992</td>
<td>Journal of Applied Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>54</td>
<td>Nealey &amp; Owen</td>
<td>1970</td>
<td>Organizational Behavior and Human Decision Processes</td>
<td>USA</td>
</tr>
<tr>
<td>55</td>
<td>Thornton, G.C.</td>
<td>1968</td>
<td>Personnel Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>56</td>
<td>Van Dyne &amp; LePine</td>
<td>1998</td>
<td>Academy of Management Journal</td>
<td>USA</td>
</tr>
<tr>
<td>57</td>
<td>McFarlane Shore &amp; Thornton</td>
<td>1986</td>
<td>Academy of Management Journal</td>
<td>USA</td>
</tr>
<tr>
<td>58</td>
<td>Schmitt, Noel, &amp; Goltschank</td>
<td>1986</td>
<td>Academy of Management Journal</td>
<td>USA</td>
</tr>
<tr>
<td>59</td>
<td>Williams &amp; Levy</td>
<td>1992</td>
<td>Personnel Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>60</td>
<td>Moser, Donat, Schuler, Funke, &amp; Roloff</td>
<td>1994</td>
<td>Zeitschrift für experimentelle und angewandte Psy.</td>
<td>Germany</td>
</tr>
<tr>
<td>61</td>
<td>Prien &amp; Luise</td>
<td>1962</td>
<td>Personnel Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>62</td>
<td>Parker, Taylor, Barnett, &amp; Martens</td>
<td>1959</td>
<td>Personnel Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>63</td>
<td>Somers &amp; Birnbaum</td>
<td>1991</td>
<td>Human Relations</td>
<td>USA</td>
</tr>
<tr>
<td>64</td>
<td>Furnham &amp; Stringfield</td>
<td>1998</td>
<td>Human Relations</td>
<td>UK</td>
</tr>
<tr>
<td>65</td>
<td>Wheeler &amp; Knop (a)</td>
<td>1982</td>
<td>Journal of Educational Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>66</td>
<td>Wheeler &amp; Knop (b)</td>
<td>1982</td>
<td>Journal of Educational Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>67</td>
<td>Church, A.H.</td>
<td>1997</td>
<td>Journal of Applied Social Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>68</td>
<td>Blackburn &amp; Clark</td>
<td>1975</td>
<td>Sociology of Education</td>
<td>USA</td>
</tr>
<tr>
<td>69</td>
<td>Ketter &amp; Holland</td>
<td>1982</td>
<td>IEEE Transactions on EM</td>
<td>USA</td>
</tr>
<tr>
<td>70</td>
<td>Ketter &amp; Holland</td>
<td>1982</td>
<td>IEEE Transactions on EM</td>
<td>USA</td>
</tr>
<tr>
<td>71</td>
<td>Webb &amp; Nicon</td>
<td>1955</td>
<td>Journal of Educational Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>72</td>
<td>Kacmar, Carlson, Wright, &amp; McMahar</td>
<td>1996</td>
<td>Unpublished Document</td>
<td>USA</td>
</tr>
<tr>
<td>73</td>
<td>Kacmar, Carlson, Wright, &amp; McMahar</td>
<td>1996</td>
<td>Unpublished Document</td>
<td>USA</td>
</tr>
<tr>
<td>Number</td>
<td>Author</td>
<td>Year</td>
<td>Source</td>
<td>Country</td>
</tr>
<tr>
<td>--------</td>
<td>-------------------------</td>
<td>------</td>
<td>---------------------------------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>74</td>
<td>Shapiro &amp; Dessler</td>
<td>1985</td>
<td>Public Personnel Management</td>
<td>USA</td>
</tr>
<tr>
<td>75</td>
<td>Shapiro &amp; Dessler</td>
<td>1985</td>
<td>Public Personnel Management</td>
<td>USA</td>
</tr>
<tr>
<td>76</td>
<td>Porter &amp; Lawler</td>
<td>1998</td>
<td>Book</td>
<td>USA</td>
</tr>
<tr>
<td>77</td>
<td>Maika</td>
<td>1990</td>
<td>Evaluation and Program Planning</td>
<td>Israel</td>
</tr>
<tr>
<td>78</td>
<td>Pym &amp; Auld</td>
<td>1965</td>
<td>Journal of Occupational and Organizational Psychology</td>
<td>UK</td>
</tr>
<tr>
<td>79</td>
<td>Pym &amp; Auld</td>
<td>1965</td>
<td>Journal of Occupational and Organizational Psychology</td>
<td>UK</td>
</tr>
<tr>
<td>80</td>
<td>Pym &amp; Auld</td>
<td>1965</td>
<td>Journal of Occupational and Organizational Psychology</td>
<td>UK</td>
</tr>
<tr>
<td>81</td>
<td>Becker &amp; Vance</td>
<td>1993</td>
<td>Journal of Management</td>
<td>USA</td>
</tr>
<tr>
<td>82</td>
<td>Piercy</td>
<td>1974</td>
<td>ERIC Document Reproduction Service</td>
<td>USA</td>
</tr>
<tr>
<td>83</td>
<td>Brett &amp; Alwater</td>
<td>2001</td>
<td>Journal of Applied Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>84</td>
<td>Warr &amp; Bourne</td>
<td>1999</td>
<td>Human Performance</td>
<td>UK</td>
</tr>
<tr>
<td>85</td>
<td>Warr &amp; Bourne</td>
<td>1999</td>
<td>Human Performance</td>
<td>UK</td>
</tr>
<tr>
<td>86</td>
<td>Maurer, Mitchell, &amp; Barbeite</td>
<td>2002</td>
<td>Journal of Occupational and Organizational Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>87</td>
<td>Beehr et al.</td>
<td>2001</td>
<td>Journal of Organizational Behavior</td>
<td>USA</td>
</tr>
<tr>
<td>88</td>
<td>Van der Heijden</td>
<td>2001</td>
<td>International Journal of Selection and Assessment</td>
<td>Netherlands</td>
</tr>
<tr>
<td>89</td>
<td>Brutus, Fleenor, &amp; London</td>
<td>1998</td>
<td>Journal of Management Development</td>
<td>USA</td>
</tr>
<tr>
<td>90</td>
<td>Carless, Mann, &amp; Wearing</td>
<td>1998</td>
<td>Applied Psychology: An International Review</td>
<td>Australia</td>
</tr>
<tr>
<td>91</td>
<td>Warr &amp; Hoare</td>
<td>2002</td>
<td>International Journal of Selection and Assessment</td>
<td>UK</td>
</tr>
<tr>
<td>92</td>
<td>Atkins &amp; Wood</td>
<td>2002</td>
<td>Personnel Psychology</td>
<td>Australia</td>
</tr>
<tr>
<td>93</td>
<td>Hilario, F.L.</td>
<td>1998</td>
<td>Dissertation</td>
<td>USA</td>
</tr>
<tr>
<td>94</td>
<td>Schuler, Hell, Muck, Becker, &amp; Diemand</td>
<td>2003</td>
<td>Zeitschrift fuer Personalphysiology</td>
<td>Germany</td>
</tr>
<tr>
<td>95</td>
<td>Schuler, Hell, Muck, Becker, &amp; Diemand</td>
<td>2003</td>
<td>Zeitschrift fuer Personalphysiology</td>
<td>Germany</td>
</tr>
<tr>
<td>96</td>
<td>Kirchner, W.K.</td>
<td>1965</td>
<td>Journal of Industrial Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>97</td>
<td>Sundstrom, Town, Rice, Osborn, &amp; Bril</td>
<td>1994</td>
<td>Environment and behavior</td>
<td>USA</td>
</tr>
<tr>
<td>98</td>
<td>Sundstrom, Town, Rice, Osborn, &amp; Bril</td>
<td>1994</td>
<td>Environment and behavior</td>
<td>USA</td>
</tr>
<tr>
<td>99</td>
<td>Sundstrom, Town, Rice, Osborn, &amp; Bril</td>
<td>1994</td>
<td>Environment and behavior</td>
<td>USA</td>
</tr>
<tr>
<td>100</td>
<td>Schrader &amp; Steiner</td>
<td>1996</td>
<td>Journal of Applied Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>101</td>
<td>Williams &amp; Johnson</td>
<td>2000</td>
<td>Journal of Applied Social Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>102</td>
<td>Yammarino &amp; Waldman</td>
<td>1993</td>
<td>Journal of Applied Psychology</td>
<td>USA</td>
</tr>
<tr>
<td>103</td>
<td>Warr &amp; Bourne</td>
<td>2000</td>
<td>European Journal of Work and Organizational Psychology</td>
<td>UK</td>
</tr>
<tr>
<td>104</td>
<td>Sundvik &amp; Lindeman</td>
<td>1998</td>
<td>Journal of social behavior and personality</td>
<td>Finland</td>
</tr>
<tr>
<td>Number</td>
<td>Asian</td>
<td>Managerial</td>
<td>360_degree</td>
<td>College degrees &gt; 80%</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
<td>------------</td>
<td>------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>1</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>2</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>3</td>
<td>no</td>
<td>.</td>
<td>no</td>
<td>.</td>
</tr>
<tr>
<td>4</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>5</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>.</td>
</tr>
<tr>
<td>6</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>7</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>8</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>9</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>10</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>.</td>
</tr>
<tr>
<td>11</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>12</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>13</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>14</td>
<td>no</td>
<td>.</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>15</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>16</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>17</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>18</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>19</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>20</td>
<td>yes</td>
<td>.</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>21</td>
<td>yes</td>
<td>.</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>22</td>
<td>yes</td>
<td>.</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>23</td>
<td>yes</td>
<td>.</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>24</td>
<td>yes</td>
<td>.</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>25</td>
<td>yes</td>
<td>.</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>26</td>
<td>yes</td>
<td>.</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>27</td>
<td>yes</td>
<td>.</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>28</td>
<td>yes</td>
<td>.</td>
<td>no</td>
<td>.</td>
</tr>
<tr>
<td>29</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>30</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>31</td>
<td>no</td>
<td>.</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>32</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>33</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>34</td>
<td>no</td>
<td>.</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>35</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>36</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>37</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>38</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>39</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>40</td>
<td>no</td>
<td>.</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>41</td>
<td>no</td>
<td>.</td>
<td>yes</td>
<td>.</td>
</tr>
<tr>
<td>42</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>43</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>44</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>.</td>
</tr>
<tr>
<td>45</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>.</td>
</tr>
<tr>
<td>46</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>47</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>48</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>49</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>50</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>51</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>52</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>53</td>
<td>no</td>
<td>.</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>54</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>55</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>56</td>
<td>no</td>
<td>.</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>57</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>58</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>59</td>
<td>no</td>
<td>.</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>60</td>
<td>no</td>
<td>.</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>61</td>
<td>no</td>
<td>.</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>62</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>63</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>64</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>65</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>66</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>67</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>68</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>69</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>70</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>71</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>72</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>73</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>.</td>
</tr>
<tr>
<td>Number</td>
<td>Asian</td>
<td>Managerial</td>
<td>360_degree</td>
<td>College_degrees &gt; 80%</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
<td>-----------</td>
<td>------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>74</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>75</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>76</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>77</td>
<td>no</td>
<td>.</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>78</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>.</td>
</tr>
<tr>
<td>79</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>80</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>81</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>82</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>83</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>84</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>85</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>86</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>87</td>
<td>no</td>
<td>.</td>
<td>yes</td>
<td>.</td>
</tr>
<tr>
<td>88</td>
<td>no</td>
<td>.</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>89</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>90</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>91</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>92</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>93</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>94</td>
<td>no</td>
<td>.</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>95</td>
<td>no</td>
<td>.</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>96</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>97</td>
<td>no</td>
<td>.</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>98</td>
<td>no</td>
<td>.</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>99</td>
<td>no</td>
<td>.</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>100</td>
<td>no</td>
<td>.</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>101</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>102</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>103</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>104</td>
<td>no</td>
<td>.</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Number</td>
<td>Confidential</td>
<td>Validation</td>
<td>Purpose</td>
<td>n subjects</td>
</tr>
<tr>
<td>--------</td>
<td>--------------</td>
<td>------------</td>
<td>---------</td>
<td>------------</td>
</tr>
<tr>
<td>1</td>
<td>no</td>
<td>no</td>
<td>research</td>
<td>75</td>
</tr>
<tr>
<td>2</td>
<td>yes</td>
<td>admin.</td>
<td></td>
<td>69.5</td>
</tr>
<tr>
<td>3</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>92</td>
</tr>
<tr>
<td>4</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>74</td>
</tr>
<tr>
<td>5</td>
<td>no</td>
<td>no</td>
<td>research</td>
<td>148</td>
</tr>
<tr>
<td>6</td>
<td>no</td>
<td>no</td>
<td>research</td>
<td>40</td>
</tr>
<tr>
<td>7</td>
<td>no</td>
<td>no</td>
<td>research</td>
<td>190</td>
</tr>
<tr>
<td>8</td>
<td>no</td>
<td>no</td>
<td>develop.</td>
<td>200</td>
</tr>
<tr>
<td>9</td>
<td>no</td>
<td>no</td>
<td>research</td>
<td>89</td>
</tr>
<tr>
<td>10</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>117</td>
</tr>
<tr>
<td>11</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>443</td>
</tr>
<tr>
<td>12</td>
<td>no</td>
<td>no</td>
<td>research</td>
<td>382</td>
</tr>
<tr>
<td>13</td>
<td>no</td>
<td>no</td>
<td>research</td>
<td>88</td>
</tr>
<tr>
<td>14</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>161</td>
</tr>
<tr>
<td>15</td>
<td>yes</td>
<td>yes</td>
<td>research</td>
<td>32</td>
</tr>
<tr>
<td>16</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>153</td>
</tr>
<tr>
<td>17</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>74</td>
</tr>
<tr>
<td>18</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>202</td>
</tr>
<tr>
<td>19</td>
<td>no</td>
<td>no</td>
<td>develop.</td>
<td>1374</td>
</tr>
<tr>
<td>20</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>274</td>
</tr>
<tr>
<td>21</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>223</td>
</tr>
<tr>
<td>22</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>112</td>
</tr>
<tr>
<td>23</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>41</td>
</tr>
<tr>
<td>24</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>39</td>
</tr>
<tr>
<td>25</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>36</td>
</tr>
<tr>
<td>26</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>36</td>
</tr>
<tr>
<td>27</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>33</td>
</tr>
<tr>
<td>28</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>188</td>
</tr>
<tr>
<td>29</td>
<td>no</td>
<td>no</td>
<td>develop.</td>
<td>30</td>
</tr>
<tr>
<td>30</td>
<td>yes</td>
<td>yes</td>
<td>admin.</td>
<td>178.5</td>
</tr>
<tr>
<td>31</td>
<td>no</td>
<td>no</td>
<td>develop.</td>
<td>782</td>
</tr>
<tr>
<td>32</td>
<td>no</td>
<td>no</td>
<td>develop.</td>
<td>2142</td>
</tr>
<tr>
<td>33</td>
<td>yes</td>
<td>yes</td>
<td>develop.</td>
<td>36</td>
</tr>
<tr>
<td>34</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>729</td>
</tr>
<tr>
<td>35</td>
<td>no</td>
<td>no</td>
<td>develop.</td>
<td>45</td>
</tr>
<tr>
<td>36</td>
<td>no</td>
<td>no</td>
<td>develop.</td>
<td>316</td>
</tr>
<tr>
<td>37</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>95</td>
</tr>
<tr>
<td>38</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>55</td>
</tr>
<tr>
<td>39</td>
<td>no</td>
<td>no</td>
<td>research</td>
<td>113</td>
</tr>
<tr>
<td>40</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>165</td>
</tr>
<tr>
<td>41</td>
<td>no</td>
<td>no</td>
<td>research</td>
<td>186.5</td>
</tr>
<tr>
<td>42</td>
<td>yes</td>
<td>yes</td>
<td>develop.</td>
<td>538</td>
</tr>
<tr>
<td>43</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>67</td>
</tr>
<tr>
<td>44</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>212</td>
</tr>
<tr>
<td>45</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>367</td>
</tr>
<tr>
<td>46</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>353</td>
</tr>
<tr>
<td>47</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>398</td>
</tr>
<tr>
<td>48</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>108</td>
</tr>
<tr>
<td>49</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>142</td>
</tr>
<tr>
<td>50</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>80</td>
</tr>
<tr>
<td>51</td>
<td>no</td>
<td>no</td>
<td>develop.</td>
<td>140</td>
</tr>
<tr>
<td>52</td>
<td>no</td>
<td>no</td>
<td>research</td>
<td>81</td>
</tr>
<tr>
<td>53</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>388</td>
</tr>
<tr>
<td>54</td>
<td>no</td>
<td>no</td>
<td>research</td>
<td>25</td>
</tr>
<tr>
<td>55</td>
<td>yes</td>
<td>yes</td>
<td>develop.</td>
<td>64</td>
</tr>
<tr>
<td>56</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>524.71</td>
</tr>
<tr>
<td>57</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>70</td>
</tr>
<tr>
<td>58</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>153</td>
</tr>
<tr>
<td>59</td>
<td>no</td>
<td>no</td>
<td>research</td>
<td>73</td>
</tr>
<tr>
<td>60</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>142.6</td>
</tr>
<tr>
<td>61</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>96</td>
</tr>
<tr>
<td>62</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>117</td>
</tr>
<tr>
<td>63</td>
<td>yes</td>
<td>yes</td>
<td>admin.</td>
<td>97</td>
</tr>
<tr>
<td>64</td>
<td>no</td>
<td>no</td>
<td>research</td>
<td>47</td>
</tr>
<tr>
<td>65</td>
<td>no</td>
<td>no</td>
<td>research</td>
<td>47</td>
</tr>
<tr>
<td>66</td>
<td>yes</td>
<td>yes</td>
<td>develop.</td>
<td>152</td>
</tr>
<tr>
<td>67</td>
<td>no</td>
<td>no</td>
<td>research</td>
<td>45</td>
</tr>
<tr>
<td>68</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>256</td>
</tr>
<tr>
<td>69</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>208</td>
</tr>
<tr>
<td>70</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>51</td>
</tr>
<tr>
<td>71</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>203</td>
</tr>
<tr>
<td>72</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>233</td>
</tr>
<tr>
<td>Number</td>
<td>Confidential</td>
<td>Validation</td>
<td>Purpose</td>
<td>n subjects</td>
</tr>
<tr>
<td>--------</td>
<td>--------------</td>
<td>------------</td>
<td>---------</td>
<td>------------</td>
</tr>
<tr>
<td>74</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>86</td>
</tr>
<tr>
<td>75</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>41</td>
</tr>
<tr>
<td>76</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>635</td>
</tr>
<tr>
<td>77</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>196</td>
</tr>
<tr>
<td>78</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>85</td>
</tr>
<tr>
<td>79</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>35</td>
</tr>
<tr>
<td>80</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>89</td>
</tr>
<tr>
<td>81</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>265</td>
</tr>
<tr>
<td>82</td>
<td>no</td>
<td>no</td>
<td>research</td>
<td>27</td>
</tr>
<tr>
<td>83</td>
<td>.</td>
<td>yes</td>
<td>developm.</td>
<td>122</td>
</tr>
<tr>
<td>84</td>
<td>no</td>
<td>yes</td>
<td>research</td>
<td>362</td>
</tr>
<tr>
<td>85</td>
<td>.</td>
<td>yes</td>
<td>developm.</td>
<td>136</td>
</tr>
<tr>
<td>86</td>
<td>.</td>
<td>no</td>
<td>developm.</td>
<td>150</td>
</tr>
<tr>
<td>87</td>
<td>.</td>
<td>.</td>
<td>developm.</td>
<td>1831</td>
</tr>
<tr>
<td>88</td>
<td>no</td>
<td>no</td>
<td>research</td>
<td>115</td>
</tr>
<tr>
<td>89</td>
<td>.</td>
<td>no</td>
<td>developm.</td>
<td>1059</td>
</tr>
<tr>
<td>90</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>249</td>
</tr>
<tr>
<td>91</td>
<td>no</td>
<td>no</td>
<td>research</td>
<td>198</td>
</tr>
<tr>
<td>92</td>
<td>.</td>
<td>yes</td>
<td>developm.</td>
<td>63</td>
</tr>
<tr>
<td>93</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>23</td>
</tr>
<tr>
<td>94</td>
<td>no</td>
<td>no</td>
<td>research</td>
<td>55</td>
</tr>
<tr>
<td>95</td>
<td>no</td>
<td>no</td>
<td>research</td>
<td>240</td>
</tr>
<tr>
<td>96</td>
<td>no</td>
<td>.</td>
<td>research</td>
<td>92</td>
</tr>
<tr>
<td>97</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>514</td>
</tr>
<tr>
<td>98</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>654</td>
</tr>
<tr>
<td>99</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>180</td>
</tr>
<tr>
<td>100</td>
<td>no</td>
<td>no</td>
<td>research</td>
<td>106</td>
</tr>
<tr>
<td>101</td>
<td>yes</td>
<td>no</td>
<td>research</td>
<td>125</td>
</tr>
<tr>
<td>102</td>
<td>.</td>
<td>.</td>
<td>admin.</td>
<td>140</td>
</tr>
<tr>
<td>103</td>
<td>.</td>
<td>.</td>
<td>developm.</td>
<td>247</td>
</tr>
<tr>
<td>104</td>
<td>.</td>
<td>no</td>
<td>admin.</td>
<td>102</td>
</tr>
<tr>
<td>Number</td>
<td>Aggregated / homogeneous</td>
<td>Aggregated / heterogeneous</td>
<td>Judgmental (r)</td>
<td>Non-judgmental (r)</td>
</tr>
<tr>
<td>--------</td>
<td>--------------------------</td>
<td>---------------------------</td>
<td>----------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>0.74</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>0.39</td>
<td>0.39</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td>-0.09</td>
<td>-0.09</td>
</tr>
<tr>
<td>5</td>
<td>0.25</td>
<td></td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td>0.16</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td>0.2</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.3</td>
<td></td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td>0.23</td>
<td>0.23</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td>0.16</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td>0.47</td>
<td>0.41</td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>20</td>
<td>0.07</td>
<td></td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>0.16</td>
<td></td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>0.13</td>
<td></td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>0.08</td>
<td></td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>0.15</td>
<td></td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>0.13</td>
<td></td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>0.09</td>
<td></td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>0.15</td>
<td></td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td></td>
<td></td>
<td>0.14</td>
<td>0.32</td>
</tr>
<tr>
<td>29</td>
<td></td>
<td></td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td></td>
<td></td>
<td>0.29</td>
<td>0.25</td>
</tr>
<tr>
<td>31</td>
<td></td>
<td></td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>32</td>
<td></td>
<td></td>
<td>0.17</td>
<td>0.17</td>
</tr>
<tr>
<td>33</td>
<td>0.17</td>
<td></td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td></td>
<td></td>
<td></td>
<td>0.15</td>
</tr>
<tr>
<td>35</td>
<td></td>
<td></td>
<td></td>
<td>0.16</td>
</tr>
<tr>
<td>36</td>
<td></td>
<td></td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>37</td>
<td></td>
<td></td>
<td>0.27</td>
<td>0.27</td>
</tr>
<tr>
<td>38</td>
<td></td>
<td></td>
<td>0.09</td>
<td>0.09</td>
</tr>
<tr>
<td>39</td>
<td></td>
<td></td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td></td>
<td></td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>0.23</td>
<td></td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>0.22</td>
<td></td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td></td>
<td></td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>44</td>
<td></td>
<td></td>
<td>0.19</td>
<td>0.19</td>
</tr>
<tr>
<td>45</td>
<td></td>
<td></td>
<td>0.6</td>
<td>0.6</td>
</tr>
<tr>
<td>46</td>
<td>0.06</td>
<td></td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td></td>
<td></td>
<td>0.18</td>
<td>0.18</td>
</tr>
<tr>
<td>48</td>
<td></td>
<td></td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td></td>
<td></td>
<td>0.27</td>
<td>0.27</td>
</tr>
<tr>
<td>50</td>
<td></td>
<td></td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>51</td>
<td></td>
<td></td>
<td></td>
<td>0.02</td>
</tr>
<tr>
<td>52</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>0.22</td>
<td></td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>0.28</td>
<td></td>
<td>0.16</td>
<td>0.22</td>
</tr>
<tr>
<td>55</td>
<td></td>
<td></td>
<td>0.23</td>
<td>0.26</td>
</tr>
<tr>
<td>56</td>
<td>0.25</td>
<td></td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>57</td>
<td></td>
<td></td>
<td></td>
<td>0.29</td>
</tr>
<tr>
<td>58</td>
<td></td>
<td></td>
<td>0.31</td>
<td>0.31</td>
</tr>
<tr>
<td>59</td>
<td></td>
<td></td>
<td>0.26</td>
<td>0.32</td>
</tr>
<tr>
<td>60</td>
<td></td>
<td></td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td>61</td>
<td></td>
<td></td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>62</td>
<td></td>
<td></td>
<td>0.29</td>
<td>0.53</td>
</tr>
<tr>
<td>63</td>
<td>0.31</td>
<td></td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>64</td>
<td></td>
<td></td>
<td>0.3</td>
<td>0.13</td>
</tr>
<tr>
<td>65</td>
<td>0.14</td>
<td></td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>66</td>
<td>0.24</td>
<td></td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>67</td>
<td></td>
<td></td>
<td>0.21</td>
<td>0.21</td>
</tr>
<tr>
<td>68</td>
<td></td>
<td></td>
<td>0.13</td>
<td>0.13</td>
</tr>
<tr>
<td>69</td>
<td>0.11</td>
<td></td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>0.08</td>
<td></td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>71</td>
<td></td>
<td></td>
<td>0.16</td>
<td>0.16</td>
</tr>
<tr>
<td>72</td>
<td></td>
<td></td>
<td>0.24</td>
<td>0.24</td>
</tr>
<tr>
<td>Number</td>
<td>Aggregated / Homogeneous</td>
<td>Aggregated / Heterogeneous</td>
<td>Judgmental (r)</td>
<td>Non-judgmental (r)</td>
</tr>
<tr>
<td>--------</td>
<td>--------------------------</td>
<td>-----------------------------</td>
<td>----------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>73</td>
<td>0.26</td>
<td>.</td>
<td>0.17</td>
<td>.</td>
</tr>
<tr>
<td>74</td>
<td>.</td>
<td>.</td>
<td>0.44</td>
<td>.</td>
</tr>
<tr>
<td>75</td>
<td>.</td>
<td>.</td>
<td>0.13</td>
<td>.</td>
</tr>
<tr>
<td>76</td>
<td>.</td>
<td>.</td>
<td>0.12</td>
<td>.</td>
</tr>
<tr>
<td>77</td>
<td>0.28</td>
<td>.</td>
<td>0.28</td>
<td>.</td>
</tr>
<tr>
<td>78</td>
<td>.</td>
<td>.</td>
<td>0.52</td>
<td>.</td>
</tr>
<tr>
<td>79</td>
<td>.</td>
<td>.</td>
<td>0.46</td>
<td>.</td>
</tr>
<tr>
<td>80</td>
<td>.</td>
<td>.</td>
<td>0.59</td>
<td>.</td>
</tr>
<tr>
<td>81</td>
<td>0.22</td>
<td>.</td>
<td>0.22</td>
<td>.</td>
</tr>
<tr>
<td>82</td>
<td>0.08</td>
<td>.</td>
<td>0.08</td>
<td>.</td>
</tr>
<tr>
<td>83</td>
<td>.</td>
<td>.</td>
<td>0.13</td>
<td>.</td>
</tr>
<tr>
<td>84</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>.</td>
</tr>
<tr>
<td>85</td>
<td>0.32</td>
<td>.</td>
<td>0.32</td>
<td>.</td>
</tr>
<tr>
<td>86</td>
<td>.</td>
<td>0.11</td>
<td>0.11</td>
<td>.</td>
</tr>
<tr>
<td>87</td>
<td>0.08</td>
<td>.</td>
<td>0.08</td>
<td>.</td>
</tr>
<tr>
<td>88</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>89</td>
<td>.</td>
<td>0.11</td>
<td>0.11</td>
<td>.</td>
</tr>
<tr>
<td>90</td>
<td>0.18</td>
<td>.</td>
<td>0.18</td>
<td>.</td>
</tr>
<tr>
<td>91</td>
<td>0.28</td>
<td>.</td>
<td>0.28</td>
<td>.</td>
</tr>
<tr>
<td>92</td>
<td>.</td>
<td>0.15</td>
<td>0.15</td>
<td>.</td>
</tr>
<tr>
<td>93</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>94</td>
<td>.</td>
<td>0.3</td>
<td>0.3</td>
<td>.</td>
</tr>
<tr>
<td>95</td>
<td>.</td>
<td>0.25</td>
<td>0.25</td>
<td>.</td>
</tr>
<tr>
<td>96</td>
<td>.</td>
<td>.</td>
<td>0.23</td>
<td>.</td>
</tr>
<tr>
<td>97</td>
<td>.</td>
<td>.</td>
<td>0.73</td>
<td>0.73</td>
</tr>
<tr>
<td>98</td>
<td>.</td>
<td>0.35</td>
<td>0.35</td>
<td>.</td>
</tr>
<tr>
<td>99</td>
<td>.</td>
<td>0.15</td>
<td>0.15</td>
<td>.</td>
</tr>
<tr>
<td>100</td>
<td>.</td>
<td>0.43</td>
<td>.</td>
<td>0.43</td>
</tr>
<tr>
<td>101</td>
<td>.</td>
<td>.</td>
<td>0.05</td>
<td>.</td>
</tr>
<tr>
<td>102</td>
<td>.</td>
<td>0.03</td>
<td>0.07</td>
<td>.</td>
</tr>
<tr>
<td>103</td>
<td>0.21</td>
<td>.</td>
<td>0.21</td>
<td>.</td>
</tr>
<tr>
<td>104</td>
<td>.</td>
<td>0.17</td>
<td>0.3</td>
<td>.</td>
</tr>
<tr>
<td>Number</td>
<td>Specific (r)</td>
<td>Global (r)</td>
<td>Task-p. (r)</td>
<td>Contextual p. (r)</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td>------------</td>
<td>------------</td>
<td>------------------</td>
</tr>
<tr>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>0.74</td>
<td>0.74</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>-</td>
<td>-0.09</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>-</td>
<td>0.25</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>-</td>
<td>-</td>
<td>0.23</td>
<td>0.18</td>
</tr>
<tr>
<td>8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>9</td>
<td>-</td>
<td>-0.21</td>
<td>0.21</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>11</td>
<td>-</td>
<td>0.16</td>
<td>0.12</td>
<td>0.19</td>
</tr>
<tr>
<td>12</td>
<td>-</td>
<td>0.2</td>
<td>0.2</td>
<td>-</td>
</tr>
<tr>
<td>13</td>
<td>-</td>
<td>-</td>
<td>-0.09</td>
<td>-</td>
</tr>
<tr>
<td>14</td>
<td>0.16</td>
<td>0.16</td>
<td>0.12</td>
<td>0.19</td>
</tr>
<tr>
<td>15</td>
<td>-</td>
<td>0.2</td>
<td>0.2</td>
<td>-</td>
</tr>
<tr>
<td>16</td>
<td>-</td>
<td>-</td>
<td>0.23</td>
<td>0.18</td>
</tr>
<tr>
<td>17</td>
<td>-</td>
<td>0.26</td>
<td>0.22</td>
<td>0.21</td>
</tr>
<tr>
<td>18</td>
<td>0.48</td>
<td>0.24</td>
<td>0.6</td>
<td>0.29</td>
</tr>
<tr>
<td>19</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>20</td>
<td>-</td>
<td>-</td>
<td>0.09</td>
<td>0.01</td>
</tr>
<tr>
<td>21</td>
<td>-</td>
<td>0.19</td>
<td>0.14</td>
<td>-</td>
</tr>
<tr>
<td>22</td>
<td>-</td>
<td>0.12</td>
<td>0.17</td>
<td>-</td>
</tr>
<tr>
<td>23</td>
<td>-</td>
<td>0.1</td>
<td>0.02</td>
<td>-</td>
</tr>
<tr>
<td>24</td>
<td>-</td>
<td>0.15</td>
<td>0.17</td>
<td>-</td>
</tr>
<tr>
<td>25</td>
<td>-</td>
<td>0.15</td>
<td>0.05</td>
<td>-</td>
</tr>
<tr>
<td>26</td>
<td>-</td>
<td>0.1</td>
<td>0.03</td>
<td>-</td>
</tr>
<tr>
<td>27</td>
<td>-</td>
<td>0.16</td>
<td>0.09</td>
<td>-</td>
</tr>
<tr>
<td>28</td>
<td>0.42</td>
<td>0.32</td>
<td>0.32</td>
<td>0.14</td>
</tr>
<tr>
<td>29</td>
<td>0.18</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>30</td>
<td>-</td>
<td>0.25</td>
<td>0.45</td>
<td>0.28</td>
</tr>
<tr>
<td>31</td>
<td>-</td>
<td>0.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>32</td>
<td>-</td>
<td>0.17</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>33</td>
<td>-</td>
<td>0.15</td>
<td>0.25</td>
<td>0.11</td>
</tr>
<tr>
<td>34</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>35</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>36</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>37</td>
<td>-</td>
<td>0.27</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>38</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>39</td>
<td>-</td>
<td>0.15</td>
<td>0.07</td>
<td>0.3</td>
</tr>
<tr>
<td>40</td>
<td>0.2</td>
<td>0.16</td>
<td>0.16</td>
<td>0.13</td>
</tr>
<tr>
<td>41</td>
<td>-</td>
<td>0.27</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>42</td>
<td>-</td>
<td>0.22</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>43</td>
<td>-</td>
<td>0.15</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>44</td>
<td>-</td>
<td>0.21</td>
<td>0.17</td>
<td>-</td>
</tr>
<tr>
<td>45</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>46</td>
<td>-</td>
<td>-</td>
<td>0.06</td>
<td>-</td>
</tr>
<tr>
<td>47</td>
<td>0.19</td>
<td>0.19</td>
<td>-</td>
<td>0.15</td>
</tr>
<tr>
<td>48</td>
<td>0.31</td>
<td>0.31</td>
<td>-</td>
<td>0.27</td>
</tr>
<tr>
<td>49</td>
<td>0.3</td>
<td>0.3</td>
<td>-</td>
<td>0.17</td>
</tr>
<tr>
<td>50</td>
<td>0.15</td>
<td>0.18</td>
<td>-0.01</td>
<td>-</td>
</tr>
<tr>
<td>51</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>52</td>
<td>0.02</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>53</td>
<td>0.19</td>
<td>-</td>
<td>0.22</td>
<td>-</td>
</tr>
<tr>
<td>54</td>
<td>-</td>
<td>0.22</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>55</td>
<td>0.22</td>
<td>0.23</td>
<td>0.19</td>
<td>0.24</td>
</tr>
<tr>
<td>56</td>
<td>-</td>
<td>0.21</td>
<td>0.28</td>
<td>-</td>
</tr>
<tr>
<td>57</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>58</td>
<td>-</td>
<td>-0.31</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>59</td>
<td>0.26</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>60</td>
<td>0.42</td>
<td>0.32</td>
<td>0.36</td>
<td>0.22</td>
</tr>
<tr>
<td>61</td>
<td>0.21</td>
<td>0.24</td>
<td>0.22</td>
<td>0.32</td>
</tr>
<tr>
<td>62</td>
<td>0.35</td>
<td>0.32</td>
<td>0.34</td>
<td>0.27</td>
</tr>
<tr>
<td>63</td>
<td>-</td>
<td>0.29</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>64</td>
<td>-</td>
<td>-0.08</td>
<td>0.2</td>
<td>-</td>
</tr>
<tr>
<td>65</td>
<td>0.11</td>
<td>0.18</td>
<td>0.08</td>
<td>-</td>
</tr>
<tr>
<td>66</td>
<td>0.32</td>
<td>0.21</td>
<td>0.32</td>
<td>-</td>
</tr>
<tr>
<td>67</td>
<td>-</td>
<td>0.22</td>
<td>0.2</td>
<td>-</td>
</tr>
<tr>
<td>68</td>
<td>0.15</td>
<td>0.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>69</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>70</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>71</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>72</td>
<td>0.23</td>
<td>0.24</td>
<td>0.24</td>
<td>-</td>
</tr>
<tr>
<td>73</td>
<td>-</td>
<td>0.15</td>
<td>0.11</td>
<td>0.17</td>
</tr>
<tr>
<td>Number</td>
<td>Specific (r)</td>
<td>Global (r)</td>
<td>Task-p. (r)</td>
<td>Contextual p. (r)</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td>------------</td>
<td>-----------</td>
<td>------------------</td>
</tr>
<tr>
<td>74</td>
<td>0.44</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>75</td>
<td>0.13</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>76</td>
<td>0.03</td>
<td>0.2</td>
<td>0.2</td>
<td>-</td>
</tr>
<tr>
<td>77</td>
<td>-</td>
<td>-</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>78</td>
<td>0.52</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>79</td>
<td>0.46</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>80</td>
<td>-</td>
<td>0.53</td>
<td>0.5</td>
<td>0.57</td>
</tr>
<tr>
<td>81</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.22</td>
</tr>
<tr>
<td>82</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>83</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>84</td>
<td>-</td>
<td>-</td>
<td>0.33</td>
<td>0.31</td>
</tr>
<tr>
<td>85</td>
<td>-</td>
<td>-</td>
<td>0.3</td>
<td>0.36</td>
</tr>
<tr>
<td>86</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>87</td>
<td>-</td>
<td>-</td>
<td>0.1</td>
<td>0.07</td>
</tr>
<tr>
<td>88</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>89</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>90</td>
<td>-</td>
<td>-</td>
<td>0.21</td>
<td>-</td>
</tr>
<tr>
<td>91</td>
<td>-</td>
<td>-</td>
<td>0.28</td>
<td>0.24</td>
</tr>
<tr>
<td>92</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>93</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>94</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>95</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>96</td>
<td>0.22</td>
<td>0.2</td>
<td>0.08</td>
<td>0.26</td>
</tr>
<tr>
<td>97</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>98</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>99</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>100</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>101</td>
<td>0.05</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>102</td>
<td>-</td>
<td>0.14</td>
<td>0.07</td>
<td>0.06</td>
</tr>
<tr>
<td>103</td>
<td>-</td>
<td>-</td>
<td>0.16</td>
<td>0.26</td>
</tr>
<tr>
<td>104</td>
<td>-</td>
<td>0.36</td>
<td>0.3</td>
<td>-</td>
</tr>
<tr>
<td>Number</td>
<td>Single-item (d)</td>
<td>Aggregated / homogeneous (d)</td>
<td>Aggregated / heterogeneous (d)</td>
<td>Judgmental (d)</td>
</tr>
<tr>
<td>--------</td>
<td>----------------</td>
<td>-------------------------------</td>
<td>-------------------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>0.04</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.04</td>
<td>1.04</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>0.37</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.89</td>
<td>0.89</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-0.03</td>
<td>-0.03</td>
<td>-0.03</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.53</td>
<td>0.53</td>
<td>0.53</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.42</td>
<td>0.42</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td>0.29</td>
</tr>
<tr>
<td>14</td>
<td>0.29</td>
<td>-0.25</td>
<td>-0.25</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td>0.4</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>-0.24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td>0.38</td>
<td>0.38</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td>-0.03</td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td>-0.34</td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td>-0.27</td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td>-0.16</td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td>0.11</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td>0.14</td>
</tr>
<tr>
<td>26</td>
<td></td>
<td></td>
<td></td>
<td>-0.06</td>
</tr>
<tr>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td>-0.3</td>
</tr>
<tr>
<td>28</td>
<td>-0.56</td>
<td>0.03</td>
<td>-0.44</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td></td>
<td></td>
<td></td>
<td>-0.03</td>
</tr>
<tr>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td>1.4</td>
</tr>
<tr>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td></td>
<td></td>
<td></td>
<td>0.41</td>
</tr>
<tr>
<td>35</td>
<td></td>
<td></td>
<td></td>
<td>-0.41</td>
</tr>
<tr>
<td>36</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td></td>
<td></td>
<td></td>
<td>0.46</td>
</tr>
<tr>
<td>38</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td></td>
<td></td>
<td></td>
<td>0.47</td>
</tr>
<tr>
<td>42</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>43</td>
<td></td>
<td></td>
<td></td>
<td>1.18</td>
</tr>
<tr>
<td>44</td>
<td></td>
<td></td>
<td></td>
<td>1.07</td>
</tr>
<tr>
<td>45</td>
<td></td>
<td></td>
<td></td>
<td>0.3</td>
</tr>
<tr>
<td>46</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>47</td>
<td></td>
<td></td>
<td></td>
<td>1.22</td>
</tr>
<tr>
<td>48</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>49</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td>0.12</td>
</tr>
<tr>
<td>51</td>
<td></td>
<td></td>
<td></td>
<td>0.23</td>
</tr>
<tr>
<td>52</td>
<td></td>
<td></td>
<td></td>
<td>-0.06</td>
</tr>
<tr>
<td>53</td>
<td></td>
<td></td>
<td></td>
<td>0.55</td>
</tr>
<tr>
<td>54</td>
<td></td>
<td></td>
<td></td>
<td>-0.45</td>
</tr>
<tr>
<td>55</td>
<td></td>
<td></td>
<td></td>
<td>0.15</td>
</tr>
<tr>
<td>56</td>
<td></td>
<td></td>
<td></td>
<td>0.71</td>
</tr>
<tr>
<td>57</td>
<td></td>
<td></td>
<td></td>
<td>0.47</td>
</tr>
<tr>
<td>58</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>59</td>
<td></td>
<td></td>
<td></td>
<td>0.37</td>
</tr>
<tr>
<td>60</td>
<td>-0.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>61</td>
<td>0.63</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>0.69</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>63</td>
<td></td>
<td></td>
<td></td>
<td>0.26</td>
</tr>
<tr>
<td>64</td>
<td></td>
<td></td>
<td></td>
<td>0.28</td>
</tr>
<tr>
<td>65</td>
<td>1.06</td>
<td></td>
<td></td>
<td>0.92</td>
</tr>
<tr>
<td>66</td>
<td>1.25</td>
<td></td>
<td></td>
<td>1.09</td>
</tr>
<tr>
<td>67</td>
<td></td>
<td></td>
<td></td>
<td>0.28</td>
</tr>
<tr>
<td>68</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>69</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>71</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>72</td>
<td>0.84</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number</td>
<td>Single-item (d)</td>
<td>Aggregated / homogeneous (d)</td>
<td>Aggregated / heterogeneous (d)</td>
<td>Judgmental (d)</td>
</tr>
<tr>
<td>--------</td>
<td>----------------</td>
<td>-------------------------------</td>
<td>-------------------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>73</td>
<td>0.07</td>
<td>0.11</td>
<td>-</td>
<td>0.07</td>
</tr>
<tr>
<td>74</td>
<td>1.07</td>
<td>-</td>
<td>1.07</td>
<td>-</td>
</tr>
<tr>
<td>75</td>
<td>0.24</td>
<td>-</td>
<td>0.24</td>
<td>-</td>
</tr>
<tr>
<td>76</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>77</td>
<td>1.06</td>
<td>-</td>
<td>1.06</td>
<td>-</td>
</tr>
<tr>
<td>78</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>79</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>80</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>81</td>
<td></td>
<td>-0.03</td>
<td>-0.03</td>
<td>-</td>
</tr>
<tr>
<td>82</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>83</td>
<td></td>
<td>-</td>
<td>-0.25</td>
<td>-</td>
</tr>
<tr>
<td>84</td>
<td></td>
<td>0.28</td>
<td>0.28</td>
<td>-</td>
</tr>
<tr>
<td>85</td>
<td></td>
<td>0.28</td>
<td>-</td>
<td>0.28</td>
</tr>
<tr>
<td>86</td>
<td></td>
<td>0.08</td>
<td>0.08</td>
<td>-</td>
</tr>
<tr>
<td>87</td>
<td></td>
<td>0.52</td>
<td>-</td>
<td>0.52</td>
</tr>
<tr>
<td>88</td>
<td></td>
<td>0.16</td>
<td>-</td>
<td>0.16</td>
</tr>
<tr>
<td>89</td>
<td></td>
<td>-</td>
<td>0.68</td>
<td>-</td>
</tr>
<tr>
<td>90</td>
<td></td>
<td>0.8</td>
<td>-</td>
<td>0.8</td>
</tr>
<tr>
<td>91</td>
<td></td>
<td>0.75</td>
<td>-</td>
<td>0.75</td>
</tr>
<tr>
<td>92</td>
<td></td>
<td>-</td>
<td>0.57</td>
<td>-</td>
</tr>
<tr>
<td>93</td>
<td></td>
<td>-</td>
<td>-0.74</td>
<td>-</td>
</tr>
<tr>
<td>94</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>95</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>96</td>
<td></td>
<td>0.13</td>
<td>-</td>
<td>0.13</td>
</tr>
<tr>
<td>97</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>98</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>99</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>100</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>101</td>
<td></td>
<td>0.3</td>
<td>-</td>
<td>0.3</td>
</tr>
<tr>
<td>102</td>
<td></td>
<td>0.07</td>
<td>0.32</td>
<td>0.23</td>
</tr>
<tr>
<td>103</td>
<td></td>
<td>0.04</td>
<td>-</td>
<td>0.04</td>
</tr>
<tr>
<td>104</td>
<td></td>
<td>-0.02</td>
<td>0.39</td>
<td>0.11</td>
</tr>
<tr>
<td>Number</td>
<td>Behavioral (d)</td>
<td>Broad (d)</td>
<td>Specific (d)</td>
<td>Global (d)</td>
</tr>
<tr>
<td>--------</td>
<td>----------------</td>
<td>----------</td>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td>1</td>
<td>0.04</td>
<td></td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0.04</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-0.01</td>
<td>-0.06</td>
<td>0.07</td>
<td>-0.07</td>
</tr>
<tr>
<td>6</td>
<td>0.42</td>
<td></td>
<td>0.42</td>
<td>0.29</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.24</td>
<td></td>
<td>0.24</td>
<td>0.24</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0.29</td>
<td>0.3</td>
<td>0.28</td>
<td>0.24</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>-0.24</td>
<td>-0.18</td>
<td>-0.25</td>
<td>-0.27</td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td>-0.04</td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td>-0.36</td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td>-0.24</td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td>-0.19</td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td>-0.29</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td>0.15</td>
</tr>
<tr>
<td>26</td>
<td></td>
<td></td>
<td></td>
<td>-0.06</td>
</tr>
<tr>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td>-0.34</td>
</tr>
<tr>
<td>28</td>
<td>-0.44</td>
<td>-0.67</td>
<td>-0.44</td>
<td>-0.44</td>
</tr>
<tr>
<td>29</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>43</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>44</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>1.22</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td>48</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>49</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.12</td>
<td>0.08</td>
<td>0.05</td>
<td>0.26</td>
</tr>
<tr>
<td>51</td>
<td></td>
<td></td>
<td>0.23</td>
<td>0.24</td>
</tr>
<tr>
<td>52</td>
<td></td>
<td></td>
<td>-0.06</td>
<td></td>
</tr>
<tr>
<td>53</td>
<td></td>
<td></td>
<td>0.55</td>
<td>-0.45</td>
</tr>
<tr>
<td>54</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>0.13</td>
<td>0.17</td>
<td>0.14</td>
<td>0.15</td>
</tr>
<tr>
<td>56</td>
<td></td>
<td></td>
<td>0.99</td>
<td>0.51</td>
</tr>
<tr>
<td>57</td>
<td>0.47</td>
<td></td>
<td>0.47</td>
<td>0.46</td>
</tr>
<tr>
<td>58</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>0.37</td>
<td>0.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>-0.18</td>
<td>0.08</td>
<td>-0.1</td>
<td>-0.1</td>
</tr>
<tr>
<td>61</td>
<td></td>
<td>0.56</td>
<td>1.12</td>
<td>0.5</td>
</tr>
<tr>
<td>62</td>
<td>0.74</td>
<td>0.32</td>
<td>0.72</td>
<td>0.71</td>
</tr>
<tr>
<td>63</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>64</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>1.06</td>
<td></td>
<td>1.08</td>
<td>0.72</td>
</tr>
<tr>
<td>66</td>
<td></td>
<td>1.25</td>
<td>1.07</td>
<td>0.96</td>
</tr>
<tr>
<td>67</td>
<td></td>
<td></td>
<td>0.31</td>
<td>0.24</td>
</tr>
<tr>
<td>68</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>69</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>71</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>72</td>
<td>0.77</td>
<td>0.88</td>
<td>0.77</td>
<td>0.77</td>
</tr>
<tr>
<td>73</td>
<td></td>
<td>0.07</td>
<td>0.09</td>
<td>0.11</td>
</tr>
<tr>
<td>74</td>
<td></td>
<td>1.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td></td>
<td>0.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number</td>
<td>Behavioral (d)</td>
<td>Broad (d)</td>
<td>Specific (d)</td>
<td>Global (d)</td>
</tr>
<tr>
<td>--------</td>
<td>----------------</td>
<td>----------</td>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td>76</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>77</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>78</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>79</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>80</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>81</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>82</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>83</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>84</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>85</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>86</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>87</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>88</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>89</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>90</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>91</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>92</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>93</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>94</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>95</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>96</td>
<td>0.13</td>
<td>.</td>
<td>0.13</td>
<td>0.13</td>
</tr>
<tr>
<td>97</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>98</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>99</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>100</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>101</td>
<td>.</td>
<td>.</td>
<td>0.3</td>
<td>.</td>
</tr>
<tr>
<td>102</td>
<td>.</td>
<td>.</td>
<td>0.07</td>
<td>.</td>
</tr>
<tr>
<td>103</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>104</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>
Appendix
A Deutschsprachige Zusammenfassung

Zusammenfassung


A.1 Einleitung


### A.2 Ziel der Arbeit

Ziel der Arbeit ist es, die Konvergenz von Selbst- und Vorgesetztenbeurteilungen zu untersuchen. Dabei liegt das Hauptinteresse darin, Kontextfaktoren von Leistungsbeurteilungen zu identifizieren, die einen moderierenden Einfluss auf die Höhe der Urteilsübereinstimmung ausüben. Drei Gruppen solcher Faktoren werden unterschieden: (a) Eigenheiten des Instruments und der Skalen, die zur Leistungsbeurteilung eingesetzt werden, (b) situative Rahmenbedingungen, wie etwa der Zweck der Leistungsbeurteilung, und (c) Eigenschaften der untersuchten Personen und ihrer beruflichen Position.

Durch die Technik der Sekundäranalyse werden primäre Forschungsergebnisse integriert und die empirische Evidenz zusammengetragen, die für den Einfluss verschiedener Kontextfaktoren auf die Übereinstimmung von Selbst- und Vorgesetztenbeurteilungen besteht. So ergibt sich eine Übersicht über den Forschungsstand. Die vorgestellte Metaanalyse richtet sich sowohl an Forscher, die sich für Selbsteinschätzungen im Kontext beruflicher Leistung oder für die
Übereinstimmung von Urteilen auf Ratingskalen generell interessieren, als auch an Praktiker, die einen Überblick über relevante Merkmale des Kontextes von Leistungsbeurteilungen gewinnen möchten, welche die Ergebnisse von subjektiven Urteilen auf Ratingskalen beeinflussen - insbesondere, wenn Selbsturteile beteiligt sind.


A.3 Hypothesen


A.4 Methode

Eine Literatursuche nach veröffentlichten und unveröffentlichten primären Forschungsarbeiten wurde durch Abfrage elektronischer Datenbanken (PsycLit, PsychInfo, ERIC, Dissertation Ab-
Stracks Database – UMI ProQuest), das Durchsehen der Literaturverzeichnisse frühere Metaanalysen und anderer Forschungsarbeiten aus dem Themenbereich, sowie die manuelle Durchsicht der wichtigsten relevanten Zeitschriften durchgeführt.

Vier Kriterien wurden angewendet um über die Aufnahme von Forschungsstudien in die Sekundäranalyse zu entscheiden: (1) die Studien berichten ein quantitatives Maß der Urteilerübereinstimmung – entweder Korrelationskoeffizienten oder Mittelwertsdifferenzen der beiden Urteilerperspektiven "Selbst" und "Vorgesetzter", wobei es sich (2) um Urteile auf Ratingskalen handelt und (3) die Ergebnisse durch Feldforschung in (4) realen Arbeitsumgebungen gewonnen wurden.


### A.5 Ergebnisse

82 Forschungsarbeiten konnten gefunden werden, die aus dem Zeitraum von 1955 bis 2003 stammten und Ergebnisse von 104 unabhängigen Stichproben enthielten. Insgesamt 96 Stichproben berichteten Korrelationskoeffizienten (417 Koeffizienten, 22287 untersuchte Personen), 70 Mittelwertsdifferenzen von Selbst- und Vorgesetztenurteilen (324 Koeffizienten, 29386 Personen). Die meisten der Studien wurden mit Stichproben aus den USA durchgeführt (70), während 34 der Studien aus anderen Ländern stammten.

Auf der Grundlage von 96 Stichproben ergab sich eine meta-analytische Schätzung der korrelativen Übereinstimmung von Selbst- und Vorgesetztenbeurteilungen von \( r = .22 \) (\( r = .36 \) nach Artefaktkorrektur). Interpretiert man diese Korrelation als ein Maß der kriteriumsbezogenen Validität, so deutet sich an, dass diese für Selbsturteile eher niedrig ausfällt. Die Höhe der Urteile unterschied sich substantiell zwischen Selbst- und Vorgesetztenurteilen (\( d = .33 / .41 \)) bei 70 zugrundeliegenden Stichproben. Insgesamt bestätigte sich damit die Annahme, dass Selbsturteile eine Tendenz zur Milde zeigen.

Die beiden Maße der Urteilerübereinstimmung (Index \( r \) und \( d \)) wurden in unterschiedlicher Weise durch moderierende Variable beeinflusst. Drei der untersuchten Variablen erwiesen sich

A.6 Integration der Ergebnisse in einem Modell der Urteiler-Konvergenz

B  LEBENSLAUF