Within- and Between-group Agreement in Supervisor's Evaluative Behaviours: Do evaluative 'styles' exist?

Jan Noeverman

ERIM REPORT SERIES *RESEARCH IN MANAGEMENT*	
ERIM Report Series reference number	ERS-2010-002-F&A
Publication	January 2010
Number of pages	38
Persistent paper URL	http://hdl.handle.net/1765/17700
Email address corresponding author	noeverman@ese.eur.nl
Address	Erasmus Research Institute of Management (ERIM) RSM Erasmus University / Erasmus School of Economics Erasmus Universiteit Rotterdam P.O.Box 1738 3000 DR Rotterdam, The Netherlands Phone: + 31 10 408 1182 Fax: + 31 10 408 9640 Email: info@erim.eur.nl Internet: www.erim.eur.nl

Bibliographic data and classifications of all the ERIM reports are also available on the ERIM website: www.erim.eur.nl

ERASMUS RESEARCH INSTITUTE OF MANAGEMENT

REPORT SERIES

RESEARCH IN MANAGEMENT

ABSTRACT AND KEYWORDS	
Abstract	Several management accounting studies have investigated the behavioural impact of evaluative style, a concept that generally refers to the manner in which supervisors use accounting information to evaluate the performance of subordinates. Although most studies study this behavioural impact at the individual level of the subordinate, the term "evaluative style" suggests that evaluative behaviours and attitudes of single supervisors will show (some) consistency across subordinates. This paper investigates whether "evaluative styles" exist by examining within-group and between-group agreement in evaluative behaviours by subordinates reporting to the same supervisor. The findings from two empirical studies indicate that evaluative behaviours in both organisations show both within-group and between-group variability. These findings suggest that evaluative behaviours of supervisors are more appropriately analysed at the level of individual subordinates than at the level of groups, although a dyadic level of analysis should be considered as well. An implication of these findings is that the concept of "evaluative style" is misleading. A suggestion is made to use the term "evaluatorship" instead as an umbrella concept to refer to evaluative behaviours and attitudes of supervisors at different levels of analysis in future research.
Free Keywords	performance evaluation, evaluative style, within-group agreement, WABA, levels of analysis
Availability	The ERIM Report Series is distributed through the following platforms: Academic Repository at Erasmus University (DEAR), DEAR ERIM Series Portal Social Science Research Network (SSRN), SSRN ERIM Series Webpage Research Papers in Economics (REPEC), REPEC ERIM Series Webpage
Classifications	The electronic versions of the papers in the ERIM report Series contain bibliographic metadata by the following classification systems: Library of Congress Classification, (LCC) LCC Webpage Journal of Economic Literature, (JEL), JEL Webpage ACM Computing Classification System CCS Webpage Inspec Classification scheme (ICS), ICS Webpage

Within- and between-group agreement in supervisor's evaluative behaviours: do evaluative 'styles' exist?

Jan Noeverman*
Department of Business Economics

*Erasmus School of Economics
Erasmus University Rotterdam
Room W H 14-28
P.O. Box 1738
3000 DR Rotterdam
The Netherlands
Tel +31 104081317
Fax +31 104089171
Email: noeverman@ese.eur.nl

Within- and between-group agreement in supervisor's evaluative behaviours: do evaluative 'styles' exist?

Abstract

Several management accounting studies have investigated the behavioural impact of evaluative style, a concept that generally refers to the manner in which supervisors use accounting information to evaluate the performance of subordinates. Although most studies study this behavioural impact at the individual level of the subordinate, the term "evaluative style" suggests that evaluative behaviours and attitudes of single supervisors will show (some) consistency across subordinates. This paper investigates whether "evaluative styles" exist by examining within-group and between-group agreement in evaluative behaviours by subordinates reporting to the same supervisor. The findings from two empirical studies indicate that evaluative behaviours in both organisations show both within-group and between-group variability. These findings suggest that evaluative behaviours of supervisors are more appropriately analysed at the level of individual subordinates than at the level of groups, although a dyadic level of analysis should be considered as well. An implication of these findings is that the concept of "evaluative style" is misleading. A suggestion is made to use the term "evaluatorship" instead as an umbrella concept to refer to evaluative behaviours and attitudes of supervisors at different levels of analysis in future research.

KEYWORDS: performance evaluation, evaluative style, within-group agreement, WABA, levels of analysis

1 Introduction

In the management accounting and control literature research level-of-analysis issues have hardly been addressed. This is surprising given the attention that is given to these issues in fields like leadership and organizational behaviour over the past two decades. In this paper, I argue that the following statement of Schriesheim at al. (2001, p. 516) made in the context of leadership also applies to management accounting and control research in general:

> "We believe that it is absolutely critical that scholars specify the level of analysis at which their hypotheses, frameworks, models, and/or theories hold so that they may be adequately tested. We also believe that it is absolutely necessary that tests of any hypothesis, framework, model, and/or theory be conducted at the proper level(s) of analysis and that tests explicitly rule out inappropriate or competing (rival) levels of analysis."

One particular area of the management accounting and control literature to which this statement applies is research on the role of supervisors in performance evaluation and its effect on subordinate managers. An important part of the accounting literature around this topic has investigated the behavioural impact of evaluative style, or performance evaluation style, a construct generally referring to (perceived) differences in the manner in which managers use accounting information, particularly budgetary information, to evaluate the performance of their subordinate managers. Although these studies have been classified under the general heading of RAPM (reliance on accounting performance measures), suggesting the existence of a coherent research stream, recently, it has been demonstrated that the RAPM-literature contains a wide variety of concepts and measures of evaluative style without appropriate consideration of validity and reliability issues (Otley & Fakiolas, 2000; Vagneur & Peiperl, 2000; Hartmann, 2000; Noeverman, Koene, & Williams, 2005). This is only one of the numerous problems of the RAPM-literature that have been spelled out in reviews of the work conducted in this area (see Briers & Hirst, 1990; Hartmann, 2000; Hartmann & Moers, 1999; 2003; Dunk, 2003; Otley & Pollanen, 2000; Otley & Fakiolas, 2000; Vagneur & Peiperl, 2000), and several recommendations have been made to restore the flaws in existing RAPM-studies. However, in my opinion, even the latest criticisms and recommendations have not sufficiently dealt with an issue that is fundamental for understanding the role of supervisors in performance evaluation and its behavioural and motivational effects on subordinates. This is the issue whether "evaluative styles" do exist, and if so, at what level of analysis.

Most RAPM-studies have used subordinates' perceptions of their supervisor's evaluative style as a measure of evaluative style, and many studies have used data collected from subordinates within one single organisation, often with multiple subordinates from one supervisor. By analysing such data at the individual subordinate level-of-analysis, these analyses assume that "evaluative style" and its effects on attitudinal outcomes such as job related tension and job satisfaction are independent of group effects, thus that the relationships found hold both within-groups and between-groups. This implicit level-of-analysis assumption has never been tested, while progress in research methods in leadership and organisational research over the past few decades allow testing this assumption.

Thus, the purpose of this paper is to address level of analysis issues around the concept of "evaluative style" by exploring and testing within- and between-group agreement in evaluative behaviours of supervisors within the context of a single organisation. Specifically, this study will answer the following three research questions: Do differences

in perceived evaluative behaviours across subordinates reporting to different supervisors occur at the level of individual subordinates or do they reflect differences at a higher level, i.e. between-group differences? Do subordinates who are evaluated by the same supervisor report differences in perceived evaluative behaviours (i.e., within-group differences), or do they report similar evaluative behaviours (i.e., within-group-agreement)? And if there are within-group differences in perceived evaluative behaviours, do these reflect differences that are intended by the supervisor? To answer these questions, I conducted two exploratory studies, each using a different research methodology. The first study was a small qualitative study, based on interview data obtained from two supervisors and three subordinates reporting to each of these supervisors. The second study was a quantitative analysis of within- and between-group differences in the perceived quality of performance feedback from their supervisor as reported by 51 subordinates reporting to eight different supervisors. The results from both studies indicated that both within- and between-group differences existed. These findings indicate that the group level-of-analysis seems inappropriate, and support an individual subordinate level of analysis. However, there is also evidence that dyads rather than groups or individual subordinates could be the appropriate level of analysis. Implications of these findings for future research in management accounting are discussed.

The remainder of this paper is organized as follows. The second section provides an overview of the literature on levels of analysis terminology, on the role of leaders in performance evaluation, and on evaluative style. This overview results in four hypotheses. The third and fourth section describe the two small empirical studies. Section 3 reports the results from the qualitative study; section 4 reports the results from the quantitative analysis of data on feedback quality. Both sections are closed with a brief discussion of the most important findings from each study. Finally, in section 5, implications for future research are discussed and an overall conclusion is drawn.

2. Background and hypothesis development

2.1 Levels of theory, measurement and data analysis

Although explicit consideration of levels-of-analysis issues in management accounting literature is rare, these issues have received increasing attention in leadership research since the seminal work of Dansereau, Alutto, & Yammarino (1984).

Dansereau, Alutto, & Yammarino (1984) developed a conceptual framework and empirical approach to address levels of analysis. They distinguish between entities, level of analysis and unit of analysis. Entities are the specific objects of interest to a researcher. Levels of analysis refer to categories or classifications of entities arranged in a hierarchical order

such that higher levels include lower levels of analysis. And the unit of analysis refers to the characterisation of a single entity in terms of parts or wholes at a specific level of analysis. I will clarify these terms using insights from leadership research.

In a recent review of levels of analysis in leadership research, Yammarino et al.(2005) distinguish four different levels of analysis that are relevant for studying leadership. The first level is the level of individuals or persons, independent of one another. The second level is the level of dyads, defined as two interdependent (on a one-to-one basis) individuals. The third level is the level of groups or teams, usually understood as a collection of individuals who are interdependent and interact on a face-to-face or virtual basis with one another. The fourth level is the level of collectives, which are clusters of individuals, larger than groups, who are interdependent based on some form of clustering, whether hierarchical (such as managerial level, functional areas, departments, or business units) or based on common/shared expectations.

Within each of these levels of analysis it is possible to focus on wholes or on parts as the unit of analysis. For example, depending on the theory used to explain leadership, at the individual level of the supervisor the supervisor (the entity) can be viewed as a whole person or the focus may be on parts within the supervisor. In earlier research on leadership, researchers –often implicitly- assumed that each supervisor had a certain "leadership style", treating each subordinate reporting to him in exactly the same way. In this view, called the Average Leadership Style (ALS) approach, the supervisor was viewed as a whole person with regard to his style of leadership (Dansereau, Alutto, & Yammarino, 1984). In contrast, in the Vertical Dyad Linkage (VDL) approach (Graen & Schiemann, 1978), supervisors are expected to vary their leadership behaviours across subordinates, depending on the relative position of one subordinate to other subordinates. Thus, the leadership behaviours towards one subordinate are dependent on a comparison between subordinates that occurs within a supervisor. In this view, a supervisor is viewed in terms of parts rather than wholes (Dansereau, Alutto, & Yammarino., 1984). The choice which of the two approaches is most appropriate in a specific study should be informed by theory, but also depends on the leadership dimensions of interest. This is commonly referred to as the level of theory, which refers to making explicit what the target level of analysis is (e.g., individual, dyad, group, organization) that the researcher aims to depict and explain (Klein, Dansereau, & Hall, 1994; Luft & Shields, 2003). Besides the level of theory, two other terms are also important, which are the level of measurement and the level of data analysis. The level of measurement refers to the level at which data is collected or obtained. For example, we may ask the supervisor to fill out one questionnaire with one set of questions about his leadership behaviours. This self-reported data is collected at the

level of the supervisor as a whole. We may also ask a supervisor to fill out several questionnaires about his leadership behaviours, each for one specific subordinate. This self-reported data is then collected at the level of supervisor parts. Finally, the level of statistical analysis refers to the level at which the data (measured at a certain level) is actually analysed. For example, data on leadership behaviours of supervisors obtained from subordinates may be aggregated and the group mean may be used as an indication of "Average Leadership Style" (ALS). Although the data is collected at the level of individual subordinates, by aggregating the data it is analysed at the group level.

2.2 Guidelines to address levels-of-analysis issues

In their review of the management accounting literature, Luft & Shields (2003) asked researchers to pay attention to levels of analysis issues in management accounting research. They presented a number of guidelines to address levels issues (pp. 199-200):

1. Indicate whether the variable of interest varies across individuals, organizational subunits, organizations, or beyond-organization entities like markets and societies.
2. Align the level of theory (what is being explained), level of variable measurement (source of evidence), and level of data analysis (unit of data).
3. If theoretical variables at multiple levels affect the observable measures, then separate the effects from multiple levels.
4. If cross-level effects are proposed, then use an interaction causal-model form, with at least one interacting (independent or moderator) variable at the level of the dependent variable.
5. If the variation of interest in a variable is variation in its value relative to a subset of other values in the sample, then use an individual-within-group-level model.

Although these guidelines may indeed be helpful in addressing levels issues, unfortunately, they do not provide much guidance on how to determine the appropriate level of theory or the expected level of variance in the construct of interest (guideline 1). This however is a very important step that needs to be taken before the other guidelines can be taken into account. This step requires a consideration of the theoretical and methodological issues involved in determining the appropriate level(s) of analysis of a construct. Klein, Dansereau, & Hall (1994, p.196) presented a theory-based framework based on the assertion that "precise articulation of the level of one's constructs is an important priority for all organizational scholars whether they propose single- or mixed-level theories". Thus, the main focus of their framework is on the level of theory. Assuming the level of measurement is individual subordinates as members of groups within a single organisation, the researcher has three basic alternatives as to the appropriate level of analysis. The researcher may predict that the responses from individuals are homogeneous, independent

or heterogeneous responses with regard to the construct of interest. When homogeneity is specified, the researcher predicts that the construct of interest does not vary between individual group members. Thus, within-group agreement or low variance in responses within the group is expected. Any within-group variance is considered to be (measurement) error, so the group is treated as a whole (cf. Dansereau, Alutto, & Yammarino, 1984). In contrast, the researcher may also specify that individual responses are independent of group membership. In this view group membership is irrelevant. Thus, the appropriate level of analysis is the individual subordinate, while the group level of analysis is inappropriate. Variability in the constructs of interest is pure between-individual variance, and within- and between-group variances are irrelevant because groups in itself are irrelevant. The third alternative is to specify heterogeneity or view the individual responses as parts within the group. In this view, the researcher predicts that subordinates are compared or ranked within the group, focusing on the relative position of individual responses as compared to the group mean. In this view, individual responses are expected to vary within groups, but the groups are still relevant and even necessary to interpret individual responses on the construct of interest. So, the focus is on within-group variance (deviances from the average group score). The Vertical Dyad Linkage (VDL) approach to which I referred earlier provides a good example of a leadership construct in which heterogeneity within groups will be specified.

2.3 Role of leaders in performance evaluation

In accounting research, performance measurement and performance evaluation are popular topics. Most accounting research focuses on the design of effective performance measurement systems, or on the incentive effects of performance measures. Only a small part of the accounting literature has investigated the role of supervisors in performance evaluation. This is surprising, given the fact that in many organisations the formal performance evaluation system allows supervisors (some) discretion in how they apply elements of the system when evaluating the performance of subordinates. Such discretion introduces subjectivity in the performance evaluation of subordinates. This subjectivity may take different forms. For example, in the context of bonus determination, Gibbs, Merchant, & Vargus (2004) stated that subjectivity may arise from (1) all or part of a bonus is based on subjective judgments about performance; (2) the weights on some or all quantitative measures are determined subjectively, or (3) a subjective performance threshold or "override" is used, in which case a subjective determination as to whether or not to pay a bonus is made based on measured performance and other factors. These three forms of subjectivity can also be distinguished in the context of performance

evaluation in general. Depending on how much discretion the supervisor has given the formal design of the performance evaluation system and the wider organisational context in which performance evaluation takes place, subjectivity in a particular organisation may be limited to just one of these three possible forms, or any combination of these forms. Whatever the exact form in which subjectivity occurs within an organisation, if supervisors have discretion in determining performance ratings and bonuses, it is interesting to investigate whether supervisors within a particular organisation show variability in the level of subjectivity used to evaluate the performance of subordinates (for example, one supervisor uses more subjective performance criteria than another supervisor). Additionally, even more interesting is the question whether there is variability (heterogeneity) or lack thereof (homogeneity) in the level of subjectivity used by a single supervisor in the performance evaluation of each of his subordinates.

2.4 RAPM and 'evaluative style'

Even the small part of the management accounting literature that focuses on the role of supervisors in performance evaluation lacks explicit consideration of such issues, albeit necessary to interpret and understand the findings of these studies and their practical relevance. This is particularly true for, although certainly not limited to, the part of management accounting research that has become known as the RAPM-literature. This literature contains studies that focused on "evaluative style", referring to the manner in which a supervisor uses accounting information or accounting performance measures for the purpose of evaluating his subordinates. Several reviews of the RAPM-literature have appeared over the last decade, indicating the problematic state of this literature. One of these problems is the lack of a uniform definition and conceptualisation of evaluative style (Otley & Fakiolas, 2000; Otley & Pollanen, 2000; Merchant & Otley, 2007), and the lack of reliability, validity and comparability of the measures that were used to capture different evaluative styles (Vagneur & Peiperl, 2000; Noeverman, Koene, & Williams, 2005). Since the focus of my study is on levels-of-analysis issues, I will not discuss these problems in RAPM in detail here, nor do I want to discuss the different concepts and measures of evaluative style. I merely refer to the RAPM literature here for two reasons. First, despite its problematic state, the RAPM literature is a good example of management accounting research in which consideration of levels-of-analysis issues is essential, given the focus on "evaluative style" as the focal construct of interest. Second, this literature has evolved from two early studies that focused explicitly on the role of supervisors in performance evaluation, that of Hopwood (1972; 1973; 1974) and Otley (1978). In these two studies, levels-of-analysis issues were raised which have been completely ignored in later research,

but are still relevant for contemporary research on supervisors' discretionary choices and behaviours in managerial performance evaluation.

Hopwood (1972; 1973; 1974)

Hopwood (1972; 1973; 1974) studied the plurality of ways in which managers within one large Chicago-based company used accounting information for managerial performance evaluation. Hopwood distinguished three "styles of evaluation" within this organisation: a Budget-Constrained style, a Profit-Conscious style, and a Non-accounting style. His sample consisted of 167 cost centre heads that were evaluated by 26 departmental supervisors. These departmental supervisors, in turn, reported to a number of area managers. Cost centre heads were asked to indicate the style of evaluation of their departmental supervisor, while the departmental supervisors were asked to indicate the style of evaluation of their area manager. In addition, the departmental supervisors were also asked to report on their own intended style of evaluation.

To test for contagion effects, an effect referring to the tendency of managers to evaluate their subordinates in the same way as they themselves perceive to be evaluated, Hopwood aggregated the individual responses of cost centre heads to obtain a measure of the dominant style of evaluation of the departmental supervisor. The departmental supervisor's dominant style was categorised as one of the three styles when more than 40% of the cost centre heads reported that they perceived that particular style.

Interestingly, Hopwood (1974) reported that in 19 of the 27 departments, there was considerable agreement between subordinates in their perception of the departmental supervisor's style. In 8 departments, however, there was considerable disagreement. Hopwood's analysis of these 8 departments provided some evidence that the divisional supervisors emphasised the Budget Constrained style especially with cost centres which were likely to have the greatest impact on the departmental results, i.e. those cost centres that most likely affected his own performance the most. These were the cost centre heads that were less successful in meeting the budget and were heads of relatively large cost centres within the department. Thus, Hopwood (1974, p. 491) concluded, "managers do not necessarily use one personal style of management irrespective of their circumstances and subordinates".

Finally, Hopwood (1974) also found that there was considerable disagreement between the perceived dominant style of the divisional managers and their intended style of evaluation. Thus, self-reported styles of evaluation are not necessarily equivalent with the subordinates' perceived styles of evaluation. However, it should be noted that each of the divisional managers was asked to report one single intended style of evaluation, while

Hopwood's (1974) analysis of the data provided evidence that some departmental supervisors tried to adapt their styles to the circumstances and subordinates, as stated above.

Otley (1978)

Otley (1978) replicated Hopwood's (1972) study in a single organisation in which the use of budgetary accounting information was well suited, in contrast to Hopwood's setting in which accounting information had severe shortcomings. Otley's sample consisted of 41 unit managers from three geographically different groups headed by three different group managers. Otley explicitly stated that the unit of analysis of his study was the individual unit manager who reported to his group manager. Accordingly, Otley (1978, p. 127) defines the style of evaluation as "the way in which a unit manager perceives the budget to be used in evaluating his performance, in relation to other relevant information." Based on the responses of the unit managers, Otley distinguished five evaluative styles. But even though his main focus was on the unit managers' perceptions, Otley also addressed the question whether group managers choose their evaluative style, and whether they differentiate between units and unit managers. His results (pp.140-141) indicated that

> 'Despite considerable differences in perceived evaluative style perceived by managers in each group, there were significant differences between groups… These perceived differences correspond very closely to the nuances of style that senior group managers intend to transmit to their subordinates.'
>
> "It is evident that unit managers were perceiving differences in perceived evaluative style that were intended to be transmitted by group managers. However, it is also clear that there was a significant degree of bias in either group managers' projection of their intended style or of unit managers' perceptions of it."

Otley (1978) provided some evidence that the intended style of evaluation was affected by three factors: the personal philosophy and personality of the group manager, environmental and economic circumstances at the group level, and the relative size of operating units within the group. While the first two factors help explain differences between intended styles of evaluation at the level of the group managers, the third factor suggests that there is not one intended style of evaluation, but that group managers may choose to differentiate their intended style of evaluation. Both Hopwood (1974) and Otley (1978) found that such variation in intended style may be attributable in part to the relative impact of a single unit to the overall group performance on which the group managers are evaluated themselves.

2.5 Development of hypotheses

Based on the findings from Hopwood (1974) and Otley (1978) I draw the following conclusions. First, a distinction should be made between **intended evaluative behaviours** and **perceived evaluative behaviours**. Intended evaluative behaviours reflect the general manner in which supervisors say they are evaluating their subordinates, while perceived evaluative behaviours reflect the manner in which individual subordinates say their manager is evaluating their performance. The perceived evaluative behaviours may or may not reflect the intended evaluative behaviours. Second, between-group differences in perceived evaluative behaviours are likely to exist, a conclusion that is in line with Hopwood's (1974) findings. Furthermore, such between-group differences seem to correspond with differences between supervisors in intended style of evaluation. Apparently, different supervisors do evaluate their subordinates in different ways, which partly explains between-group differences in perceived evaluative behaviours. Third, although between-group differences exist, there are considerable within-group differences in perceived evaluative behaviours as well. These differences may occur because a supervisor may intentionally evaluate different subordinates in different ways and/or because individuals may perceive the intended style differently (i.e., measurement error). Fourth, Hopwood's finding that in 8 of the 27 departments there was (far) less agreement on perceived style of evaluation among the cost centre heads indicates that evaluators may differ in the extent to which they differentiate their intended styles of evaluation. These conclusions, primarily based on empirical evidence, are also supported by theoretical arguments from a study by Schriesheim, Cogliser, & Neider (1995) on the level of analysis of the Ohio State leadership model. This leadership model distinguishes two dimensions of leadership, consideration and initiating structure, which have been used in several accounting studies (DeCoster & Fertakis, 1968; Hopwood, 1973; 1974; Pratt & Jiambalvo, 1981; 1982; Jiambalvo & Pratt, 1982; Brownell, 1983; Otley & Pierce, 1995). In their study, Schriesheim, Cogliser, & Neider (1995, pp. 117-118) argue that leadership could be treated as both a between-groups and a within-group phenomenon. They argue that the between-groups view is supported by notions of consistencies in behaviour across persons and situations in many areas of psychology, especially in the fields of personality and clinical psychology. Furthermore, they argue, that

> "Equity theory (e.g., Adams, 1963) would predict at least some general behavioural consistencies from supervisors, arising from such causes as the need to maintain equity among work unit members and limitations in the resources (e.g., time, energy, etc.) available to supervisors for them to tailor different behaviours toward different subordinates." (p. 117)

But there is also much evidence and theory that supports viewing leadership as a dyadic or within-groups phenomenon:

> "For example, expectancy theory (e.g., Vroom, 1964) would predict that leaders will behave differently toward subordinates in their work unit, depending upon how instrumental each is to furthering his or her desired goals or outcomes." (p. 118)

These same theoretical arguments can also be applied to evaluative behaviours of supervisors. Thus, there is both empirical and theoretical support for the view that evaluative behaviours of supervisors are likely to contain some consistencies across subordinates (i.e., constitute "styles"), while at the same time subordinates from the same group or work unit may be treated differently. Thus, I expect both between-groups and within-group differences in evaluative behaviours.
This leads to the following hypotheses that will be explored empirically in this study:

Hypothesis 1. There are significant between-group differences in perceived evaluative behaviours of the supervisor (where groups are subordinates reporting to the same supervisor).

Hypothesis 2. There are significant differences in perceived evaluative behaviours of the supervisor within groups of subordinates reporting to the same supervisor.

It should be noted however that these theoretical arguments are based on the implicit assumption that the supervisor actually has some discretion in determining how to evaluate the performance of subordinates. But given the existence of this discretion, it is likely that supervisors will use this discretion in different ways. Each supervisor, for example, has to make a decision on the resources he or she is willing to spend on performance evaluation, which will possibly also affect the opportunity to behave differently towards subordinates. Thus, in choosing how to evaluate the performance of individual subordinates, supervisors may implicitly or explicitly trade-off costs and benefits of alternative evaluative behaviours. Therefore, it is likely that there will occur differences in the level of within-group variability between supervisors, and, consequently, between groups of subordinates.

This leads to the following hypotheses that will be explored empirically in this study:

Hypothesis 3. There will be significant differences in the level of within-group agreement between groups of subordinates.

Hypothesis 4. Within- and between-group differences in perceived evaluative behaviours are related to differences in intended evaluative behaviours.

3. Empirical study 1

3.1 Research setting and methodology

The first empirical study took place within the business unit Letters of TPG Post, the Dutch Royal mail company. This research setting is the same setting as the study conducted by Wiersma (2008). The organisation was chosen because it was a large branch-type organisation, which made it relatively easy to interview subordinate managers at the same level in the organisation and with similar job responsibilities, i.e. branch managers, who report to and are evaluated by the same supervisor, i.e. an area manager. I was able to get access to the organisation through the Head of Management Development.
Data were collected through different sources. I conducted two interviews with the Head of Management Development and collected a number of documents that described the rules and procedures of the performance evaluation system. The Head of Management Development gave us the names of two area managers that I could talk to. Both immediately agreed to participate in the study.

In one area (area X) I interviewed the area manager and three branch managers (X1, X2 and X3). In the other area (area Y) I interviewed the area manager (Y), two branch managers (Y1 and Y2) and the manager of Customer Service (Y3). Although the area manager evaluates the latter's performance too, it is important to realize that manager Y3 was in a staff function, not in a line management function. His position was not directly comparable to the position of the branch managers. The Customer Service manager replaced the Area Manager in case of absence. Furthermore, the Customer Service department provided monthly and quarterly management information on the performance of the area as a whole, and all branches within the area. The Service Manager was interviewed to get an independent view on how the area manager evaluated the branch managers. The branch managers were the lowest management level (of four levels) in the organisation, supervising employees in non-managerial positions.
All interviews were held in the summer of 1997, a period that falls within the same time period as that on which Wiersma's (2008) data analysis was based. Although the evaluative behaviours described in the interviews are certainly not representative for the current evaluation practice within the organisation, for the purpose of this paper, i.e. to test within- and between-group agreement, I believe the interviews still provide valuable information.

An interview protocol was developed to guide the interviews and to collect data in a semi-structured format. After some introductory questions, the interview protocol focused on

the evaluation criteria that were used in the organisation to evaluate performance in general, which of these performance criteria were (perceived to be) emphasised the most by the area manager when evaluating the performance of branch managers, and the manner in which supervisors used these evaluation criteria. The interviews were not tape-recorded. Notes made during the interviews were written out in as much detail as possible directly after the interview. The interview-transcriptions were sent back to interviewees to check the validity of statements.

3.2 Results

Introduction

The interview data on how each of the two area managers intends to evaluate and is perceived to evaluate the performance of the branch managers is summarized in Table 1 for Area A and table 2 for Area Y. <INSERT TABLE 1 HERE> <INSERT TABLE 2 HERE> The first column in both tables summarises the answers given by the area manager, and thus refers to intended evaluative behaviours. The other three columns in each table summarises the answers given by the branch managers, referring to perceived evaluative behaviours. The last column in table 2 summarises the answers of the Service Centre Manager in Area Y on the perceived evaluative behaviours of the area manager in evaluating the branch managers in general.

Differences and similarities in evaluative behaviours between areas

As shown in table 1 and 2, all respondents indicated that the performance evaluation of branch managers is primarily based on the extent to which branch managers meet the targets that are set as part of the management contract. Costs and quality are the two most important performance areas on which targets are included in the contract. Quality refers mainly to delivering mail in time and correctly, and is measured by on-time delivery of mail. The emphasis on quality in evaluating the performance of branch managers is in line with the quality-oriented strategy (Wiersma, 2008) pursued by the company, while at the same time there was a high emphasis on controlling costs, especially labour costs, because of the labour intensive processes. However, the tables reveal some interesting differences in how the performance of branch managers is evaluated between area X and area Y. First, as can be seen in table 1, in area X, the area manager and the three branch managers all indicated that besides costs and quality customer satisfaction and employee satisfaction (or worker satisfaction as Wiersma (2008) calls it) were part of the management contract of branch managers. Furthermore, both the area manager himself

and the three branch managers stressed that the area manager attached importance to these targets, although the primary emphasis remained on cost efficiency, which was given a weight of 50%, and on-time delivery (a weight of approximately 35%). The interviews revealed that the inclusion of customer satisfaction and employee satisfaction measures in the performance evaluation of branch managers reflected a recent organizational-wide change of emphasis. Both aspects had been measured for only two years, i.e. since 1995, as part of the stronger emphasis on quality in addition to cost efficiency. In contrast, in area Y, as shown in table 2, all four respondents only referred to cost and quality targets as part of the management contract. The interviews revealed that the area manager did consider customer satisfaction and employee motivation in the performance evaluation of branch managers, but merely as an indication of whether the branch managers were 'people managers'. The service manager also said that employee satisfaction was part of the performance evaluation, although this was a more subjective part in his opinion than cost efficiency (productivity) and quality targets. Yet the two branch managers did not mention either customer satisfaction or employee satisfaction in the interviews. These results suggest that, in 1997, there were important differences in how the recent emphasis of the organization on a broad concept of quality, including internal processes, was dispersed and emphasised in different areas across the organization. A second difference in performance evaluation that appears from the interview data is the emphasis on results (pre-set targets) versus on the processes by which these results were obtained (managerial behaviour). Area manager Y seems to adhere more strictly to pre-set targets than area manager X does. Although branch managers Y1 and Y2 may be able to explain why they did not achieve their targets, this does not change the fact that area manager Y still thinks that they should have attained the targets: managers have either met agreements or they have not; explanations can not change that. Yet, when managers provide a reasonable explanation this does influence what the consequences of not meeting the agreements will be for the overall evaluation. One of the branch managers in area Y described the manner in which his performance would be evaluated as follows:

> "Initially, the conversation will focus on the items in the contract, but he will ask for the story behind it. Often he does already know the story anyway. The items in the contract are rigorous, and will be checked item by item. All items will be considered. And when I do not meet certain items, then at first it is not good. And even if I have a good story, it remains wrong. But it does not mean that the evaluation will be negative. If you are able to show what you did, then it will be taken into account in the evaluation".

While area manager X also stresses that the agreements in the contract have to be met, the interviews suggest that area manager X puts a lot more emphasis on the process by which branch managers achieve the results than area manager Y, also in the daily practice between formal evaluation moments. The importance of how a branch manager operates is an ever-recurring theme during the interview with area manager X. First, when describing his own function, the area manager says: "I regard it as my most important responsibility to determine what has to be realised, but above all how it has to be realised." Second, concerning his own evaluation, the Area Manager says he slightly disagrees with how his boss evaluates him, because the evaluation is limited to the items in the contract, while he feels dedication is more important. "In evaluating my performance, my boss runs through almost every item on which an agreement has been made. I do not feel that's necessary." Third, the area manager has even introduced a special term (a verb) in his area to describe how he expects a (branch) manager to behave: "a manager has to ZIP". "ZIP" is an abbreviation of the Dutch words "Zelfstandig", "Initiatief nemend", and "Probleemoplossend", which can be translated as Autonomous, Taking initiative, and Problem solving (ATP). I will use the abbreviation of the English equivalents, ATP, instead of the Dutch abbreviation ZIP in the remainder of this paper. ATP is explicitly included in the contract as part of the agreements. A branch manager's contract in area X therefore contains agreements on expected outcomes and agreements on how the manager should achieve these outcomes. Both types of agreement are part of branch managers' performance evaluation. The area manager stated:

> "A branch manager who does not show ATP puts problems at my desk that do not belong there. A branch manager will come across that because I will point it out to him. He will also come across it at the end of the year during the performance evaluation meeting, because ATP is a part of the contract that we both signed at the beginning of the year... When I believe a branch manager does not show ATP, his evaluation will be somewhat negative. If the branch manager disagrees, he has to illustrate that he did show ATP. For example, if he claims to have shown problem-solving behaviour, then he should name, for example, five problems he solved in the past year."

The purpose of introducing ATP, however, is not to provide branch managers with an excuse for not reaching expected outcomes or to decrease the importance of meeting the targets. Quite the contrary, as the following quote from the area manager illustrates.

"The contract has to be observed. All agreements in the contract are realistic. They are obtainable with reasonable effort. However, in the past you could always explain why you did not attain these targets. People began to focus on finding good excuses to allow them not to attain their targets. I was very annoyed by that. I thought: "I work hard, but I do not attain my targets. What are my managers doing to attain them?" It had to change. The story that a branch manager puts forward for not attaining his targets has to be realistic. There have to be good reasons and you must be able to show what you have done to solve problems, or to attain certain targets. If a manager can show that he really managed everything well, deviations from the contract are allowed. Initially it is necessary to head for the targets; there should not be any discussion about these targets."

Differences in evaluative behaviours within areas (between branch managers)

The interviews with the branch managers in area X for the greater part supported the description of the (intended) evaluative behaviours that area manager X provided. According to two of the branch managers, although the area manager attaches importance to outcomes, he has defined another important component of performance, which is how you attain these outcomes. As one branch manager said:

"The area manager calls it -and it becomes increasingly clear- showing ATP, he calls it moneymaking, and he calls it centralising the customer process. As to moneymaking, the area manager says: "you have a contract, but actually, you should be able to achieve even more than what's in it." Thus, you get rid off: "now I have to spend my whole budget, otherwise I will have less to spend next year." ATP stands for Autonomous, Taking initiative, and Problem solving, so "don't bother me with your problems". As to customer process: do you have the customer between the ears and can you bring it between the ears of your own subordinates".

Similarly, another branch manager said:

"Our contract includes not just targets, but also agreements on the manner of functioning: a number of behavioural characteristics that fit the management style that the area manager propagates. Then you talk about a number of concepts. What is essential to him is so-called ATP. That is autonomous, taking initiative, and problem solving... We have a culture -from the past, and a little exaggerated, and black and white- that we do what we are appointed to, and we try to do that well. However, do not ask me too much, too much own initiative, too many problems to solve, because than it becomes difficult. That applies especially to the shop floor. Therefore, the aim is that we start to do things, that we signalise things, that we embrace opportunities, and that we do not talk in terms of

problems, but in terms of solutions. It is not interesting whether something is a problem, but it is interesting whether you can settle it, and solve it."

The descriptions that branch managers X1 and X2 gave of area manager X's evaluative behaviours seem to be consistent with each other, as well as with the description given by area manager X himself. However, the description of branch manager X3 shows some differences. According to branch manager X3, behaviour and effort were not part of performance evaluation. This branch manager stated that performance evaluation focuses on figures only; performance evaluation is too straightforward. Yet, this seems to contradict the impression the other three interviews provide: effort seems to be important; explanations for not attaining targets seem to be possible, lest they are realistic. All three branch managers referred to ATP, which is clearly referring to effort and behaviour, several times during the interviews; however, branch manager X3 did not mention it once. A number of reasons may explain this apparent contradiction.

1. The interviews revealed that area manager X had only been employed in area X for a couple of months when I conducted the interviews. Thus, area manager X had not formally evaluated the branch managers yet. With the former area manager, as all three branch managers in area X emphasised in the interviews, performance was evaluated very rigidly against pre-set targets. Possibly, branch manager X3's perception of his current area manager's evaluative behaviours may be coloured by past experiences. Although this was not apparent from the interview with branch manager X3 himself, a comment from one of his colleagues, branch manager X1, concerning the visit he got from area manager X supports this explanation:

   "As part of the control cycle, we have progresses conversations, and two weeks ago the area manager was here. He had said: "I will come to have a progresses conversation". With the other boss it was always like this: prepare, and demonstrate to decimal places. So, thus I did. He immediately said: "Hey, it goes really well. I do not need to hear that at all. I would really like to know: how do you manage employee motivation?" So I sat here stuttering. I actually knew that he would do it like that, but you are still used to the cycle, in this case progresses conversation."

2. Branch manager X3 experienced trouble in meeting this year's agreement. In particular, the branch manager explained that he was using more labour than he was supposed to. He knew there was only a slight chance that he would meet his agreement on this item. He was sure that he would get a negative evaluation at the end of the year, and felt threatened by it. In contrast, branch managers X1 and X2

were quite sure they were doing a good job. The current performance may thus influence the perception of evaluative behaviours. Possibly, managers who experience trouble in achieving their targets may perceive their supervisor's evaluative behaviours in a different way than managers who know they are performing well. In this particular case, this effect may even be stronger because of past experiences (see the first reason above).

3. While the two former reasons assume the difference in description occurred because of a difference in the perception of the performance evaluation, it is possible that this difference in perception reflected an actual difference in evaluative behaviours. In other words, area manager X may have differentiated his evaluative behaviours with branch manager X3 than with the other two branch managers. In this particular case, it is possible that area manager X did stress the importance of meeting the targets precisely because branch manager X3 had trouble meeting them. If the area manager at this stage would accept that branch manager X3 will not meet his targets, branch manager X3 may not do the best he can to meet them. Since both branch managers X1 and X2 were likely to meet their targets anyway, there was no need to stress the importance of these targets. With these two managers the emphasis could be on what they did to attain these targets; these managers had to show that they attained their targets because of the way they managed, and not just because they got lucky.

In area Y, the descriptions that branch managers Y1 and Y2 gave of area manager Y's evaluative style seem to be consistent with one another. Furthermore, these descriptions also appeared to be reasonably similar to the Although there were some small differences with the descriptions given by the service centre manager and the area manager himself, although, as explained above, there were some small differences regarding the importance of customer satisfaction and employee satisfaction. All three managers evaluated by area manager Y seemed to agree with the way in which they were evaluated. In contrast to the managers in area X, all managers interviewed in area Y had been employed in their current function for the past three years at least. The area manager selects the branch managers working under his supervision himself. The area manager explicitly stated that if managers received a formal evaluation mark "passable", they would have to do better in the following year. If they fail to do better, they have to find themselves another job. However, in contrast to the ATP notion in area X, the area manager Y does not provide clear guidelines on how performance can be improved.

3.3 Conclusion empirical study 1

Although the qualitative nature of this study does not allow testing the statistical significance of within-group and between-group differences, overall, the results do support the hypotheses. As to hypothesis 1, the results indicate that there were similarities, but also differences in the perceived evaluative behaviours of their area manager between branch managers in area X and branch managers in area Y. Furthermore, in area X, the branch managers also experienced significant differences in evaluative behaviours of the former and the current area manager. These results support hypothesis 1, indicating that subordinate manager's perception of how their performance is evaluated depends on the supervisor who evaluates their performance. There also appeared to be differences within groups, although these differences were apparent in area X, but not in area Y. This partly supports hypothesis 2, and fully supports hypothesis 3. These results are interesting, as combined they suggest that the level of within-group variability (or reversed, the level of within-group agreement) may not be homogeneous across groups in the same organisation, but may differ between groups. Yet, in the current study, the difference could also be due to methodological artifacts, as the respondents from area Y had worked together with their area manager for some years as compared to just a couple of months in area X. This difference in tenure with the supervisor may also reflect differences in the level of trust in the supervisor and/or in the quality of the supervisor-subordinate relationship. Leader-member-exchange theory (LMX) (Dansereau, Graen, & Haga, 1975; Graen & Schiemann, 1978), for example, suggests that leaders will differentiate leadership behaviours towards subordinates within the work-group, based on the quality of the supervisor-subordinate relationship (Graen & Uhl-Bien, 1995; Scandura & Lankau, 1996; Schriesheim, Castro, & Cogliser, 1999). Thus, in area Y the quality of the supervisor-subordinate relationship may be high, while in area X this relationship may be low as there has hardly been time to develop this relationship.

Finally, the results also seem to support hypothesis 4, indicating that there appears to be congruence in perceived and intended evaluative style. However, it should be noted that in the interviews with the area managers, area managers were asked to indicate how they in general evaluated the performance of branch managers. A different approach could have been to take a more dyadic approach, asking the area manager how he evaluates one or more specific branch managers. Such an approach would not only allow a distinction between intended differentiation in evaluative behaviours and perceived differentiation in evaluative behaviours, but also matching intended and perceived evaluative style for each superior-subordinate relationship.

4. Empirical study 2

4.1 Research setting and methodology

Sample

As a complement to the qualitative approach of the first study relying on interview data, I conducted a second study using survey data with the purpose to statistically test for within- and between-group differences in evaluative behaviours. Since the RAPM literature lacks a reliable measure of evaluative behaviour that has external validity (Vagneur & Peiperl, 2000; Noeverman, Koene, & Williams, 2005), in this study I focused on feedback quality, an aspect of supervisors' behaviour in the context of performance evaluation that recently has received attention in the management (Steelman, Levy, & Snell, 2004; Steelman & Rutkowski, 2004) and accounting literature (Hartmann & Slapnicar, 2009). Data was collected on the quality of feedback provided by supervisors from 60 employees employed in a Big Four Accountancy firm located in a large city in the Netherlands. The data reported here was collected in the summer of 2009 as part of a broader study (reported elsewhere). The data was collected using convenience sampling rather than random sampling, as I did not get permission to use a written questionnaire to collect data from all employees. A student working on his master thesis under the supervision of the author was allowed, however, to collect data from employees who were present at the office on three separate visits to the office. Practically all employees who were approached by the student agreed to fill out and hand in the questionnaire immediately. This data collection process did not allow testing for the possibility of non-response bias. Even though some bias may exist due to the sampling method, there is no reason to suspect that such bias would be different across groups of subordinates reporting to one supervisor. Therefore, I believe the potential bias due to the sampling method does not cause problems in interpreting the results on within- and between-group variance. To be considered a group, at least 3 employees should report on the feedback quality of the same supervisor. This definition of a group implied that nine employees, reporting to five different supervisors, had to be removed from the sample, leaving a final sample of 51 employees reporting to eight different supervisors. These eight groups varied in size from three employees to ten employees.

Variable measurement

Feedback quality was measured in this study using the instrument from Steelman, Levy, & Snell (2004). This instrument has been used in various management studies before, but also in a recent accounting study by Hartmann & Slapnicar (2009). The instrument contains four items, asking respondents to indicate the extent to which feedback provided by their

supervisor was consistent and useful using a 5-point Likert scale. This measure was chosen, because it seems to have acceptable psychometric properties and equal interval measurement, which are characteristics required for using r_{WG} (James, Demaree, & Wolf, 1984; see also Castro, 2002) and WABA (Castro, 2002). Furthermore, in previous studies (Steelman, Levy, & Snell, 2004; Whitaker, 2007) the scale has been shown to have high internal validity, i.e. Cronbach's alpha > .87. In my study, the mean score on feedback quality was 3.73 and the standard deviation was .51, with scores ranging from 2.5 to 4.5. The internal reliability coefficient (Cronbach's alpha) was .69, which is satisfactory, but lower than expected based on previous studies.

Data analysis technique

Because only data from subordinates' was collected, hypothesis 4 was not addressed in this second study. To test hypotheses 1 and 2, I used within- and between-entity analysis (WABA) (Dansereau, Alutto, & Yammarino., 1984; Yammarino & Markham, 1992). WABA is a test to assess statistically and practically whether variance in the data is primarily due to within-group differences (group parts level of analysis), between-group differences (whole group level of analysis), or both (equivocal condition, group level of analysis is inappropriate). However, WABA only allows testing within- and between-group variance for the sample as a whole. I was also interested in possible differences between groups of subordinates (hypothesis 3). To test whether such differences occurred, I computed the within-group agreement (r_{WG}) coefficient (James, Demaree, & Wolf, 1984; 1993; see Lebreton & Senter, 2008 for details on this coefficient). Since the r_{WG} coefficient is calculated for each group separately, in contrast to WABA, this coefficient allows me to compare within-group agreement or lack thereof across groups.

One of the most complicated factors in using r_{WG} -based indices is choosing the null distribution (LeBreton & Senter, 2008). The null distribution is the theoretical distribution used to assess the expected variance in item-scores or scale-scores when there is a total lack of agreement, thus assuming that all respondents (raters) respond randomly (James, Demaree, & Wolf, 1984; LeBreton & Senter, 2008). Although typically a uniform distribution is used as a null distribution to estimate the expected variance, LeBreton & Senter (2008) emphasise, following James, Demaree, & Wolf (1984) and Schriesheim et al. (2001), that the null distribution needs to be justified theoretically. Since a uniform distribution assumes the absence of any cognitive and affective biases in responding to questionnaire items, such a distribution, given the overwhelming evidence of the existence of such biases, is almost certainly biased in itself, inflating r_{WG}. Therefore, rather than simply relying on the assumption of a uniform null distribution, the expected

variance to be used with the r_{WG} index was based on previous empirical findings (cf. Schriesheim, Cogliser, & Neider, 1995) as summarised in table 3. <INSERT TABLE 3 HERE> Table 3 shows that only two studies (Steelman, Levy, & Snell, 2004; Whitaker, 2007) report the mean and standard deviation of feedback quality. The reported standard deviation and corresponding variance in these two studies are much smaller than the variance expected under a uniform distribution. Instead, the variance reported in both studies suggest that a normal distribution is more likely, although the reported mean is slightly higher than the mid-scale point, suggesting a slight bias towards the higher end of the scale (negative skew). Therefore, using the table with variance estimates under different distributions as provided by LeBreton & Senter (2008), I will show r_{WG} values under the usual assumption of a uniform distribution (with expected variance of 2.00), under the assumption of a slight negatively skewed null distribution (with expected variance of 1.34), and under the more realistic assumption of a normal distribution (with expected variance of 1.04).

4.2 Results

<INSERT TABLE 4 HERE>

WABA I

The right –hand side of table 4 reports the outcomes of the WABA approach. As pointed out by Schriesheim, Neider, & Scandura (1998, p. 308), the decision rules to draw inferences from WABA are outlined in detail in Dansereau, Alutto, & Yammarino, 1984, pp.169-185 and Yammarino & Markham, 1992, pp. 171-172. Following these decision rules, first the practical significance of within- and between-group variance was assessed, based on the E-test. This test indicates that within-group is significantly greater than the between-group variance using the 15^{o} level of practical significance, but not when using the more conservative 30^{o} level. Because the E-test indicates that within-group variance is larger than the between-group variance, a corrected F-test should be used (Dansereau, Alutto, & Yammarino, 1984). The corrected F-test is calculated as the inverse of the F-ratio with N-J degrees of freedom for the numerator and J-1 degrees of freedom for the denominator (Dansereau, Alutto, & Yammarino, 1984; Yammarino & Markham, 1992; Schriesheim, Cogliser, & Neider, 1995), where N is the number of subordinates and J is the number of groups. So, in the test, critical values for the F-ratio with 43 and 7 degrees of freedom of approximately 3.33 and 5.88 were obtained from a table at the .05 and .01 levels of significance respectively. Taking the inverse, this leads to critical values of the

corrected F-test of 0.300 ($p < .05$) and 0.170 ($p < .01$). Thus, the F-ratio obtained in the sample of .404 is not statistically significant at the .05 level of significance. Overall, these results should be interpreted as equivocal at the group level, with a weak tendency towards a parts condition because of the practical significance at 15°. The equivocal condition at the group level supports the view that a group level of analysis is not appropriate, and suggests that the appropriate level of analysis could be that of the individual subordinate. Overall, the results from WABA I confirm hypothesis 1 and 2.

r_{WG}

Table 4 reports the r_{WG} values for each of the eight managers on feedback quality. To interpret r_{WG}, as a rule of thumb, George (1990) suggested that values above .70 are necessary to demonstrate high within-group agreement. This criterion is not without debate in organisational research (LeBreton, James, & Lindell, 2005; LeBreton & Senter, 2008), and recently several other criteria have been developed to interpret r_{WG}. LeBreton & Senter (2008), for example, suggest the following guidelines: r_{WG} values from 0 to .30 indicate lack of agreement, from .31 to .50 weak agreement, from .51 to .70 moderate agreement, from .71 to .90 strong agreement, and values above .90 show very strong agreement. Following these guidelines, table 4 shows that under the assumption of a uniform null distribution there is strong within-group agreement in two groups (group 4 and 5), and very strong agreement in the other six groups. Using a slightly skewed null distribution, in five groups the within-group agreement remains very strong, while in three groups (groups 4, 5 and 8) the within-group agreement is strong. Finally, under the more realistic assumption of a null distribution with expected variance of 1.04 only in two groups (group 1 and 2) the within-group agreement is very strong, in four groups (group 3, 6, 7, and 8) the within-group agreement is strong, and in two groups (groups 4 and 5) within-group agreement is moderate. These results illustrate the importance of defining a realistic null distribution in calculating and interpreting the within-group coefficient, as there appear to be substantial differences in values of r_{WG} under different null distributions. Yet, the results also show that only the absolute value of the within-group coefficient changes as a result of specifying different null distributions. The ranking of within-group agreement coefficients within the sample does not change, and, except under the assumption of a uniform null distribution, the results suggest that there seem to be considerable differences in r_{WG} -coefficients across groups. In particular, overall we may conclude that groups 4 and 5 have relatively low r_{WG} coefficients, while group 1, 2, 3 and 7 have relatively high coefficients, with group 6 and 8 somewhere in between.

It should be noted that instead of the guidelines used here, more advanced tests have been developed to evaluate the statistical significance of r_{WG} (Charnes & Schriesheim, 1995; Dunlap, Burke, & Smith-Crowe, 2003; Cohen, Doveh, & Nahum-Shani, 2009) or to estimate the expected variance if each individual responded at random creating pseudo-groups (Bliese & Halverson, 2002; Lüdtke & Robitzsch, 2009). However, to test hypothesis 3, I was not interested in evaluating the (significance of) individual r_{WG} values per se, but rather in assessing whether the apparent differences in r_{WG} values between groups are statistically significant. For this purpose, I used a test recently developed by Pasisz & Hurtz (2009). This test is based on O'Brien's (1979; 1981) method to test variances between groups which is robust for non-normality and unbalanced design. In this method, the original raw data are transformed so that the transformed values equal the group variances of the raw data (O'Brien, 1981). These transformed data are then used in a mixed-model ANOVA where the supervisors are entered as a between-group variable and the transformed ratings of feedback quality are entered as a within-group variable (see Pasisz & Hurtz, 2009 for more details about the procedure). To conduct this test, I used the SPSS syntax provided by Pasisz & Hurtz (2009) in the appendix of their paper. The test revealed that there was no significant interaction effect between supervisor and the transformed ratings, indicating that there were no significant differences in the pattern of agreement across supervisors across the four items of feedback quality, $F(21, 129)=.657$, $p=.398$. Since there is no significant interaction between groups and ratings, the between-subject effects can be interpreted unambiguously (i.e., does not vary across the four items comprising the scale), but was found to be not significant, $F(7,43)=1.098$, $p=.382$. Thus, no significant differences in the average variances across groups were found, and hypothesis 3 is rejected.

4.3 Conclusion empirical study 2

The findings of the second study indicate that feedback quality varies both within- and between-groups, but no significant differences are found in within-group variability between groups. These findings support the view that feedback quality, as perceived by subordinates, is an individual level variable, independent of groups. These results should, first of all, be viewed in light of the small sample size. WABA I also provided practical significance tests, which are independent of sample size. The practical significance E-test indicated a tendency towards a parts within groups level of analysis, which could not be confirmed with an F-test of statistical significance. The small sample size makes it impossible to detect statistically significant differences in between-group variances as compared to within-group variances unless such differences are large. A second point of

concern when interpreting these findings is that the combination of WABA I and r_{WG} is not without problems. WABA I assumes homogeneity of variance, while the test to determine significant differences in r_{WG} coefficients is in fact a test of homogeneity of variance (Pasisz & Hurtz, 2009). Thus, if significant differences were found in r_{WG} coefficients, this would have casted doubt on the appropriateness of the WABA I analysis. Although I did not find significant differences in r_{WG} coefficients, it is important to note that testing hypotheses 1 and 2 thus assumed that hypothesis 3 should be rejected. If, however, hypothesis 3 cannot be rejected, it is important to develop theory to explain the differences in r_{WG} coefficients between the original groups. Based on such theory, different homogeneous (sub)groups within the sample should then be identified that allow re-testing hypotheses 1 and 2 using WABA. In this way, WABA I and r_{WG} can be used as complementary tests, also in future studies.

5 Discussion and conclusion

Although limited in scope, both empirical studies reported in the previous sections provide evidence that significant within- and between-group differences in evaluative behaviours of supervisors exist within single organisations. However, the findings with regard to differences in within-group variability are mixed. In the first study notable differences occurred between the two groups of subordinates, but in the second study no statistically significant differences in within-group variability were found. Finally, there is some evidence that at least part of the within-group differences in evaluative behaviours that subordinates perceived were intended by the superior. As a whole, these findings, especially when taking into account some of the limitations of the two empirical studies, have a number of implications for future research.

First, these findings suggest that "evaluative styles", as a consistent pattern of evaluative behaviours of one supervisor, are not likely to exist. Although an average evaluative style - whether perceived by subordinates, intended by the supervisor, or perhaps both- can be assessed for a single supervisor, such an average style ignores important within-group variances, which makes the concept of style as a characteristic of a supervisor rather meaningless. Following leadership literature, where the concept of average leadership style (ALS) is "seriously out of vogue" (Schriesheim, Cogliser, & Neider, 1995, p.117) and leadership style has been replaced by a focus on different leadership dimensions, I suggest to replace the concept of "evaluative style" with that of "evaluatorship". The term "evaluatorship" can be used as an umbrella concept in future research that refers to a range of evaluative behaviours and attitudes of supervisors displayed when evaluating the performance of subordinates. In the context of management accounting, important

evaluative behaviours (as part of evaluatorship) could be the use of accounting information such as budgets, performance measurement diversity, budgetary participation or other forms of voicing opportunities, or different forms of subjectivity.

Second, the empirical evidence in this study suggest that perceived or intended evaluatorship is more likely operating at the individual level of analysis than at a group-level. This finding may be dependent on the particular evaluative behaviour considered in this study. Feedback quality as one aspect of evaluatorship may vary primarily at the individual level of analysis, while other characteristics may vary primarily at the level of parts within groups. For example, Moers (2005) investigated bias and discretion in performance evaluation. But if there is a forced distribution, a supervisor may base performance ratings also on the relative performance of each subordinate in comparison to other subordinates within the same group. Thus, perhaps the most appropriate level of analysis to study the use of bias in performance ratings could be that of the within-group level. Furthermore, the data reported here was collected at the level of either the supervisor or the subordinate as a whole. The results from the first empirical study, however, partly support a dyadic approach. A dyadic approach would imply that data on evaluatorship should be collected from matched pairs of supervisors and subordinates. Since in the second empirical study data was collected from subordinates only, the data did not allow me to test the dyad as an appropriate level of analysis. This would be a fruitful avenue for future research, with many research opportunities. For example, even at the dyadic level, the theoretical question is whether we expect wholes or parts at this level. Furthermore, the hierarchical nestedness of dyads within groups is also of interest. For example, are some dyads within groups more homogeneous than other dyads within the same group? Finally, studying the development of evaluatorship in the context of dyadic relationships over time would also be an interesting topic for further research. The leadership literature on LMX is just starting to consider such issues (see, for example, Henderson et al., 2009; Nahrgang, Morgeson, & Ilies, 2009). Future research should try to incorporate recent findings from this literature into management accounting research on performance evaluation.

Third, consideration of levels of analysis as in the current study is relevant for many other concepts included in management accounting research, and particularly research on performance evaluation. Consider performance for example, or task uncertainty, or environmental uncertainty, which may all occur at different levels-of-analysis, depending on the exact theorising of the construct. Again, measurement and data analysis in these cases should fit the level of theory.

Fourth, even though evaluatorship could be viewed as an individual level variable, the relationship with other variables may occur at different levels than the individual level of analysis. Thus, although it is important to consider level of analysis issues for single constructs as in this study, future research should extend the analyses to levels-of-analysis issues in relationships, both bivariate, multivariate, and moderated relationships. For example, does the relationship between feedback quality and procedural justice perceptions (Hartmann & Slapnicar, 2009) hold at the individual level or (also) at the dyadic level or even at the group level? Addressing questions like this should be guided by theory, and the data analysis should confirm the hypothesised relationships at the specified level of analysis.

Notwithstanding the limitations of this study, in particular the limited scope of the investigation and the small sample size, this study discloses fruitful avenues for future research on the role of supervisors in performance evaluation. When careful consideration is given to the level of theory, level of measurement, and level of analysis, the ambiguity of findings that characterises earlier research on this topic, with the RAPM literature as the most well known example, may be prevented in the future. In particular, this study shows that future research would benefit from combining insights from the leadership literature with insights from the accounting literature. Such integration is rare in both fields. Yet, the management and leadership literature may profit from accounting studies that study the role of supervisors in performance evaluation as a complement to the design of the performance evaluation system, taking into account the appropriateness of the performance measures used within the evaluation. In the accounting literature, in turn, a stronger focus on the role of supervisors, taking a more managerial approach, is likely to contribute to a more thorough understanding of the effects of performance measures on the motivation and behaviour of lower level managers and employees. Such research would truly contribute to a better understanding of performance evaluation as an important tool from the management control toolbox.

References

Adams, J. S. (1963). Towards an understanding of inequity. *Journal of Abnormal and Social Psychology, 67*(5), 422-436.

Bliese, P. D., & Halverson, R. R. (2002). Using random group resampling in multilevel research: An example of the buffering effects of leadership climate. *The Leadership Quarterly, 13*(1), 53-68.

Briers, M., & Hirst, M. (1990). The role of budgetary information in performance evaluation. *Accounting, Organizations and Society, 15*(4), 373-398.

Brownell, P. (1983). Leadership style, budgetary participation and managerial behavior. *Accounting, Organizations and Society, 8*(4), 307-321.

Castro, S. L. (2002). Data analytic methods for the analysis of multilevel questions: A comparison of intraclass correlation coefficients, rwg(j), hierarchical linear modeling, within- and between-analysis, and random group resampling. *The Leadership Quarterly, 13*(1), 69-93.

Charnes, J. M., & Schriesheim, C. A. (1995). Estimation of quantiles for the sampling distribution of the rWG within group agreement index. *Educational and Psychological Measurement, 55*(3), 435-437.

Cohen, A., Doveh, E., & Nahum-Shani, I. (2009). Testing agreement for multi-item scales with the Indices rWG(J) and AD M(J). *Organizational Research Methods, 12*(1), 148-164.

Dansereau, F., Alutto, J. A., & Yammarino, F. J. (1984). *Theory testing in organizational behavior: The varient approach*. Englewood Cliffs, N.J. : Prentice-Hall.

Dansereau, F., Graen, G., & Haga, W. J. (1975). A vertical dyad linkage approach to leadership within formal organizations : A longitudinal investigation of the role making process. *Organizational Behavior and Human Performance, 13*(1), 46-78.

DeCoster, D.T., & Fertakis, J. P. (1968). Budget-induced pressure and its relationship to supervisory behavior. *Journal of Accounting Research, 6*(2), 237-246.

Dunk, A. S. (2003). Moderated regression, constructs and measurement in management accounting: A reflection. *Accounting, Organizations and Society, 28*(7-8), 793-802.

Dunlap, W. P., Burke, M. J., & Smith-Crowe, K. (2003). Accurate tests of statistical significance for rWG and average deviation interrater agreement indexes. *Journal of Applied Psychology, 88*(2), 356-362.

George, J. M. (1990). Personality, affect, and behavior in groups. *Journal of Applied Psychology, 75*(2), 107-116.

Gibbs, M., Merchant, K. A., van, d. S., & Vargus, M. E. (2004). Determinants and effects of subjectivity in incentives. *Accounting Review, 79*(2), 409-436.

Graen, G. B., & Uhl-Bien, M. (1995). Relationship-based approach to leadership: Development of leader-member exchange (LMX) theory of leadership over 25 years: Applying a multi-level multi-domain perspective. *The Leadership Quarterly, 6*(2), 219-247.

Graen, G., & Schiemann, W. (1978). Leader-member agreement: A vertical dyad linkage approach. *Journal of Applied Psychology, 63*(2), 206-212.

Hartmann, F. G. H., & Moers, F. (1999). Testing contingency hypotheses in budgetary research: An evaluation of the use of moderated regression analysis. *Accounting, Organizations and Society, 24*(4), 291-315.

Hartmann, F. G. H., & Moers, F. (2003). Testing contingency hypotheses in budgetary research using moderated regression analysis: A second look. *Accounting, Organizations and Society, 28*(7-8), 803-809.

Hartmann, F. G. H. (2000). The appropriateness of RAPM: Toward the further development of theory. *Accounting, Organizations and Society, 25*(4-5), 451-482.

Hartmann, F., & Slapničar, S. (2009). How formal performance evaluation affects trust between superior and subordinate managers. *Accounting, Organizations and Society, 34*(6-7), 722-737.

Henderson, D. J., Liden, R. C., Glibkowski, B. C., & Chaudhry, A. (2009). LMX differentiation: A multilevel review and examination of its antecedents and outcomes. *The Leadership Quarterly, 20*(4), 517-534.

Hopwood, A. G. (1973). *An accounting system and managerial behaviour*. Farnborough [Hants.] (D.C. Heath Ltd., Westmead, Farnborough, Hants.) : Saxon House ; 1973.

Hopwood, A. G. (1972). An empirical study of the role of accounting data in performance evaluation. *Journal of Accounting Research, 10*(SUPPL.), 156-182.

Hopwood, A. G. (1974). Leadership climate and the use of accounting data in performance evaluation. *The Accounting Review, 49*(3), 485-495.

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69*(1), 85-98.

James, L. R., Demaree, R. G., & Wolf, G. (1993). Rwg: An assessment of within-group interrater agreement. *Journal of Applied Psychology, 78*(2), 306-309.

Jiambalvo, J., & Pratt, J. (1982). Task complexity and leadership effectiveness in CPA firms. *The Accounting Review, 57*(4), 734-750.

Klein, K. J., Dansereau, F., & Hall, R. I. (1994). Levels issues in theory development, data collection, and analysis. *Academy of Management Review, 19*(2), 195-229.

LeBreton, J. M., James, L. R., & Lindell, M. K. (2005). Recent issues regarding rWG, rWG, rWG(J), and rWG(J). *Organizational Research Methods, 8*(1), 128-138.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*(4), 815-852.

Ludtke, O., & Robitzsch, A. (2009). Assessing within-group agreement: A critical examination of a random-group resampling approach. *Organizational Research Methods, 12*(3), 461-487.

Luft, J., & Shields, M. D. (2003). Mapping management accounting: Graphics and guidelines for theory-consistent empirical research. *Accounting, Organizations & Society, 28*(2), 169-249.

Moers, F. (2005). Discretion and bias in performance evaluation: The impact of diversity and subjectivity. *Accounting, Organizations and Society, 30*(1), 67-80.

Nahrgang, J. D., Morgeson, F. P., & Ilies, R. (2009). The development of leader–member exchanges: Exploring how personality and performance influence leader and member relationships over time. *Organizational Behavior and Human Decision Processes, 108*(2), 256-266.

Noeverman, J., Koene, B. A. S., & Williams, R. (2005). Construct measurement of evaluative style: A review and proposal. *Qualitative Research in Accounting and Management, 2*(1), 77.

O'Brien, R. G. (1979). A general ANOVA method for robust tests of additive models for variances. *Journal of the American Statistical Association, 74*(368), 877-880.

O'Brien, R. G. (1981). A simple test for variance effects in experimental designs. *Psychological Bulletin, 89*(3), 570-574.

Otley, D. T. (1978). Budget use and managerial performance. *Journal of Accounting Research, 16*(1), 122-149.

Otley, D. T., & Pierce, B. J. (1995). The control problem in public accounting firms: An empirical study of the impact of leadership style. *Accounting, Organizations and Society, 20*(5), 405-420.

Otley, D., & Fakiolas, A. (2000). Reliance on accounting performance measures: Dead end or new beginning? *Accounting, Organizations and Society, 25*(4-5), 497-510.

Otley, D., & Pollanen, R. M. (2000). Budgetary criteria in performance evaluation: A critical appraisal using new evidence. *Accounting, Organizations and Society, 25*(4-5), 483-496.

Pasisz, D. J., & Hurtz, G. M. (2009). Testing for between-group differences in within-group interrater agreement. *Organizational Research Methods, 12*(3), 590-613.

Pratt, J., & Jiambalvo, J. (1981). Relationships between leader behaviors and audit team performance. *Accounting, Organizations and Society, 6*(2), 133-142.

Pratt, J., & Jiambalvo, J. (1982). Determinants of leader behavior in an audit environment. *Accounting, Organizations and Society, 7*(4), 369-379.

Scandura, T. A., & Lankau, M. J. (1996). Developing diverse leaders: A leader-member exchange approach. *The Leadership Quarterly, 7*(2), 243-263.

Schriesheim, C. A., Castro, S. L., Zhou, X., & Yammarino, F. J. (2001). The folly of theorizing "A" but testing "B" A selective level-of-analysis review of the field and a detailed leader-member exchange illustration. *Leadership Quarterly, 12*(4), 515-551.

Schriesheim, C. A., Cogliser, C. C., & Neider, L. L. (1995). Is it "trustworthy"? A multiple-levels-of-analysis reexamination of an ohio state leadership study, with implications for future research. *Leadership Quarterly, 6*(2), 111-145.

Schriesheim, C. A., Castro, S. L., & Cogliser, C. C. (1999). Leader-member exchange (LMX) research: A comprehensive review of theory, measurement, and data-analytic practices. *The Leadership Quarterly, 10*(1), 63-113.

Schriesheim, C. A., Neider, L. L., & Scandura, T. A. (1998). Delegation and leader-member exchange: Main effects, moderators, and measurement issues. *The Academy of Management Journal, 41*(3), 298-318.

Steelman, L. A., Levy, P. E., & Snell, A. F. (2004). The feedback environment scale: Construct definition, measurement, and validation. *Educational and Psychological Measurement, 64*(1), 165-184.

Steelman, L. A., & Rutkowski, K. A. (2004). Moderators of employee reactions to negative feedback. *Journal of Managerial Psychology, 19*(1), 6-18.

Vagneur, K., & Peiperl, M. (2000). Reconsidering performance evaluative style. *Accounting, Organizations and Society, 25*(4-5), 511-525.

Vroom, V. H. (1964). *Work and motivation* Oxford, England: Wiley.

Whitaker, B. G. (2007). Explicating the links between the feedback environment, feedback seeking, and job performance. The University of Akron/ OhioLINK. Dissertation retrieved (November 10, 2009) from http://rave.ohiolink.edu/etdc/view?acc_num=akron1187456858

Wiersma, E. (2008). An exploratory study of relative and incremental information content of two non-financial performance measures: Field study evidence on absence frequency and on-time delivery. *Accounting, Organizations and Society, 33*(2-3), 249-265.

Yammarino, F. J., Dionne, S. D., Uk Chun, J., & Dansereau, F. (2005). Leadership and levels of analysis: A state-of-the-science review. *The Leadership Quarterly, 16*(6), 879-919.

Yammarino, F. J., & Markham, S. E. (1992). On the application of within and between analysis: Are absence and affect really group-based phenomena? *Journal of Applied Psychology, 77*(2), 168-176.

Table 1. Intended and perceived role of area manager X in performance evaluation of branch managers

	Area Manager	Branch managers		
	X	X1	X2	X3
Description of function of branch manager	Manner of functioning. Not just long and hard working, but especially ATP as a characteristic of the management style. Emphasis on managerial behaviour.	Realise company goals, but preferably do more than that. Responsible that mail is delivered, no matter the circumstances (keep customers satisfied). Control costs, but also meet quality targets. Keep my employees motivated.	Question not asked in interview.	Realise company goals in such a way that all employees like their jobs
Performance evaluation in general:	Contract needs to be met. Targets are realistic. In general: no excuses and no discussion about it. If targets are not met, branch managers need to have a real good and realistic explanation, and provide evidence about what they have done about it.	Based on management contract (50%), and 50% on how you attain the targets, your effort, management behaviour. But you need to attain the targets.	Based on management contract, but a few items are highlighted by AM. How you do things is important. But you need to attain the targets.	Based on targets in contract, no agreements on desired behaviour.
- Performance areas (targets) in contract	1. Costs; 2. Quality; 3. Customer satisfaction; 4. Employee satisfaction	1. Costs; 2. Quality; 3. Customer satisfaction; 4. Employee satisfaction	1. Costs; 2. Quality; 3. Customer satisfaction; 4. Employee satisfaction; 5. Contribute to area as a whole	1. Costs; 2. Quality; 3. Customer satisfaction; 4. Employee satisfaction
- Role of area manager	In addition to contract emphasise management behaviour: ATP, management style. “A manager who shows ATP will without any doubt meet his targets.”	AM emphasises management behaviour. This is a general agreement, but not a formal part of the performance evaluation cycle. AM calls it ATP: “don’t bother me with your problems”. “He does not like to deal with someone else’s problems”.	AM emphasises how you manage your branch. This should contribute to attaining your targets, also in the long run. And he has emphasised a few items from the contract that are controllable.	AM emphasises in meetings that there is more than numbers and costs in meetings. But the contract only contains what needs to be achieved, not how.
- Performance areas to which area manager attaches most	Targets from the contract (management by exception). The	1. Agreements in contract; 2. Management behaviour: ATP	1. Costs: 50%; 2. Quality, but not sure of the	1. 50% costs; 2. 35% quality;

importance when evaluating branch managers?	targets are the boundaries within which the branch managers need to operate. But not just meeting targets, they also have to show ATP.		percentage; 3. management style, my behaviour: ATP	3. 15% others, such as client and employee satisfaction

Table 2. Intended and perceived role of area manager Y in performance evaluation of branch managers

	Area Manager	Branch managers		Service manager
	Y	Y1	Y2	Y3
Description of function of branch manager	Support adequate management of mail delivery, control costs and quality, people management.	Take care that customers in our geographical area receive their mail in time and correctly, with high quality standards	Making money for the company; being on the shopping floor; deliver mail with high productivity (mail volume divided by labour capacity)	
Performance evaluation in general:	Based on management contracts. Benchmarking is very important. Some branch managers are not able or willing to perform. Benchmarking will show that. These managers get a second chance, but not a third.	Based on management contract. No agreements on how targets are achieved.	Based on management contract. Contract is contract. Once the targets are set, they cannot be changed.	Based on management contract. For branch managers: contract. Also on how targets need to be achieved: customer satisfaction and employee motivation.
- Performance areas (targets) in contract	1. Costs; 2. Quality.	1. Costs; 2. Quality.	All targets have to be met, and all targets are important.	For branch managers: 1. Labour costs and productivity; 2. Other costs; 3. Quality.
- Role of area manager	Costs and quality targets are company norms. In setting targets for customer satisfaction and employee motivation, it is important to take into account current situation within the branch. More meetings with branches with lower performance than with good performing branches. In distinguishing between good and excellent branch managers, it is important how the manager achieved his results: is it really good in the long run? AM makes standard contract.	AM is very to the point. “If you have to do something, he simply says: “Solve it!” How you do that is up to you”. AM focuses on the numbers. If you do not meet your targets, he will ask what you are going to do to meet them. Every quarter of a year the items from the contract are discussed in a meeting with AM.	AM allows branch managers to make an additional agreement on one item in particular. For this item a stretching target is agreed upon. “The AM says: “if I do not visit the branch, it is good”.” (management by exception) “If AM visits the branches, he will sit with the employees, work with them for a short period, to grasp the work atmosphere”.	AM evaluates performance primarily on basis of objectified targets. Targets have to be met. Targets will differ between branch managers, because they are personalized. AM will visit all branches at least once a year. “But if performance is below target, he will visit a branch more often and ask the branch manager: “what are you going to do to still meet your contract? And how can we help?”.”

- Performance areas to which area manager attaches most importance when evaluating branch managers?	Costs and quality. If you don't meet these targets, you have a problem. Customer satisfaction, employee motivation and absenteeism are important indicators to see if a manager besides paying attention to costs and quality also is a people manager.	Costs and quality.	Costs (labour), productivity and quality.	1. Productivity; 2. Quality: same weighting as productivity. 3. Employee motivation. This has a lower weight, and is more subjective than the first two.

Table 3. Descriptive statistics of feedback quality in earlier studies

Study	Mean	SD	Scale	n	Cronbach's alpha
Steelman, Levy, & Snell (2004); Steelman & Rutkowski (2004)	4.76	1.40	7-point Likert scale	405	.92
Whitaker (2007)	4.22	0.93	7-point Likert scale	202	.87
Hartmann & Slapnicar (2009)	NA	NA			

Table 4. Within-group agreement coefficients and WABA I results for feedback quality

Groups	n	r_{WG}			Eta		E	Corrected F-test[a]	Inference
		Un	**SS**	**Emp**	Within	Between			
1	3	0.99	0.98	0.98					
2	6	0.96	0.93	0.91					
3	6	0.95	0.92	0.89					
4	8	0.88	0.78	0.68			NA		
5	5	0.88	0.77	0.66					
6	6	0.94	0.90	0.86					
7	7	0.95	0.91	0.89					
8	10	0.92	0.86	0.81					
	51				.844	.537	.637†	.404	Equivocal

[a] Degrees of freedom are 43, 7

† Significant by the 15° test

†† Significant by the 30° test

* $p < .05$

** $p < .01$

Publications in the Report Series Research* in Management

ERIM Research Program: "Finance and Accounting"

2010

Within- and Between-group Agreement in Supervisor's Evaluative Behaviours: Do evaluative 'styles' exist?
Jan Noeverman
ERS-2010-002-F&A
http://hdl.handle.net/1765/17700

---

* A complete overview of the ERIM Report Series Research in Management:
https://ep.eur.nl/handle/1765/1

ERIM Research Programs:
LIS Business Processes, Logistics and Information Systems
ORG Organizing for Performance
MKT Marketing
F&A Finance and Accounting
STR Strategy and Entrepreneurship