Effectiveness of Performance Appraisal: An Integrated Framework

Abstract

Based on a robust analysis of the existing literature on performance appraisal, this paper makes a case for an integrated framework of effectiveness of performance appraisals. To achieve this, it draws on the expanded view of measurement criteria of effectiveness of performance appraisal, i.e., utilization criteria (purposefulness), qualitative criteria (fairness), and quantitative criteria (accuracy), and identifies their relationships with the PA outcomes criteria, i.e., ratee reactions. The analysis reveals that the expanded view of utilization criteria includes more theoretical anchors for the purposes of performance appraisal and relates to various aspects of human resource functions, e.g., feedback and goal-orientation. The expansion in the qualitative criteria suggests certain newly established nomological networks, which were ignored in the past (e.g., the relationship between distributive justice and organization-referenced outcomes). Further, refinements in quantitative criteria reveal a more comprehensive categorization of rating biases. Coherence amongst measurement criteria has resulted in a ratee reactions-based integrated framework, which should be useful for both researchers and practitioners.

Keywords: Performance appraisal, utilization criteria, qualitative criteria, quantitative criteria, purposefulness, fairness, accuracy

Introduction

Effectiveness of performance appraisal (EPA) has remained one of the most vital subjects in the theory and practice of performance appraisal (PA). Earlier on, it merely referred to how well the complex process of assessing employee work performance was operated (Lawler et al. 1984; Lee 1985; Keeping and Levy 2000). Now it has grown into a comprehensive evaluative approach to managing the PA system (Chiang and Birtch 2010). This approach uses certain 'measurement' and 'outcome' criteria and assesses the antecedent-outcome relationships that manifest EPA.

During the last three decades, PA literature has revealed a range of subordinate measurement and outcome criteria, albeit piecemeal. While developing the concept of EPA, Jacobs et al. (1980) proposed a system that established three categories of measurement criteria, i.e., utilization, qualitative, and quantitative criteria. According to PA researchers (e.g., Hedge and Teachout 2000; Kudisch et al. 2006; Roch 2006; Wood and Marshall 2008; Chiang and Birtch 2010; Linna et al. 2012) the utilization criteria address the question: why performance appraisals are conducted. Hence, it deals with the purposes and uses of performance appraisal. The qualitative criteria relate to a set of rules and practices that ensure fairness in the performance appraisal system. The quantitative criteria refer to rating accuracy.

In addition, researchers maintain that PA is considered effective when its key stakeholders (i.e., ratees) reckon it useful (Giles and Mossholder 1990; Keeping and Levy 2000; Levy and Williams 2004; Walsh and Fisher 2005; Roberson and Stewart 2006), i.e., ratee reactions criteria. Pichler (2012, p. 710) defines them as "individual-level attitudinal evaluations of and responses to the performance appraisal process." In the light of this definition, this paper focuses on ratee reactions-based EPA outcomes, and thus, uses Greenberg's (1990) taxonomy that categorizes ratee reactions into two groups, i.e., person - referenced outcomes (ratee satisfaction with reward, the rater, rating system, ratings, and feedback) and organization-

referenced outcomes (organizational commitment, self-evaluation, feedback seeking behaviour, role-clarity, and perceived detriments to EPA).

Although organizations have instilled one set of measurement criteria or another, they seem to be discontented with their choices. Their complaint is that most PAs are ineffective, as they result in decreased employee performance (Latham et al. 2005) and increased employee dissatisfaction (Shrivastava and Purang 2011). This indicates that, by and large, PAs fail to contribute to human resource functions (Chiang and Birtch 2010) and organizational effectiveness (Taylor et al. 1995). Thus, responding to calls in the literature to propose a theoretically sound and broader view of each measurement criteria that may ensue desirable ratee reactions (e.g., Dipboye 1985; Griffeth and Bedeian 1989; Woehr and Huffcutt 1994; Cardy and Dobbins 1994; Murphy and Cleveland 1995; Fletcher 1995, 2001; Haines and St-Onge 2012; Roch et al. 2007), this paper aims to make three contributions to the field of PA. First, it presents the expanded view of measurement criteria of EPA, making a two-fold contribution to PA literature and practice. One, it highlights the rarely used PA purposes, i.e., role-definition and strategic; and two, it promulgates a classification of rating errors, i.e., rater-centric, ratee-centric, system-centric, and relation-centric. Second, the paper identifies relationships between measurement criteria of EPA and ratee reactions. Ratee reactions are considered as the most important PA criteria, as in practice these are deemed to be more important than other outcome criteria (Pichler 2012). Thus, this paper attempts to provide a ratee reaction-based view of EPA. Third, it proposes an integrated framework of EPA by two mechanisms. Firstly, by suggesting integration between all the measurement criteria and ratee reactions criteria, and secondly, by discussing the integration amongst the measurement criteria.

Method

Given the dispersed nature of the EPA literature, we adopted a structured review (Tranfield et

al. 2003) undertaking three decisive factors for search and selection of published literature, i.e., quality, relevance and recentness (see Figures 1, 2, and 3 for details). Unlike searching through databases (e.g., de Menezes and Kelliher 2011; Claus and Briscoe 2009), we targeted quality journals listed in the academic journal quality guide of Association of Business Schools (ABS) and Social Science Citation Index (SSCI). However, few articles published in two- and one-grade journals were also included in the sample and these articles were reviewed while carrying out the initial literature survey. On the homepage of each journal, the advanced search options were used to elicit relevant results. As a first step, main search terms of 'performance appraisal,' 'performance rating,' and 'performance evaluation' were applied. Afterwards, for searching within the results, major search terms were used. For example, for utilization criteria, the search terms were 'purpose', 'administrative', 'developmental', 'strategic', and 'role-definition'; for qualitative criteria, search terms of 'justice', 'fair', 'distributive,' 'procedural', 'interactional', 'interpersonal', and 'informational' were applied; and for quantitative criteria, the search terms were 'accuracy,' 'bias,' and 'error.' Search terms for employee reactions were 'reward', 'organizational commitment', 'feedback', 'self-monitor', 'self-appraisal', 'self-evaluation', and 'satisfaction'. All search terms were applied to the full text using the truncation symbol (*).

(Figure 1, 2 and 3 here)

The process produced 549 articles, which were skim read (rapid scanning of the entire article) to select the most relevant ones (Thomas 2004). Concentrating more on the concepts of PA relating to the theme of our study, i.e., EPA in general, and utilization, qualitative, quantitative and ratee reactions criteria made selection of relevant articles in particular (see Figure 2). A total of 127 articles, published in 37 journals falling under four subject categories, i.e., general management, human resource management, psychology, and organization studies, met the criteria per se. The selected journal articles include 104

empirical studies, 20 review papers, two triangulation studies, and one conceptual paper. With regard to periodization, we focused more on studies published in the year 2000 – 2012. However, keeping in mind the inconsistent research attention being paid to each of the EPA criteria during this timeline, studies published before 2000 were also included. Thus, selected articles include 57% of papers published in the year 2000 – 2012. Although 245 authors from 18 countries authored selected articles, 72% of the literature was contributed by US researchers. Thus, the assessment of Chiang and Birtch (2010) that most of the PA literature was US-oriented was found to be true. Figure 4 shows details about the principal author's country affiliation and countries where the studies were carried out.

(Figure 4 here)

Integration amongst the EPA Criteria

The proposed ratee reactions-based integrated framework of EPA is presented in Figure 5. In this section, the integrated framework is discussed in four parts. The first three parts discuss the relationships between the measurement criteria, i.e., utilization, qualitative, and quantitative, and ratee reactions criteria. The fourth part discusses the correlates amongst the measurement criteria.

(Figure 5 here)

Building on research that highlights ratees' perceptions as the most important criteria for determining the effectiveness of PA systems (e.g., Keeping and Levy 2000; Levy and Williams 2004; Roberson and Stewart 2006; Roch et al. 2007; Pichler 2012), the center of our analysis is ratee reactions. Our review of 127 studies provides both a theoretical rationale and sufficient empirical evidence that measurement criteria (i.e., utilization, qualitative, and quantitative), lead to ratee reactions. It translates that purposeful and fair PA practices result in positive person- and organizational-referenced ratee reactions (e.g., ratee satisfaction and organizational commitment), whereas rating errors/biases ensue negative outcomes (i.e.,

detriments to EPA), which manifest ratee dissatisfaction, low organizational commitment, etc. As the focus of our proposed integrated framework is on ratee reactions, therefore, PA professionals will find it useful to use it as a (felt) needs assessment approach to PA, i.e., employees'/ratees' needs.

Utilization criteria and ratee reactions criteria

Researchers have provided certain theoretical reflections on the PA purposes. These theories help lay a pathway for the PA purposes to be utilized as an EPA criteria. However, attention being paid to their empirical examination has been patchy. During the last three decades, most of the empirical research has been confined only to the administrative and developmental purposes (e.g., Dorfman et al. 1986; Farh et al. 1991; Zimmerman et al. 2008; Selvarajan and Cloninger 2011; Varma et al. 2008). As a result, very little research has discussed the role-definition and strategic purposes of PA (e.g., Youngcourt et al. 2007, for the former; Noe et al. 2003, for the latter).

Cleveland et al. (1989) inventoried, and then categorized 20 purposes of PA into four a priori defined factors. All purposes in the first factor, i.e., 'between individuals,' have been regarded as administrative purposes in the PA literature. These included: salary administration, promotion, retention or termination, recognition of individual performance, layoffs, and identification of poor performance. The second factor, i.e., 'within individuals,' focuses on the developmental purposes (Tziner et al. 2000; Tziner et al. 2001). These were: identification of individuals' training needs, performance feedback, determination of transfers and assignments, and identification of individuals' strengths and weaknesses. Some uses under the remaining factors (i.e., 'system maintenance' and 'documentation') relate to the strategic and role-definition purposes. These include: 'evaluate goal achievement' and 'assist in goal identification,' for the former; and 'reinforce authority structure,' for the latter. Using a self-completion questionnaire survey in 74 Jordanian organizations (36 public and 38

private), Abu-Doleh and Weir (2007) partially replicated the study by Cleveland et al. (1989). Their sample of private organizations substantiated Cleveland et al. more than the sample of public organizations. That is, private organizations' PA had a significantly greater impact than public sector on promotion, retention/termination, lay-offs, identifying individual training needs, transfers and assignments.

Administrative purposes of PA. The relationship between administrative purposes of PA and ratee reactions has gained the support of expectancy and equity theories. Expectancy theory explains that in order to raise the employees' interests in the organizational setting, they should be rewarded corresponding to their performance. This is because ratees expect that the higher the performance is, the greater the reward will be (Harder 1992; Kudisch et al. 2006). Moreover, if the amount of reward corresponds to the level of ratee performance, they may perceive the equity to be achieved (Chiang and Birtch 2010). If it is otherwise, then the ratees perceive that they are under-rewarded; hence, they might decrease their performance to balance out the equity in their own way (Harder 1992).

Supporting the above theoretical rationale, Chiang and Birtch (2010) argue that administrative purposes and financial needs of employees have always been current and short-term-oriented. Hence, a strong linkage between performance results and reward may exist (Bititci et al. 2012). Chiang and Birtch's Hong Kong and Singapore samples empirically supported this. Similarly, another study with a cultural perspective, has confirmed similar relationship, which is based on Latin America and Taiwan samples (see Milliman et al. 2002). Analysis if such studies confirm that, the more the rewards are tied to PA results, the more the EPA will be perceived (Lawler 2003).

Administrative purposes also relate to ratee satisfaction (with the rating system and the rater) and commitment. In their cross-sectional study (n = 599 employees), Youngcourt et al. (2007) report significant correlations between administrative PA, and satisfaction with the

rating system (r = .43) and affective commitment (r = .36). Using the structural equation modelling, these researchers also found administrative purposes to have an effect on the ratee reactions (β = .53 and β = .03, respectively). A longitudinal experimental study by Boswell and Boudreau (2002) reveal similar findings. These scholars divided the sample (n = 116 employees) into the treatment group (rated for administrative purposes) and the control group (rated for both administrative and developmental purposes) and found significant correlations between PA ratings about ratees in both the treatment and the control groups, and their satisfaction with the rating system (r = .38 and .29, respectively). Likewise, another longitudinal study (n = 242 dyads) by Dorfman et al. (1986) found administrative purposes of PA to have a significant effect on ratee satisfaction with the rating system and the rater (β = .22) as one factor. Thus, the PA used for administrative purposes may have a positive significant relationship of the ratees' satisfaction with reward, the rating system and the rater, and organizational commitment.

Developmental purposes of PA. Employee development is said to be amongst the primary purpose of PA (Cleveland et al. 1989; Nurse 2005). While identifying the desired emphasis on developmental purposes of PA, Milliman et al. (2002) found that a high priority was reported by samples in the American continent, Australia, and Taiwan. However, the emphasis was moderate in some Asian countries. Chiang and Birtch (2010) carried out their study in seven countries (Canada, Hong Kong, Finland, Singapore, Sweden, the UK, and the US) and found a strong consensus across the sample that PA was being used for employee development, albeit to varying degrees.

Social exchange theory explains that when individuals feel that the organization is keen for their long-term development, they try to reciprocate (Youngcourt et al. 2007; Kuvaas 2006; Chiang and Birtch 2010). The most likely return on long-term development is employee organizational commitment (Tziner et al. 2001). As assumed by the social exchange theory,

employees may feel motivated to maximize their outcomes (Roberson and Stewart 2006) and demonstrate positive attitudes (Kudisch et al. 2006). Substantiating this theory, a review by Beer (1981) and the following empirical studies suggest that developmental PA may lead to ratee commitment and satisfaction (with the rating system and the performance feedback).

Using a heterogeneous sample from three different countries (the US, Canada, and Israel), Tziner et al. (2001) estimated inter-correlations among administrative and developmental purposes, and affective commitment. They found developmental purposes to have a higher degree of corrected correlation (r = .38) with affective commitment than administrative purposes (r = .32). Youngcourt et al. (2007) found developmental purposes to have significant correlations with satisfaction with the rating system (r = .43) and affective commitment (r = .37). They also found developmental purposes to have predicted affective commitment (β = .49). In a longitudinal study by Tharenou (1995) of 172 employees of the Australian Federal Agency (108 appraised and 64 non-appraised) were surveyed, both before and after the introduction of developmental PA. With respect to ratee satisfaction with the feedback, an increase in the post-test scores was found. This increase is accounted for by the developmental performance appraisal.

Some literature prefers administrative purposes to developmental purposes and vice versa. For example, a meta-analysis of 22 studies (Jawahar and Williams 1997) reveals that administrative purposes have been the focus of research than have the developmental purposes. In contrast, a survey of 276 students (Hong Kong: 141 and UK: 135) by Snape et al. (1998) reveals that the Hong Kong sample appreciates administrative purposes more and developmental purposes less than the UK sample does. Drawing from these contrasting opinions, it is learnt that the relative importance of administrative and developmental purposes over each other may be assessed, particularly while predicting the common response variables, i.e., organizational commitment and satisfaction with the rating system.

Strategic purposes of PA. Goal-setting theory regards behaviours as goal-directed. Using the goal-setting lens, van Dierendonck et al. (2007) maintain that ratees use performance ratings about them for self-monitoring. This is for assessing whether their performance is consistent with their goals or otherwise. However, before letting this desirable state occur, organizations solicit functional relationships between the organizational goals and the goals of its employees (Aguinis 2009). This is because organizations want ratees to self-monitor so that they pursue only those goals, which are linked to organizational goals. This is why London et al. (2004) consider that 'setting goals' is better than 'assigning goals.'

Several researches have suggested the relationship between PA ratings for strategic purposes and self-monitoring (see e.g., Miller and Cardy 2000; Jawahar 2001, 2005), and thus, the latter is regarded as an integral component of the PA system (Campbell and Lee 1988). In addition, Renn and Fedor (2001) have identified that performance feedback-related research has focused largely on identifying antecedents of feedback seeking behaviour and goal orientation being one of them. Therefore, it is expected that the strategic PA may rouse ratees to self-monitor and seek performance feedback.

Role-definition purposes of PA. Role-definition purposes of PA remain the least explored ones. This paper found only one empirical study (Youngcourt et al. 2007) that even partially drew attention to it. According to Duarte et al. (1994), roots of role-definition purposes can be found in dyad organizations. In fact, the role of an employee at workplace changes over time; therefore, based on PA results, the supervisor defines and communicates roles to the subordinate. However, ideally, the process is completed only when the subordinate seeks feedback on their performance-position gaps, and this is the ratee reaction that organizations desire and researchers call for investigation (Levy and Williams 2004).

Youngcourt et al. (2007) have reported significant correlations between role-definition purposes and ratee satisfaction with the rating system (r = .49) and affective commitment (r =

.40 and β = .03). Although the existing literature provides little support to the above-mentioned relationships (see Dahling et al. 2012), it gives a lead to associating role-definition PA with feedback seeking behaviour, organizational commitment and satisfaction with the rating system.

Ratee reactions are an outcome of PA purposes that is critical for the long-term EPA (Mount 1984). However, the literature highlights that PA researchers maintain two different opinions about relationships between PA purposes and ratee reactions. One suggests that each specific PA purpose may predict a unique outcome. The other suggests simultaneous effects of a combination of PA purposes on some outcomes. In support of the former theory, Beer (1981) suggested uncoupling administrative and developmental purposes in order to improve the PA system. Providing empirical support for this, two studies (Stephan and Dorfman 1989; Zimmerman et al. 2008) have suggested administrative and developmental purposes to be unique predictors of 'task performance' and 'organizational goal performance,' respectively. The former was an experimental study (n = 72 students) and the latter was a longitudinal study (n = 396 employees).

Substantiating the latter theory, three empirical studies (Harris et al. 1995; Tziner et al. 2001; Tziner et al. 2002) found significant correlations between administrative and developmental purposes (r = .58, .72 and .16 respectively). Providing a stronger evidence, Youngcourt et al. (2007) have reported that correlations among administrative, developmental, and role-definition purposes were $r \geq .60$, at $p < 0.1$. These results help infer that if a category of PA purposes is not included in the research model of an empirical study undertaking utilization criteria as a predictor, it may affect the framework as a nuisance variable.

Qualitative criteria and ratee reactions criteria

The qualitative criteria address the fairness perceptions of ratees (Giles et al. 1997). Generally,

fairness is derived from equity theory that refers to perceived outcome-related fairness (McDowall and Fletcher 2004). However, it is based on organizational justice theory. Under the tenets of this theory, forms of justice are categorized as one-, two-, three-, and four-factor models. In the one-factor model, major forms of justice, i.e., distributive and procedural, are measured through one scale, and being highly correlated (Welbourne et al. 1995; Sweeney and McFarlin 1997). Greenberg's (1986) empirical investigation laid the foundation for the two-factor model. In his exploratory study (n = 217 employees), Greenberg showed that distributive justice and procedural justice were two distinct dimensions. Although the two-factor conceptualization incorporated distributive and procedural justice in one model, these were treated differently (Greenberg 1990).

The three-factor model was developed to address the inclusion of interactional fairness in the justice literature (e.g., Bies and Shapiro 1987; Barling and Phillips 1993; Martocchio and Judge 1995; Skarlicki and Folger 1997). In the early 2000s, the four-factor model was conceptualized and it provided a clearer expression of all forms of justice by categorizing interactional justice into two factors, i.e., interpersonal and informational justice. While propounding the dimensionality of the four-factor model, Colquitt (2001) demonstrated its construct and predictive validities adequately. Since then and until now, this conceptualization has been used in most empirical research (e.g., McDowall and Fletcher 2004; Jawahar 2007; Kass 2008; Jepsen and Rodwell 2009; Colquitt and Rodell 2011). However, without assessing the "fair process effect" (Folger et al. 1979), i.e., the outcomes of fairness/justice, it cannot be said that justice is done. Thus, the positive relationship between the four-factor justice and person- and organization-referenced ratee reactions indicates PA fairness.

Distributive justice. Initially, distributive justice used to deal with the fairness of decision outcomes (Colquitt 2001) and distribution of outcomes, e.g., reward (Jawahar 2007). Under the umbrella of the two-factor model (McFarlin and Sweeney 1992; Sweeney and McFarlin

1993), it was proposed to be related to only person-referenced outcomes, e.g., job satisfaction. However, recent research has included the evaluation of the outcomes-related fairness in its scope. This was done to embed norms of distribution, such as equity or equality (Colquitt 2001). This expanded view of distributive justice justified its measurement as a separate justice factor. The following empirical investigations support the relationships among distributive justice and person as well as organization-referenced ratee reactions.

Drawing on person-referenced outcomes, four empirical studies (Foley et al. 2005; McFarlin and Sweeney 1992; Sweeney and McFarlin 1997; Jepsen and Rodwell 2009) found distributive justice to have a positive effect on ratee job satisfaction, albeit to varying degrees, i.e., β = 0.11, 0.30, 0.18, and 0.23, respectively. It is noteworthy that Jepsen and Rodwell (2009) reported the β coefficient only for their male sample (n = 265), as it was insignificant for their female sample (n = 113). Alongside the distal variable of job satisfaction, distributive justice relates to certain proximal variables as well, e.g., ratee satisfaction with ratings, rating system, the rater, the performance feedback and reward.

Holbrook (1999) suggested a significant correlation between distributive justice and ratee satisfaction with ratings (r = .72). Later, Colquitt (2001) and Jawahar (2007) examined this relationship in artificial and actual respondents, i.e., n = 301 students and n = 163 employees respectively and found distributive justice to have a significant effect on ratee satisfaction with ratings (β = .73 and β = .83, respectively). Ratee satisfaction with the rating system is the second proximal variable that two empirical studies (Korsgaard and Roberson 1995; Elicker et al. 2006) have reported to find an association with distributive justice (r = .75 and r = .79, respectively). Ratee satisfaction with the rater and the performance feedback have been found to have influenced by distributive justice, e.g., McFarlin and Sweeney (1992) and Sweeney and McFarlin (1997) (β = .15 and .37, respectively), for the former, and Jawahar (2007) (β =

.33), for the latter. McFarlin and Sweeney (1992) and Colquitt (2001) have also found distributive justice explaining variance in rewards (β = .52 and β = .36, respectively).

Drawing on the organization-referenced outcomes, six empirical studies supporting the relationship between distributive justice and organizational commitment, two have reported correlations between them, and four have suggested that the former may predict the latter. Conducting a scenario-based experiment on 240 students, Holbrook (1999) reported a positive correlation between the two constructs (r = .73). Likewise, the correlation matrix generated from 92 matched manager-employee dyads in another study (Heslin and VandeWalle 2011) revealed a significant association between distributive justice and organizational commitment. However, while teasing apart the dimensions of organizational commitment, they reported the coefficients as r = .41 for affective commitment, and r = .33 for normative commitment.

With regard to predictive relationship, in a survey of 877 Protestant clergies in Hong Kong (see Foley et al. 2005) reported a positive effect of distributive justice on organizational commitment (β = .19). McFarlin and Sweeney (1992) also supported this relationship but in one study they reported greater effect (β = .52, p < .01) and smaller, yet more significant (β = .14, p < .001) in the other (see McFarlin and Sweeney 1992 and Sweeney and McFarlin 1997, respectively). Such a variation could be accounted for by change of environment and the sample size. The former analysis was carried out with a sample of bank employees (n = 675), whilst the latter was undertaken with a survey of civilian employees of the US federal government (n = 12,670). In another survey of 378 employees (265 male and 113 female), Jepsen and Rodwell (2009) found organizational commitment of male employees to be influenced by their perceived distributive justice (β = .27). Their results for females were insignificant. It is notable that the female sample was comparatively small. Moreover, it comprised occupationally diverse employees, which could have made it even more vulnerable to weak statistical power.

Procedural justice. The construct of procedural justice has been developed through various stages. Initially, it highlighted the significance of procedures, facilitating decision-making on outcomes and distribution of resources, to perceived fairness. Later, structural aspects of procedures were also included in its perimeter, e.g., giving weight-age to stakeholders' voice and letting them contribute to decision making, demonstrating accuracy, and practicing ethics (Leventhal 1980; Leventhal et al. 1980; Greenberg 1986; Holbrook 1999). In early 1990s, procedural justice was proposed to be used as a separate factor. Therefore, it was constructed and measured differently from distributive justice (McFarlin and Sweeney 1992; Sweeney and McFarlin 1993). These scholars maintained that it was related to evaluation of organization-referenced outcomes, e.g., organizational commitment. However, the present review has come across an interesting expansion in the literature that reveals procedural justice to have association with person-referenced ratee reactions as well, e.g., job satisfaction.

Being a distal variable, job satisfaction has been reported to be influenced by procedural justice (see Foley et al. 2005; McFarlin and Sweeney 1992, Sweeney and McFarlin 1997; Cropanzano et al. 2002). The PA literature also suggests a positive association between procedural fairness and certain proximal variables of ratee satisfaction, i.e., satisfaction with ratings, the rating system, the rater, and performance feedback. For example, a field experiment (n = 111 dyads) by Taylor et al. (1995) has suggested procedural fairness to have significant correlation with ratee satisfaction with ratings and the rating system (r = .66 and .52, respectively). Elicker et al.'s (2006) study revealed greater correlation coefficient for the latter (r = .78). In addition, a recent survey of 203 full-time Mexican employees (Selvarajan and Cloninger 2011) has suggested that procedural justice led to satisfaction with the rating system (β = .27), however, the effect size was smaller than that reported by Jawahar (2007), i.e., β = .65.

The relationship between procedural justice and satisfaction with the rater has received notable research attention (e.g., Taylor et al. 1995 reported r = .38). Some researchers (e.g., McFarlin and Sweeney 1992; Sweeney and McFarlin 1997; Cropanzano et al. 2002; Colquitt 2001) have also suggested that the latter regresses the former (β = .23, 34, 41, and .48 respectively). Although Colquitt (2001) pronounced the criterion as leader evaluation, the items used for measurement revealed satisfaction with the rater. In the recent past, Jawahar (2007) suggested that procedural justice might influence ratee satisfaction with performance feedback (β = .23). In a recent study (n = 299 teachers), Tuytens and Devos (2012) substantiated this relationship while teasing apart the criterion into two dimensions, i.e., feedback utility and feedback accuracy (r = .48 and .51, respectively). Regarding ratee satisfaction with reward, McFarlin and Sweeney (1992) have found a significant effect of procedural justice on it (β = .14, p < .01).

Procedural justice has been considered more as an organization-referenced, thus its relationship with organizational commitment has been suggested in both non-contrived and contrived environments (e.g., Brockner et al. 2003, for the former; Holbrook 1999, for the latter). The correlation coefficients reported in these studies are r = .74 and .62, respectively. Heslin and VandeWalle (2011) substantiated these results, however, they teased apart organizational commitment into affective commitment and normative commitment (r = .43 and .39, respectively). The literature also suggests that organizational commitment regresses procedural justice (e.g., Foley et al. 2005; McFarlin and Sweeney 1992; Sweeney and McFarlin 1997; Colquitt 2001).

Interactional justice (interpersonal and informational). Initially, interpersonal treatment came under the caption of procedural justice. However, later, it was constructed as a separate dimension (Kass 2008). As a result, by the addition of this newly dubbed form of justice, i.e., interactional justice, the three-factor model came into existence. In this regard, Kass (2008) sounded a strong contention that it was merely a facet of procedural justice. At that stage, an interesting debate began and the literature agreed upon the distinction between the two models (procedural and interactional). That distinction was based on 'target', where the target of procedural justice was considered to be the 'system,' whereas, that of interactional justice was believed to be the 'agent' (Cropanzano et al. 2002). Thereafter, the four-factor model was conceptualized, which maintained that interactional justice should not be deemed to be merely distinct from procedural justice, but it should also be teased apart into two components, i.e., interpersonal and informational.

Interpersonal justice refers to interpersonal treatment by the person with the authority to enact the procedures. Treating employees politely and with dignity and respect are exemplified as do's, whereas, passing improper remarks and comments is regarded as don'ts. The interpersonal treatment was further represented by the agent-system model (Bies and Moag 1986). Informational justice is considered to be done when the person with authority to enact the procedures, communicates willingly, readily, and candidly with the employees. Moreover, he or she makes sure that the practicability of the procedures is thoroughly explained in a timely manner (Colquitt 2001). Informational justice also facilitates the evaluation of structural aspects of the process (Jawahar 2007), which further helps ratees maintain perceptions of fairness with regard to the agent (rater/supervisor). Drawing from the literature, interactional fairness can be mirrored to interpersonal and informational fairness, for suggesting their associations with ratee reactions.

According to the agent-system model, interpersonal treatment of the agent (the rater/supervisor) may lead to person-referenced (ratee satisfactions with the rater, the performance feedback, and the rating system) and organization-referenced outcomes (organizational commitment). For example, Colquitt (2001) and Jawahar (2007) have suggested that interpersonal and informational fairness may relate to satisfaction with the rater (β = .23 and .50, respectively). Jawahar (2007) also suggested informational justice to have an effect on satisfaction with the performance feedback (β = .61). Moreover, results of three surveys suggest that interactional fairness may relate to ratee satisfaction with the rating system. For example, Elicker et al. (2006) reported a significant correlation between these two constructs (r = .63), whereas Selvarajan and Cloninger (2011) and Cropanzano et al. (2002) have reported interactional justice to have predicted ratee satisfaction with the rating system (β = 0.22 and .77, respectively). In addition, Jepsen and Rodwell (2009) have suggested that informational justice may lead to job satisfaction (males: β = .32 and females: β = .43), whereas interpersonal justice may predict organizational commitment (female: β = .32). The latter was also supported by Barling and Phillips's analysis (1993).

Quantitative criteria and ratee reactions criteria

The quantitative criteria refer to the accuracy and reliability of performance ratings; hence, it aims to alleviate rating errors/biases (Jacobs et al. 1980). Being on the frontier of a PA system, usually raters are held responsible for rating errors, but in fact there are certain other factors that may cause biases. The argument presented by Curtis et al. (2005) seems logical that there are some errors, which a rater commits with a political agenda, but there are many for which ratees' PA system and social factors (relations) should be held responsible. Thus, this review inventories and classifies the threats to accuracy into four groups, i.e., rater-centric, ratee-centric, relation-centric, and system-centric rating errors, to understand their sources and effects.

Rater-centric rating errors. The major influence a rater takes on is of demographic aspects. *Age bias* occurs when raters are influenced by an elder ratee or become sympathetic with a younger one. They do this to safeguard interests of such ratees. Supporting this, a study on 464 supervisor-subordinate dyads (Griffeth and Bedeian 1989) has suggested that younger raters give significantly lower ratings than older raters. However, another study with similar design, i.e., supervisor-subordinate dyads (Shore and Bleicken's 1991) shows that the age bias might not relate solely to older workers, but certain aspects of employee performance.

Gender bias takes place when raters distort true ratings to benefit the similar gender or victimize the opposite gender. Either of them may dissatisfy the affected ratees (Cook 1995; Arvey and Murphy 1998; Reichel and Mehrez 1994). In their study with 60 supervisors generating performance ratings of 220 supervisees, Varma and Stroh (2001) found that after controlling for performance, both male and female supervisors had inflated ratings about ratees of the same gender. However, two scenario-based studies have revealed diverse findings. Using a sample of 292 students, Hall and Hall (1976) found no significant effect of gender on ratings. Conversely, Lee et al. (2009) with a male sample (n = 92) found a significant impact of gender on ratings. Artificial phenomenon can be the major contributor to this contradiction. It is notable that in another study, gender was found to have an interaction effect with age (Griffeth and Bedeian 1989).

Leniency (or strictness) is considered as the backbone of most rating biases. Mainly, due to raters' own temper of mind, they set a tendency of *leniency/strictness bias*. This tendency compels them to use those categories on the rating scale that represent a lenient/strict rating (Bernardin et al. 2009; Murphy and Cleveland 1995; Noe et al. 2003). The tendency of being lenient or strict can be based on many other biases. For example, ratings can be based on the previous performance of the ratee. Hence, the *past performance error* makes a rater lenient or strict while rating the current performance of the ratee (London et al. 2004). Practitioners

pronounce it *critical incident error*. It occurs when raters rely only on some incidents during the appraisal period and disregard the rest. Similarly, raters' selectiveness about observations is found in the *recency effect*. This occurs when raters ratings are based on the recent good or poor performance of ratees (London et al. 2004).

Raters may escalate their performance ratings while getting influenced by the ratees' physical attractiveness (*attractiveness effect)* (Reichel and Mehrez 1994) or future potential (*high-potential error*). Usually, this happens when raters prefer subjective rating (trait-based) to objective rating (task-based) (Murray 1981). Similarly, raters' personal (dis)likes may lead to *interpersonal affect* that brings out inaccurate ratings (Cook 1995). It occurs when the raters rate the liked ratees by recalling their positive work behaviours and vice versa (Wayne and Liden 1995; Cardy and Dobbins 1994; Arvey and Murphy 1998; Lefkowitz 2000; Varma et al. 2005). Empirical studies with varying designs have confirmed the effect of the interpersonal affect on ratings. In an experimental study (n = 66 students), Cardy and Dobbins (1986) investigated the effect of interpersonal affect. They found that raters' ratings were less accurate when scores on their liking had variations than when liking was constant. Confirming this for multisource feedback, a survey elicited 163 downward, 103 upward, and 1027 peer ratings from 433 employees of an insurance company (Antonioni and Park 2001). These results reveal an influence of interpersonal affect in all three sources of the feedback (i.e., downward, upward, and peer).

With regard to culture, Asian raters are considered more prone to interpersonal affect than Western ones. Varma et al. (2005) carried out a cross-cultural study with two samples (the US: n = 190 and India: n = 113) and reported that interpersonal affect had a significant effect on performance ratings in India, as raters inflated the ratings of low performers. In contrast, the US raters could separate their liking for a ratee from actual performance, revealing no interpersonal affect. The results of the US sample are somewhat astonishing, where using a

supervisor-subordinate dyadic sample; Varma and Stroh (2001) have reported a high correlation between interpersonal affect and performance ratings (r = .78). However, results based on the Indian sample are substantiated by another field study in Asia (n = 172 military officers in Singapore), i.e., raters' interpersonal affect predicts leniency (β = .40) (Ng et al. 2011). *Emotional rating error* is another threat to accuracy that resides beside the interpersonal affect. This occurs when raters, being emotionally attached (or detached) to ratees, use a positive (or negative) lens to see everything about them (London et al. 2004). Sometimes, these feelings of affection/hatred can be of personal nature. Recently, in an empirical investigation, Bento et al. (2011) have identified an interesting finding about *stigma bias*. In their study, they investigated raters' perceptions about ratees' obesity and suggested that such perceptions may influence ratings.

Due to some social reasons, raters may demonstrate *avoidance to negative feedback* (Hogan 1987). Using ratings from 667 bank staff by their 101 supervisors, Wilson (2010) reported raters' tendency to make positive comments and reluctance to give negative feedback. Social desirability pressures on supervisors and/or fears of retaliation from subordinates were reported as possible reasons. Furthermore, raters may mislay motivation to rate judiciously when they realize that ratings will affect ratees' promotion, salary or any other benefit, their *low motivation* towards judicious rating comes into play (London et al. 2004). Further, low motivation toward ratings may result in an *escalation bias* (inflated ratings) (Slaughter and Greguras 2008). Tziner et al. (2008) suggest that raters' *discomfort with the rating system* could be another reason behind inflated ratings. *Similarity error or "similar to me" effect* is another behaviour-based threat to accuracy. This error is committed when raters perceive ratees similar to them, and thus give favourable ratings (London et al. 2004). This may happen the other way round when raters perceive ratees to be dissimilar.

The PA literature suggests two levels of (dis)similarity effect, i.e., deep level (behaviour-based) and surface (demographics-based) (Varma and Stroh 2001). This review includes two longitudinal studies with dyadic samples. First (Tepper et al., 2011), investigating the deep-level (dis)similarity suggested that rater perception of relationship conflict and ratee performance mediated the relationship between perceived deep-level dissimilarity and abusive supervision. Second (Wayne and Liden 1995), examining the surface similarity suggested correlation between demographic similarity and supervisor's liking of the subordinate (r = .31), the latter further related to supervisor's ratings of the subordinate's performance (r = .36).

Like demographic variables (age, gender, education level, etc.), psychological variables (self-confidence, self-efficacy, cognitive abilities, anxiety, etc.) also cause variations in ratings about ratees (Landy and Farr 1980; Wood and Marshall 2008). Psychological variables have been noticed to set raters' expectations about ratees or the position they hold. There are certain instances wherein raters compare ratees' actual performance with prior expectations, and when they find a *disconfirmation of expectations*, they deflate ratings. Endorsing this, in a field study of 49 supervisor-subordinate dyads, Hogan (1987) reported that prior expectations of raters about the ratee interact with actual performance to affect ratings (β = .32). The results of this study also revealed that relationships between prior expectations and performance ratings were strongly correlated (r = .28) than actual performance and performance ratings (r = .16).

Until recently, there were five *personality traits* (i.e., extroversion, agreeableness, conscientiousness, neuroticism and openness), which were deemed vital to variations in ratings. For example, in their empirical investigations, Tziner et al. (2002) with a heterogeneous sample of 253 managers in Israel and Randall and Sharples (2012) in an experiment with 230 government employees, found conscientiousness and agreeableness,

respectively, causing variations in ratings. In two more empirical studies using students as participants, Bernardin and colleagues investigated the effects of these two personality traits on ratings about ratees. In their experimental study (n = 111), Bernardin et al. (2000) found that agreeableness and conscientiousness scores were correlated with rating levels, though in different directions (r = .33 and -.37, respectively). These relationships were also confirmed by a further longitudinal laboratory study by Bernardin et al. (2009). This study (n = 126) reported that raters with high agreeableness and low conscientiousness made the most lenient and least accurate ratings. The extant literature has made an addition to personality traits and their effects on ratings. Using an online survey of direct support professionals (n = 269) and the actual ratings by their supervisors (n = 250), Johnson et al. (2011) explored and found honesty-humility as a sixth personality type that uniquely affected the actual ratings (β = .25).

Raters' inability to rate may lead to *logical error* and *proximity error*. The former is the tendency of giving similar ratings for performance areas that seem logically related. The latter is the tendency to rate similarly those performance areas, which are adjacent on the evaluation form (Jacobs et al. 1980). Therefore, cognitive psychologists have drawn more attention towards information processing and retrieval aspects. They maintain that *raters' memory* affects ratings (Woehr 1992). In an experiment with 70 students, Robbins and DeNisi (1993) found correlation between direct recall and ratings (r = .24). Moreover, another experimental study in a laboratory setting (n = 456 professionals in government agency) showed that participants' cognitive ability, practical intelligence, and job knowledge influence ratings about ratees (Pulakos et al. 1996).

Wong and Kwong (2007) argue that *raters' goals* influence their ratings about ratees. They studied harmony, fairness, and motivating goals. Their research was extended by Wang et al. (2010), who carried out two studies to analyze the effects of raters' goals on rating scores about low, medium and high performer ratees. The results of their study 1 (n = 103

students) revealed that raters were found to be inflating their peer ratings, in pursuance of harmony, fairness and motivation goals. As regards to non-peer ratings, study 2 (n = 120 students) revealed that, on the one hand raters deflated ratings about high performers, to demonstrate fairness on the other hand, they inflated ratings about the low performer ratees, to motivate them.

Ratee-centric rating errors. Raters cannot be held responsible on every occasion for errors; ratees also attempt to change raters' view. Ratees may utilize a family of three behaviours, i.e., impression management, ingratiation, and undeserved reputation for the purpose. Wayne and Liden (1995) suggested that ratees' *impression management* behaviour may indirectly affect the performance ratings, i.e., through self-presentation and other-enhancement. Self-presentation becomes a bias when ratees present them by out of proportionally magnifying positives or airbrushing negatives to earn inflated ratings. Other-enhancement is considered a bias when ratees 'butter up' raters to earn favourable ratings.

Ingratiation occurs when a ratee successfully manages to get undue favours from the rater. Ingratiation can be job-focused, supervisor-focused, and self-focused. The job-focused ingratiation refers to administering the credit for job-related achievements, regardless of the fact that the ratee has or even has not contributed to such an achievement. And sometimes ratees attempt to signify their role in the team's accomplishments. The supervisor-focused ingratiation refers to seeking to obtain raters' gratification by extending them favours in personal as well as professional life. The self-focused category of ingratiation reveals ratees' efforts to present them before raters as friendly, polite, sincere, etc. Ratees do this in order to create a soft corner in raters' heart (Cook 1995). The *undeserved reputation bias* appears when ratees manage to establish an undeserved reputation. This is done by developing networks within the organization, public relations, covering their back by not taking part in controversial issues, stealing credit for successes, high turnover to avoid facing appraisal at

every organization, continuously expanding unit or department, reorganization, and getting the benefit of their absence in critical times (Cook 1995).

Relation-centric rating errors. The PA literature also reveals relation-centric threats to accuracy, which are committed by both raters and ratees. *Ethnicity bias* intensifies the circle of relationships. This refers to intervention of racial discrimination instead of actual performance of ratees (Hall and Hall 1976; Cook 1995). Past literature has established that *racial differences* in PA have been found persistently (Arvey and Murphy 1998; Dewberry 2001). Using actual ratings of bank employees, Wilson (2010) found raters to be giving systematically lower ratings to black staff relative to white staff. The results of this study revealed many differences in the specific factors mentioned across ethnic groups. Similarly, in a longitudinal study (n = 3027 trainee lawyers in the UK), Dewberry (2001) reported evidence of racial discrimination by the assessors. He suggested that future research on ethnicity should focus on differences in the individual's life experiences since his or her childhood. Expanding the circle of influence further, raters may also commit *cross-cultural biases* that occur due to the difference between cultural influences on raters and ratees (Bogardus 2004).

When it comes to *dyadic quality and duration*, empirical studies emphasizing leader-member exchange provide evidence of relation-centric biases. Duarte et al. (1994) used data from 261 dyads and six-month records of their telephone company to analyze the effect of dyadic quality on ratings. They found that, both in the short and the long run, in high-quality leader-member exchange relationships, employee performance was rated high. This was apart from objective ratings about them. The ratings of employees in low-quality leader-member exchange relationships in the short-run were consistent with the objective ratings about them. However, these were high in the long run, apart from their objective ratings. They also found that correlations among leader-member exchange relationship quality, and task and relationship performance ratings were positively significant (r = .26 and .30, respectively).

Tepper et al. (2006) carried out two studies (n = 347) in which managers gave more favourable ratings about ratees with high leader-member exchange even for resistant ratees. However, ratings were higher for those ratees who resisted by negotiating than those who resisted by refusing. In another empirical study, Varma and Stroh (2001) found a positive correlation between dyadic relationship and ratings (r = .77). Sometimes, the dyadic relationships are established for political motives. Therefore, a *political culture* in which the appraisal process operates may also aggravate in-group and out-group situations resulting in favourable and unfavourable ratings, respectively (Wood and Marshall 2008). Usually it happens when team performance is replaced with a political agenda. The *political considerations* start capitalizing the PA system and the rater becomes over lenient or over strict, to extend benefits or to victimize the ratee (Cook 1995).

Relatedness within and between-ratees may also affect ratings, e.g., *halo and horn effects and stereotyping.* The halo error occurs when raters find a positive aspect of performance and then continue rating positively the remaining aspects of ratees' performance. Conversely, horn error leads to keep on rating negatively if one aspect is found to be so (Arvey and Murphy 1998; Murphy and Cleveland 1995; Noe et al. 2003; Bogardus 2004). In their experimental study (n = 170 students), Becker and Cardy (1986) found halo effect on accuracy and even statistical control of its influence could not improve the rating validity. Jackson (1996) carried out two studies, one using 100 students and in the other 323 trained interviewers rated eight video-taped interviewees in a laboratory setting. Both studies revealed that the maximum accuracy within a task was not necessarily at 'zero invalid halo.' *Stereotyping* is a tendency to generalize across groups and ignore individual differences (Bogardus 2004). It is more likely to happen when team performance is appraised.

System-centric rating errors. Findley et al. (2000) grouped certain PA aspects such as appraisal policies, procedures, and support provided by the organization, and pronounced them *appraisal system facets*. Their survey (*n* = 199 school teachers) revealed that appraisal system facets explained significant incremental variance in perceived rating accuracy. This was more than that was explained by the appraisal process facets (refer to observation, feedback/voice, and planning) ($\Delta R^2 = .04$). This shows significant impact of PA policies and procedures on rating errors. Substantiating this, Jawahar (2005) investigated the impact of *system factors* (also known as situational influences) on rating accuracy. His experimental study 1 (*n* = 186) and study 2 (*n* = 108 HR managers) revealed that some system factors (e.g. quality of equipment, availability of resources, difficulty of sales territory) are beyond the control of individual employees. Therefore, sometimes the PA system compels raters to be lenient in order to offset the anticipated effect of system factors on ratee performance. The results of these two studies indicated that both junior and senior raters altered ratings depending on the situational conditions under which ratees worked.

Some PA systems *exempt* certain employees from being evaluated. For example, using a large German sample (*n* = 7,598), Grund and Sliwka (2009) found that the performance of older employees, women, and employees with very high or very low responsibilities was often assessed less. Based on ratings generated by students from videotapes, two laboratory studies have suggested that *rating format* may cause system-centric errors. One of these was a cross-sectional study (*n* = 180) that revealed that behavioural anchors caused biased ratings, as raters focused only on those aspects of performance, which were anchored in the scale, regardless of their representativeness of ratees' actual performance (Murphy and Constans 1987). The other was a longitudinal study (*n* = 57) that revealed that consistently average ratings were less accurate than descending and ascending ratings. It was also found that the

overall ratings by the subjects were more accurate than an average of ratings made on each concluding exercise (Karl and Wexley 1989).

Available tools for descriptive analysis of PA results may also reveal errors such as *central tendency and range restriction,* and *negative and positive skew.* The former is a tendency of using rating scales representing average rating (Murphy and Cleveland 1995; Noe et al. 2003; Grote 2002; Bogardus 2004). The latter occurs when raters stick to extreme ratings on either side of the rating scale (Grote 2002). Apart from analysis, the system in which raters perform sometimes compels them to commit a *contrast error.* This is normally caused by holding a comparison between ratees instead of comparing their performance with the objective standards (Latham et al. 2008; Noe et al. 2003; Bogardus 2004). If such comparison is held within-individual, then opportunities to come across the *inappropriate substitutes for performance* become evident. This error takes place when the organization sets an inadequate criterion to determine performance (De Cenzo and Robbins 1996) and, ultimately, raters rate hypothetically (global observations).

The existing literature presents a caution that political considerations sometimes seem to intermingle with *inflationary pressures.* It also coerces raters to think that mere high ratings are not sufficient for certain ratees' promotion but the highest ratings (De Cenzo and Robbins 1996). Therefore, *purposes and uses of PA* compel raters to give the desirable PA results leading to biased ratings (Tziner et al. 2002; Farh et al. 1991). Organizations can avoid biases by holding raters accountable to the PA system, as accountability relates to rating accuracy (r = .34) (Wood and Marshall 2008). This was confirmed by a scenario-based study (Curtis et al. 2005) in which 123 students rated ratees more leniently when they were accountable to the ratee than the experimenter. However, participants rated ratees less leniently when they were accountable to both (the ratee and the experimenter) than ratees only (downwardly

accountable). In contrast, participants rated ratees more leniently when they were accountable to both (the ratee and the experimenter) than the experimenter only (upwardly accountable).

In another experimental study (*n* = 197 students), Mero et al. (2007) found that participants rated more accurately when they knew that they were accountable to 'high-ups' than when they were either accountable to ratees or had no one to account to. This might be because participants pre-empted the self-criticism and relied on more complex judgment strategies when they were answerable to high-ups. Thus, their pre-emption-based complex information processing led them to more defensible ratings, which turned out to be more accurate.

Having discussed categories of rating errors in detail, we have brought this section to a stage where, according to literature (e.g. Keeping and Levy 2000; Levy and Williams 2004; Roberson and Stewart 2006), it is suggested that rating errors limit EPA. Thus, the rater-centric, ratee-centric, relation-centric, and system-centric threat to accuracy may lead to perceived detriments to EPA. However, the relative importance of each is likely to vary.

Relationships among measurement criteria

Merely accomplishing some PA purposes, demonstrating fairness with regard to selected aspects of justice, or neutralizing effects of certain rating biases are not sufficient to demonstrate EPA, unless these measurement criteria are integrated in order to strengthen the PA system. Therefore, this section aims to identify linkages amongst the measurement criteria of EPA.

Utilization and qualitative criteria . The chances of unfairness are more likely to occur when PA is used for administrative purposes. This is because of its vital role in organizational decision making, especially when the ultimate beneficiaries of these decisions are employees. Organizations consider results of administrative PA helpful in pursuing personal agenda and/or satisfy political motives. For example, precipitating certain employees to victimize

them, or casting certain employees in the limelight to pave path for their promotion. Since such decisions directly affect the outcomes (pay, promotion, etc.), the literature suggests that administrative PA is perceived to be more prone to unfairness (distributive) than PA used for other purposes. The developmental PA is considered to have at least a neutral effect, because it is likely to have a mild effect on outcome-related organizational decisions (Selvarajan and Cloninger 2011).

Selvarajan and Cloninger (2011) further argue that employees' perceptions of distributive unfairness may prompt their perceptions about procedural unfairness, maintaining that procedures that reveal unfair outcomes must themselves be unfair. Once again, developmental PA may interact differently with procedural fairness (Jawahar 2007). Overall, this argument is in line with empirical findings. For example, an experiment (n = 195) by Bettenhausen and Fedor (1997) revealed that developmental PA resulted in more positive outcomes than administrative PA. They also found that administrative PA resulted in more negative outcomes than the developmental PA. Thus, developmental PA may have more positive relationship with perceived distributive and procedural fairness than the administrative PA.

Utilization and quantitative criteria. Empirical literature suggests that administrative and developmental PA may relate to rating accuracy. For example, a simulation-based laboratory study (n = 130) of Zedeck and Cascio (1982) has revealed that administrative and developmental PA explained more variation in rating accuracy than other variables, e.g., rater training. In addition, some empirical studies lay the foundation for establishing relationships between administrative and developmental PA, and system and rater-centric rating errors.

Based on an analysis of two datasets, one for the developmental purposes (ratings of 193 raters) and the other for the administrative purposes (ratings about 223 ratees), Harris et al. (1995) found that ratings for the administrative purposes were more biased (lenient) than for the developmental purposes. Moreover, their results revealed administrative purposes to have

a significant relationship with ratee seniority (r = .18), but developmental ratings did not have a significant relationship (r = .00). This is supported by results of a quasi-experiment (n = 65 students) by Farh et al. (1991) that revealed a propensity to contain greater halo and leniency when ratings were conducted for administrative purposes than for developmental purposes.

Curtis et al. (2005) found that in the administrative purpose condition, raters rated most leniently when they were only accountable to the ratees. Conversely, in the developmental purpose condition, raters rated least leniently when they were accountable to the experimenter. Most of the empirical investigations have revealed that administrative PA leans more towards rating errors than developmental PA. Therefore, to neutralize this effect, Selvarajan and Cloninger (2011) concluded that both administrative and developmental PAs are perceived to be more accurate than administrative PA alone. Thus, when used simultaneously, administrative and developmental PAs may explain a positive variation in system-centric rating errors. However, on teasing apart PA purposes, administrative PA would be more likely to explain variation in system-centric rating errors than developmental PA.

The PA literature maintains that certain PA purposes may cause rater-centric rating errors, e.g., Tziner et al. (2002) and Tziner et al. (2008) suggest that developmental PA may relate positively to rater's confidence in PA (r = .59 and r = .39, respectively). However, Tziner et al. (2008) also suggested that administrative PA may relate inversely to raters' confidence in PA (r = -.28). These results indicate that administrative PA is more prone to rater-centric errors than developmental PA. However, there is a caution. Based on only one aspect (i.e., rater's confidence), the possibility of rater-centric rating errors triggered by the developmental PA cannot be eliminated. Therefore, it can be expected that both administrative and developmental PAs may explain variations in rater-centric rating errors. However, on teasing apart PA purposes, administrative PA may explain more variations in rater-centric rating errors than the developmental PA.

Quantitative and qualitative criteria. Empirical literature suggests that ratees' perceived fairness might lead to perceived rating accuracy. Taylor et al. (1995) found ratees' perceived procedural fairness to be correlated with rating accuracy (r = .73). Adding to this, a survey by Elicker et al. (2006) reported that distributive, procedural and interactional justice are positively correlated with perceived accuracy (r = .81, .80, .65, respectively). Skarlicki and Folger (1997) further confirmed this by using different criteria. They found distributive, procedural and interactional justice to have a significant negative effect on ratees' organizational retaliation behaviour (β = -3.73, -2.38, -5.23 respectively). These results indicate that if ratees perceive unfairness, they may try to establish equity in their own way, e.g., showing retaliation, being counterproductive, or manipulating ratings. Thus, the higher the perceived fairness is, the lower the ratee-centric biases will be and vice versa.

Conclusions

This paper offers a two-pronged conclusion. The one part is general, about the research trends in the sub-field of EPA, and the other is specific, about the ratee reactions-based integrated framework of EPA. We have monitored four aspects of research trends in EPA literature that can be helpful for upcoming empirical research in this body of knowledge.

First, empirical studies on EPA have used a variety of research designs, e.g., cross-sectional and longitudinal, surveys and experiments or quasi experiments. With regard to study setting, of the 104 empirical studies, 64% were carried out in real actors (e.g., employees) and 27% were in artificial settings (e.g., with students). Among the latter, most were scenario-based experimental studies with effective research designs. The remaining nine percent of the studies used combination of the above two (contrived and non-contrived).

Second, EPA literature lacks a holistic view, as it is scattered in pieces. Therefore, a segment of literature considers PA a mere activity, instead of a system. Also, the effectiveness of this system is not discussed as such. This resulted in a patchy attention being paid to the

EPA criteria. In the 1980s, the quantitative criteria outweighed other criteria. However, from the early 1990s, qualitative criteria started to attract the attention of the EPA researchers, and now its coverage in the literature is almost equal to that of quantitative criteria. Thus far, the utilization criteria could manage less than a moderate appearance in the EPA research, during the last three decades.

Third, where attention being paid to the measurement criteria has been uneven, within each measurement criteria certain subordinate criteria have also been ignored. For example, regarding utilization criteria, the major focus has been on administrative purposes followed by the developmental one. A scarcity is also found with regard to the strategic and role-definition. Similarly, with regard to the quantitative criteria, the emphasis has been on rater-centric errors, followed by the system-centric one, whereas ratee-centric errors have been discussed rarely. Moreover, this paper has discussed over 40 factors as direct or indirect determinants of rating bias. Many of them so far have not been part of robust empirical investigations.

Lastly, there is a limitation of the PA literature that it largely represents the US-oriented models, approaches and theories. Since performance management is a social phenomenon, Bititci et al. (2012) raise a valid question, i.e., 'do these theoretical rationales fit globally?' On the one hand, this question challenges the external validation of the existing evidence for diverse countries and cultures. On the other hand, this draws attention towards the fact that the PA body of knowledge has been deprived of indigenous wisdom from the perspective of geographical considerations. To the best of our understanding, cross-cultural studies can offset the deficiency in geographical representation, but to only a small extent. The PA literature needs to represent those countries and cultures that represent more than two-third of the world's population, and also the emerging markets due to their growing economic dominance and increasing interest of foreign investors in them. To start with, at least the Eastern

researchers may be encouraged to replicate the models and theories propounded in the West and where possible develop their own context specific approaches to PA. This would serve a two-fold purpose: one, it would help manage the representation of the developing part of the world; two, it would help demonstrate the external validation of research models geographically and also develop context relevant models. We believe that the contradictory results would refine the existing theories or give birth to new ones.

In addition to the above-mentioned general conclusions, this paper also offers some specific conclusions.

The first objective of this analysis was to present the expanded view of measurement criteria of EPA. This paper highlighted notable refinements and expansions about utilization, qualitative, and quantitative criteria. Utilization criteria: the long-standing view of PA that has focused more on administrative and little on developmental purposes had restricted this practice to personnel, evaluation, accountability, judgement, and development functions. The addition of strategic and role-definition purposes has added more theoretical anchors and widened the scope of EPA towards more human resource functions, e.g., feedback and goal-orientation. On the face of the current PA practice and research, the latter are rapidly gaining prominence, whereas the former are becoming secondary, with the exception of development function.

Qualitative criteria: empirical literature has refined certain relationships by broadening the scope, e.g., under the two-factor model, distributive justice was thought to have affected only person-referenced outcomes. However, under the three and four-factor models, organization-referenced outcomes was added as a criterion (e.g., Foley et al. 2005; Jepsen and Rodwell 2009; Heslin and VandeWalle 2011). Quantitative criteria: traditionally, raters were held responsible for rating errors. However, this paper has mounted sufficient evidence to justify the categorization of 40 factors (errors/biases) into four groups, i.e., rater-centric, ratee-

centric, relation-centric, and system-centric errors. Expectedly, this categorization may lead PA researchers and practitioners to put directed efforts into minimizing bias and increasing accuracy.

The second objective was to identify relationships between measurement criteria and their respective outcomes. This paper provided empirical confirmations based on a priori theory or models that have suggested nomological networks for above-mentioned relationships, which are all set for empirical testing.

The final objective was to seek an integrated framework of EPA. Although the PA literature contains sufficient support for developing a ratee reactions-based integrated framework of EPA, some cautions must be borne in mind before putting this into practice. First, an uneven use of PA purposes may lead to injustice, e.g., administrative PA is more prone to distributive and procedural injustice than developmental PA. Second, an uneven use of PA purposes may also lead to rating errors, e.g., administrative PA may lead to system-centric and rater-centric rating errors more than developmental PA. Finally, any slackness in qualitative criteria can dismantle the quantitative criteria, as justice dimensions of the four-factor model are inversely related to ratee-centric errors. Thus, integration among measurement criteria of EPA is simple yet complex.

Despite certain limitations, e.g., only 'quality' journals, and a limited number of articles were reviewed, this paper made an attempt to structure the diverse literature on EPA. The authors believe that the outcome of this analysis would provide a valuable venture to researchers, fuelling more relevant and focused research on PA systems. It is expected that future empirical research on EPA would fill the research gaps highlighted in this review such as undertaking the expanded view of utilization criteria, classification of quantitative criteria, and their relationship with Greenberg's taxonomy of PA outcomes criteria. Also, by filling the highlighted gaps in the existing literature, future empirical evidence on EPA framework

would inform professionals about the required focal point in their endeavours, i.e., ratee reactions-based view, for designing an effective PA system.

For example, we suggest that on completion of a PA exercise, organizations may collect soft data (e.g., on employee perceptions about the four criteria) and analyze it using our proposed integrated framework. This will help them identify the felt needs (of their employees, e.g., negative ratee reactions such as a low level of satisfaction and commitment etc.), indicate high felt needs, and vice versa. Once employees' felt needs are identified, organizations can plan to manage and meet them, because meeting such needs will help the employees to know more about things such as their organization's view of their performance; as to how well they perform; the ways they can improve their performance; their strengths and weaknesses; their future role; and how to devise a skill supply strategy for their future role. These would prepare them for pursuing their own and the organization's goals.

References

Abu-Doleh, J. and Weir, D. (2007). Dimensions of performance appraisal systems in Jordanian private and public organizations. *The International Journal of Human Resource Management,* **18,** pp. 75-84.

Aguinis, H. (2009). An expanded view of performance management. In Smither, J.W. and London, M. (eds.), *Performance Management: Putting Research Into Action.* San Francisco: Jossey-Bass, pp. 1-44.

Antonioni, D. and Park, H. (2001). The relationship between rater affect and three sources of 360-degree feedback ratings. *Journal of Management,* **27,** pp. 479-495.

Arvey, R.D. and Murphy, K.R. (1998). Performance evaluation in work settings. *Annual Review of Psychology,* **49,** pp. 141-168.

Barling, J. and Phillips, M. (1993). Interactional, formal, and distributive justice in the workplace: an exploratory study. *Journal of Psychology,* **127,** pp. 649-656.

Becker, B.E. and Cardy, R.L. (1986). Influence of halo error on appraisal effectiveness: a conceptual and empirical reconsideration. *Journal of Applied Psychology,* **71,** pp. 662-671.

Beer, M. (1981). Performance appraisal: dilemmas and possibilities. *Organizational Dynamics,* **9,** pp. 24-36.

Bento, R.F., White, L.F. and Zacur, S.R. (2011). The stigma of obesity and discrimination in performance appraisal: a theoretical model. *The International Journal of Human Resource Management,* **iFirst 0,** pp. 1-29.

Bernardin, H.J., Cooke, D.K. and Villanova, P. (2000). Conscientiousness and agreeableness as predictors of rating leniency. *Journal of Applied Psychology,* **85,** pp. 232-234.

Bernardin, H.J., Tyler, C.L. and Villanova, P. (2009). Rating level and accuracy as a function of rater personality. *International Journal of Selection and Assessment,* **17,** pp. 300-310.

Bettenhausen, K.L. and Fedor, D.B. (1997). Peer and upward appraisals: a comparison of their benefits and problems. *Group &. Organization Management,* **22,** pp. 236-263.

Bies, R.J., & Moag, J.F. (1986). Interactional justice: Communication criteria of fairness. In Bies, R.J., Moag, J.F., Lewicki, R.J., Sheppard, B.H., and Bazerman, M.H. (eds.), *Research on Negotiations in Organizations* (Vol. 1). Greenwich, CT: JAI, pp. 43-55.

Bies, R.J. and Shapiro, D.L. (1987). Interactional fairness judgments: the influence of causal accounts. *Social Justice Research,* **1,** pp. 199-218.

Bititci, U., Garengo, P., Dörfler, V. and Nudurupati, S. (2012). Performance measurement: challenges for tomorrow. *International Journal of Management Reviews,* **14,** pp. 305-327.

Bogardus, A.M. (2004). *PHR/SPHR: Professional in Human Resources Certification Study Guide*. New York: Sybex.

Boswell, W.R. and Boudreau, J.W. (2002). Separating the developmental and evaluative performance appraisal uses. *Journal of Business and Psychology,* **16,** pp. 391-412.

Brockner, J., Heuer, L., Magner, N., Folger, R., Umphress, E., Vandenbos, K., Vermunt, R., Magner, M. and Siegel, P. (2003). High procedural fairness heightens the effect of outcome favorability on self-evaluations: An attributional analysis. *Organizational Behavior and Human Decision Processes,* **91,** pp. 51-68.

Campbell, D.J. and Lee, C. (1988). Self-appraisal in performance evaluation: development versus evaluation. *Academy of Management Review,* **13,** pp. 302-314.

Cardy, R.L. and Dobbins, G.H. (1986). Affect and appraisal accuracy: liking as an integral dimension in evaluating performance. *Journal of Applied Psychology,* **71,** pp. 672-678.

Cardy, R.L., Dobbins, G.H. (1994). *Performance Appraisal: Alternative Perspectives.* South-Western OH: Cincinnati.

Chiang, F.F.T. and Birtch, T.A. (2010). Appraising performance across borders: an empirical examination of the purposes and practices of performance appraisal in a multi-country context. *Journal of Management Studies,* **47,** pp. 1365-1393.

Claus, L. and Briscoe, D. (2009). Employee performance management across borders: a review of relevant academic literature. *International Journal of Management Reviews,* **11,** pp. 175-196.

Cleveland, J.N., Murphy, K.R. and Williams, R.E. (1989). Multiple uses of performance appraisal: prevalence and correlates. *Journal of Applied Psychology,* **74,** pp. 130-135.

Colquitt, J.A. (2001). On the dimensionality of organizational justice: a construct validation of a measure. *Journal of Applied Psychology,* **86,** pp. 386-400.

Colquitt, J.A. and Rodell, J.B. (2011). Justice, trust, and trustworthiness: a longitudinal analysis integrating three theoretical perspectives. *Academy of Management Journal,* **54,** pp. 1183–1206.

Cook, M. (1995). Performance appraisal and true performance. *Journal of Managerial Psychology,* **10,** pp. 3-7.

Cropanzano, R., Prehar, C.A. and Chen, P.Y. (2002). Using social exchange theory to distinguish procedural from interactional justice. *Group & Organization Management,* **27,** pp. 324-351.

Curtis, A.B., Harvey, R.D. and Ravden, D. (2005). Sources of political distortions in performance appraisals: appraisal purpose and rater accountability. *Group & Organization Management,* **30,** pp. 42-60.

Dahling, J.J., Chau, S.L. and O'malley, A. (2012). Correlates and consequences of feedback orientation in organizations. *Journal of Management,* **38,** pp. 531-546.

De Cenzo, D.A. and Robbins, S.P. (1996). *Human Resources Management (5th ed.).* New York: John Wiley and Sons.

de Menezes, L.M. and Kelliher, C. (2011). Flexible working and performance: a systematic review of the evidence for a business case. *International Journal of Management Reviews,* **13,** pp. 452-474.

Dewberry, C. (2001). Performance disparities between whites and ethnic minorities: real differences or assessment bias? *Journal of Occupational and Organizational Psychology,* **74,** pp. 659-673.

Dipboye, R.L. (1985). Some neglected variables in research on discrimination in appraisals. *Academy of Management Review,* **10,** pp. 116-127.

Dorfman, P.W., Stephan, W.G. and Loveland, J. (1986). Performance appraisal behaviors: supervisor perceptions and subordinate reactions. *Personnel Psychology,* **39,** pp. 579-597.

Duarte, N.T., Goodson, J.R. and Klich, N.R. (1994). Effects of dyadic quality and duration on performance appraisal. *Academy of Management Journal,* **37,** pp. 499-521.

Elicker, J.D., Levy, P.E. and Hall, R.J. (2006). The role of leader-member exchange in the performance appraisal process. *Journal of Management,* **32,** pp. 531-551.

Farh, J.L., Cannellajr., A.A. and Bedeian, A.G. (1991). Peer ratings: the impact of purpose on rating quality and user acceptance. *Group & Organization Studies,* **16,** pp. 367-386.

Findley, H.M., Giles, W.F. and Mossholder, K.W. (2000). Performance appraisal process and system facets: relationships with contextual performance. *Journal of Applied Psychology,* **85,** pp. 634-640.

Fletcher, C. (1995). New directions for performance appraisal: some findings and observations. *International Journal of Selection and Assessment,* **3,** pp. 191-196.

Fletcher, C. (2001). Performance appraisal and management: the developing research agenda. *Journal of Occupational and Organizational Psychology,* **74,** pp. 473-487.

Foley, S., Hang-Yue, N. and Wong, A. (2005). Perceptions of discrimination and justice: are there gender differences in outcomes? *Group &. Organization Management,* **30,** pp. 421-450.

Folger, R., Rosenfield, D., Grove, J. and Corkran, L. (1979). Effects of "voice" and peer opinions on responses to inequity. *Journal of Personality and Social Psychology,* **37,** pp. 2253-2261.

Giles, W.F. and Mossholder, K.W. (1990). Employee reactions to contextual and session components of performance appraisal. *Journal or Applied Psychology,* **75,** pp. 371-377.

Giles, W.F., Findley, H.M. and Feild, H.S. (1997). Procedural fairness in performance appraisal: beyond the review session. *Journal of Business and Psychology,* **11,** pp. 493-506.

Greenberg, J. (1986). Determinants of perceived fairness of performance evaluations. *Journal of Applied Psychology,* **71,** pp. 340-342.

Greenberg, J. (1990). Organizational justice: yesterday, today, and tomorrow. *Journal of Management,* **16,** pp. 399-432.

Griffeth, R.W. and Bedeian, A.G. (1989). Employee performance evaluations: effects of ratee age, rater age, and ratee gender. *Journal of Organizational Behavior,* **10,** pp. 83-90.

Grote, D. (2002). *The Performance Appraisal Question and Answer Book: A Survival Guide for Managers.* New York: AMACOM.

Grund, C. and Sliwka, D. (2009). The anatomy of performance appraisals in Germany. *The International Journal of Human Resource Management,* **20,** pp. 2049-2065.

Haines III, V.Y. and St-Onge, S. (2012). Performance management effectiveness: practices or context? *The International Journal of Human Resource Management,* **23,** pp. 1158-1175.

Hall, F.S. and Hall, D.T. (1976). Effects of job incumbents' race and sex on evaluations of managerial performance. *Academy of Management Journal,* **19,** pp. 476-481.

Harder, J.W. (1992). Play for pay: effects of inequity in a pay-for-performance context. *Administrative Science Quarterly,* **37,** pp. 321-335.

Harris, M.M., Smith, D.E. and Champagne, D. (1995). A field study of performance appraisal purpose: research- versus administrative-based ratings. *Personnel Psychology,* **48,** pp. 151-160.

Hedge, J.W. and Teachout, M.S. (2000). Exploring the concept of acceptability as a criterion for evaluating performance measures. *Group & Organization Management,* **25,** pp. 22-44.

Heslin, P.A. and Vandewalle, D. (2011). Performance appraisal procedural justice: the role of a manager's implicit person theory. *Journal of Management,* **37,** pp. 1694-1718.

Hogan, E.A. (1987). Effects of prior expectations on performance ratings: a longitudinal study. *Academy of Management Journal,* **30,** pp. 354-368.

Holbrook, Jr., R.L. (1999). Managing reactions to performance appraisal: the influence of multiple justice mechanisms. *Social Justice Research,* **12,** pp. 205-221.

Jackson, C. (1996). An individual differences approach to the halo-accuracy paradox. *Personality and Individual Differences,* **21,** pp. 947-957.

Jacobs, R., Kafry, D. and Zedeck, S. (1980). Expectations of behaviorally anchored rating scales. *Personnel Psychology,* **33,** pp. 595-640.

Jawahar, I.M. (2001). Attitudes, self-monitoring, and appraisal behaviors. *Journal of Applied Psychology,* **86,** pp. 875-883.

Jawahar, I.M. (2005). Do raters consider the influence of situational factors on observed performance when evaluating performance? Evidence from three experiments. *Group & Organization Management,* **30,** pp. 6-41.

Jawahar, I.M. (2007). The influence of perceptions of fairness on performance appraisal reactions. *Journal of Labor Research,* **28,** pp. 735-754.

Jawahar, I.M. and Williams, C.R. (1997). Where all the children are above average: the performance appraisal purpose effect. *Personnel Psychology,* **50,** pp. 905-926.

Jepsen, D.M. and Rodwell, J.J. (2009). Justice in the workplace: the centrality of social versus judgmental predictors of performance varies by gender. *The International Journal of Human Resource Management,* **20,** pp. 2066-2083.

Johnson, M.K., Rowatt, W.C. and Petrini, L. (2011). A new trait on the market: honesty–humility as a unique predictor of job performance ratings. *Personality and Individual Differences,* **50,** pp. 857-862.

Karl, K.A. and Wexley, K.N. (1989). Patterns of performance and rating frequency: influence on the assessment of performance. *Journal of Management,* **15,** pp. 5-20.

Kass, E. (2008). Interactional justice, negotiator outcome satisfaction, and desire for future negotiations: R-E-S-P-E-C-T at the negotiating table. *International Journal of Conflict Management,* **19,** pp. 319-338.

Keeping, L.M. and Levy, P.E. (2000). Performance appraisal reactions: measurement, modeling, and method bias. *Journal of Applied Psychology,* **85,** pp. 708-723.

Korsgaard, M.A. and Roberson, L. (1995). Procedural justice in performance evaluation: the role of instrumental and non-instrumental voice in performance appraisal discussions. *Journal of Management,* **21,** pp. 657-669.

Kudisch, J.D., Fortunato, V.J. and Smith, A.F.R. (2006). Contextual and individual difference factors predicting individuals' desire to provide upward feedback. *Group & Organization Management,* **31,** pp. 503-529.

Kuvaas, B. (2006). Performance appraisal satisfaction and employee outcomes: mediating and moderating roles of work motivation. *The International Journal of Human Resource Management,* **17,** pp. 504-522.

Landy, F.J. and Farr, J.L. (1980). Performance rating. *Psychological Bulletin,* **87,** pp. 72-107.

Latham, G.P., Almost, J., Mann, S. and Moore, C. (2005). New developments in performance management. *Organizational Dynamics,* **34,** pp. 77-87.

Latham, G.P., Budworth, M.H., Yanar, B. and Whyte, G. (2008). The influence of a manager's own performance appraisal on the evaluation of others. *International Journal of Selection and Assessment,* **16,** pp. 220-228.

Lawler III, E.E. (2003). Reward practices and performance management system effectiveness. *Organizational Dynamics,* **32,** pp. 396-404.

Lawler III, E.E., Mohrman, Jr., A.M. and Resnick, S.M. (1984). Performance appraisal revisited. *Organizational Dynamics,* **13,** pp. 20-35.

Lee, C. (1985). Increasing performance appraisal effectiveness: matching task types, appraisal process, and rater training. *Academy of Management Review,* **10,** pp. 322-331.

Lee, J.A., Welbourne, J.L., Hoke, W.A. and Beggs, J. (2009). Examining the interaction among likelihood to sexually harass, ratee attractiveness, and job performance. *Journal of Management,* **35,** pp. 445-461.

Lefkowitz, J. (2000). The role of interpersonal affective regard in supervisory performance ratings: a literature review and proposed causal model. *Journal of Occupational and Organizational Psychology,* **73,** pp. 67-85.

Leventhal, G. S. (1980). What should be done with equity theory? New approaches to the study of fairness in social relationships. In Gergen, K., Greenberg, M. and Willis, R. (eds.), *Social exchange: Advances in theory and research.* New York: Plenum Press, pp. 27-55.

Leventhal, G. S., Karuza, J., & Fry, W. R. (1980). Beyond fairness: A theory of allocation preferences. In Mikula G. (eds.), *Justice and social interaction*. New York: Springer-Verlag, pp. 167-218.

Levy, P.E. and Williams, J.R. (2004). The social context of performance appraisal: a review and framework for the future. *Journal of Management,* **30,** pp. 881-905.

Linna, A., Elovainio, M., Vandenbos, K., Kivimäki, M., Pentti, J. and Vahtera, J. (2012). Can usefulness of performance appraisal interviews change organizational justice perceptions? A 4-year longitudinal study among public sector employees. *The International Journal of Human Resource Management,* **23,** pp. 1360–1375.

London, M., Mone, E.M. and Scot, J.C. (2004). Performance management and assessment: methods for improved rater accuracy and employee goal setting. *Human Resource Management,* **43,** pp. 319-336.

Martocchio, J.J. and Judge, T.A. (1995). When we don't see eye to eye: discrepancies between supervisors and subordinates in absence disciplinary decisions. *Journal of Management,* **21,** pp. 251-278.

Mcdowall, A. and Fletcher, C. (2004). Employee development: an organizational justice perspective. *Personnel Review,* **33,** pp. 8-29.

Mcfarlin, D.B. and Sweeney, P.D. (1992). Distributive and procedural justice as predictors of satisfaction with personal and organizational outcomes. *Academy of Management Journal,* **35,** pp. 626-637.

Mero, N.P., Guidice, R.M. and Brownlee, A.L. (2007). Accountability in a performance appraisal context: the effect of audience and form of accounting on rater response and behavior. *Journal of Management,* **33,** pp. 223-252.

Miller, J.S. and Cardy, R.L. (2000). Self-monitoring and performance appraisal: rating outcomes in project teams. *Journal of Organizational Behavior,* **21,** pp. 609-626.

Milliman, J., Nason, S., Zhu, C. and Decieri, H. (2002). An exploratory assessment of the purposes of performance appraisals in North and Central America and the Pacific Rim. *Human Resource Management,* **41,** pp. 87-102.

Mount, M.K. (1984). Satisfaction with a performance appraisal system and appraisal discussion *Journal of Occupational Behaviour,* **5,** pp. 271-279.

Murphy, K.R. and Constans, J.I. (1987). Behavioral anchors as a source of bias in rating. *Journal of Applied Psychology,* **72,** pp. 573-577.

Murphy, K.R. and Cleveland, J.N. (1995). *Understanding Performance Appraisal: Social, Organizational and Goal-Based Perspectives.* Thousand Oaks: Sage.

Murray, R.S. (1981). Managerial perceptions of two appraisal systems. *California Management Review,* **XXIII,** pp. 92-96.

Ng, K.-Y., Koh, C., Ang, S., Kennedy, J.C. and Chan, K.-Y. (2011). Rating leniency and halo in multisource feedback ratings: testing cultural assumptions of power distance and individualism-collectivism. *Journal of Applied Psychology* **96,** pp. 1033–1044.

Noe, R.A., Hollenbeck, J.R., Gerhart, B., and Wright, P.M. (2003). *Human Resource Management*. London: McGraw Hill.

Nurse, L. (2005). Performance appraisal, employee development and organizational justice: exploring the linkages. *The International Journal of Human Resource Management,* **16,** pp. 1176-1194.

Pichler, S. (2012). The social context of performance appraisal and appraisal reactions: a meta-analysis. *Human Resource Management,* **51,** pp. 709-732.

Pulakos, E.D., Schmitt, N. and Chan, D. (1996). Models of job performance ratings: an examination of ratee race, ratee gender, and rater level effects. *Human Performance,* **9,** pp. 103-119.

Randall, R. and Sharples, D. (2012). The impact of rater agreeableness and rating context on the evaluation of poor performance. *Journal of Occupational and Organizational Psychology,* **85,** pp. 42-59.

Reichel, A. and Mehrez, A. (1994). Employee selection and performance evaluation biases and organizational efficiency: a mathematical modeling attempt. *Journal of Management Inquiry,* **3,** pp. 85-95.

Renn, R.W. and Fedor, D.B. (2001). Development and field test of a feedback seeking, selfefficacy, and goal setting model of work performance. *Journal of Management,* **27,** pp. 563-583.

Robbins, T.L. and DeNisi, A.S. (1993). Moderators of sex bias in the performance appraisal process: a cognitive analysis. *Journal of Management,* **19,** pp. 113-126.

Roberson, Q.M. and Stewart, M.M. (2006). Understanding the motivational effects of procedural and informational justice in feedback processes. *British Journal of Psychology,* **97,** pp. 281-298.

Roch, S.G. (2006). Discussion and consensus in rater groups: implications for behavioral and rating accuracy. *Human Performance,* **19,** pp. 91-115.

Roch, S.G., Sternburgh, A.M. and Caputo, P.M. (2007). Absolute vs relative performance rating formats: implications for fairness and organizational justice. *International Journal of Selection and Assessment,* **15,** pp. 302-316.

Selvarajan, T.T. and Cloninger, P.A. (2011). Can performance appraisals motivate employees to improve performance? A Mexican study. *The International Journal of Human Resource Management,* **iFirst 0,** pp. 1–22.

Shore, L.M. and Bleicken, L.M. (1991). Effects of supervisor age and subordinate age on rating congruence. *Human Relations,* **44,** pp. 1093-1105.

Shrivastava, A. and Purang, P. (2011). Employee perceptions of performance appraisals: a comparative study on Indian banks. *The International Journal of Human Resource Management,* **22,** pp. 632-647.

Skarlicki, D.P. and Folger, R. (1997). Retaliation in the workplace: the roles of distributive, procedural, and interactional justice. *Journal of Applied Psychology,* **82,** pp. 434-443.

Slaughter, J.E. and Greguras, G.J. (2008). Bias in performance ratings: clarifying the role of positive versus negative escalation. *Human Performance,* **21,** pp. 414-426.

Snape, E., Thompson, D., Yan, F.K. and Redman, T. (1998). Performance appraisal and culture: practice and attitudes in Hong Kong and Great Britain. *The International Journal of Human Resource Management,* **9,** pp. 841-861.

Stephan, W.G. and Dorfman, P.W. (1989). Administrative and developmental functions in performance appraisals: conflict or synergy? *Basic and Applied Social Psychology,* **10,** pp. 27-41.

Sweeney, P.D. and Mcfarlin, D.B. (1993). Workers′ evaluations of the "Ends" and the "Means": an examination of four models of distributive and procedural justice. *Organizational Behavior and Human Decision Processes,* **55,** pp. 23-40.

Sweeney, P.D. and Mcfarlin, D.B. (1997). Process and outcome: gender differences in the assessment of justice. *Journal of Organizational Behavior,* **18,** pp. 83-98.

Taylor, M.S., Tracy, K.B., Renard, M.K., Harrison, J.K. and Carroll, S.J. (1995). Due process in performance appraisal: a quasi-experiment in procedural justice. *Administrative Science Quarterly,* **40,** pp. 495-523.

Tepper, B.J., Moss, S.E. and Duffy, M.K. (2011). Predictors of abusive supervision: supervisor perceptions of deep-level dissimilarity, relationship conflict, and subordinate performance. *Academy of Management Journal,* **54,** pp. 279–294.

Tepper, B.J., Uhl-Bien, M., Kohut, G.F., Rogelberg, S.G., Lockhart, D.E. and Ensley, M.D. (2006). Subordinates' resistance and managers' evaluations of subordinates' performance. *Journal of Management,* **32,** pp. 185-209.

Tharenou, P. (1995). The impact of a developmental performance appraisal program on employee perceptions in an Australian federal agency. *Group & Organization Management* **20,** pp. 245-271.

Thomas, A.B. (2004). *Research Skills for Management Studies*. London: Routledge.

Tranfield, D., Denyer, D. and Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management,* **14,** pp. 207-222.

Tuytens, M. and Devos, G. (2012). The effect of procedural justice in the relationship between charismatic leadership and feedback reactions in performance appraisal. *The International Journal of Human Resource Management,* **iFirst 0,** pp. 1-16.

Tziner, A., Joanis, C. and Murphy, K.R. (2000). A comparison of three methods of performance appraisal with regard to goal properties, goal perception, and ratee satisfaction. *Group & Organization Management,* **25,** pp. 175-190.

Tziner, A., Murphy, K.R. and Cleveland, J.N. (2001). Relationships between attitudes toward organizations and performance appraisal systems and rating behavior. *International Journal of Selection and Assessment,* **9,** pp. 226-239.

Tziner, A., Murphy, K.R. and Cleveland, J.N. (2002). Does conscientiousness moderate the relationship between attitudes and beliefs regarding performance appraisal and rating behavior? *International Journal of Selection and Assessment,* **10,** pp. 218-224.

Tziner, A., Murphy, K., Cleveland, J.N., Yavo, A. and Hayoon, E. (2008). A new old question: do contextual factors relate to rating behavior: an investigation with peer evaluations. *International Journal of Selection and Assessment,* **16,** pp. 59-67.

van Dierendonck, D., Haynes, C., Borrill, C. and Stride, C. (2007). Effects of upward feedback on leadership behaviour toward subordinates. *Journal of Management Development,* **26,** pp. 228-238.

Varma, A. and Stroh, L.K. (2001). The impact of same-sex LMX dyads on performance evaluations. *Human Resource Management,* **40,** pp. 309-320.

Varma, A., Pichler, S. and Srinivas, E.S. (2005). The role of interpersonal affect in performance appraisal: evidence from two samples – the US and India. *The International Journal of Human Resource Management,* **16,** pp. 2029-2044.

Walsh, K. and Fisher, D. (2005). Action inquiry and performance appraisals: Tools for organizational learning and development. *The Learning Organization,* **12,** pp. 26-41.

Wang, X.M., Wong, K.F.E. and Kwong, J.Y.Y. (2010). The roles of rater goals and ratee performance levels in the distortion of performance ratings. *Journal of Applied Psychology* **95,** pp. 546-561.

Wayne, S.J. and Liden, R.C. (1995). Effects of impression management on performance ratings: a longitudinal study. *Academy of Management Journal,* **38,** pp. 232-260.

Welbourne, T.M., Balkin, D.B. and Gomez-Mejia, L.R. (1995). Gainsharing and mutual monitoring, A combined agency-organizational justice interpretation. *Academy of Management Journal,* **38,** pp. 881-899.

Wilson, K.Y. (2010). An analysis of bias in supervisor narrative comments in performance appraisal *Human Relations* **63,** pp. 1903-1933.

Woehr, D.J. (1992). Performance dimension accessibility: implications for rating accuracy. *Journal of Organizational Behavior,* **13,** pp. 357-367.

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

Wong, K.F.E. and Kwong, J.Y.Y. (2007). Effects of rater goals on rating patterns: evidence from an experimental field study. *Journal of Applied Psychology,* **92,** pp. 577-585.

Wood, R.E. and Marshall, V. (2008). Accuracy and effectiveness in appraisal outcomes: the influence of self-efficacy, personal factors and organisational variables. *Human Resource Management Journal,* **18,** pp. 295-313.

Youngcourt, S.S., Leiva, P.I. and Jones, R.G. (2007). Perceived purposes of performance appraisal: correlates of individual- and position-focused purposes on attitudinal outcomes. *Human Resource Development Quarterly,* **18,** pp. 315-343.

Zedeck, S. and Cascio, W.F. (1982). Performance appraisal decisions as a function of rater training and purpose of the appraisal. *Journal of Applied Psychology,* **67,** pp. 752-758.

Zimmerman, R.D., Mount, M.K. and Goff, M. (2008). Multisource feedback and leaders' goal performance: moderating effects of rating purpose, rater perspective, and performance dimension. *International Journal of Selection and Assessment,* **16,** pp. 121-133.

Figure 1. Subject categories and selected journals (journals' relevance and quality)

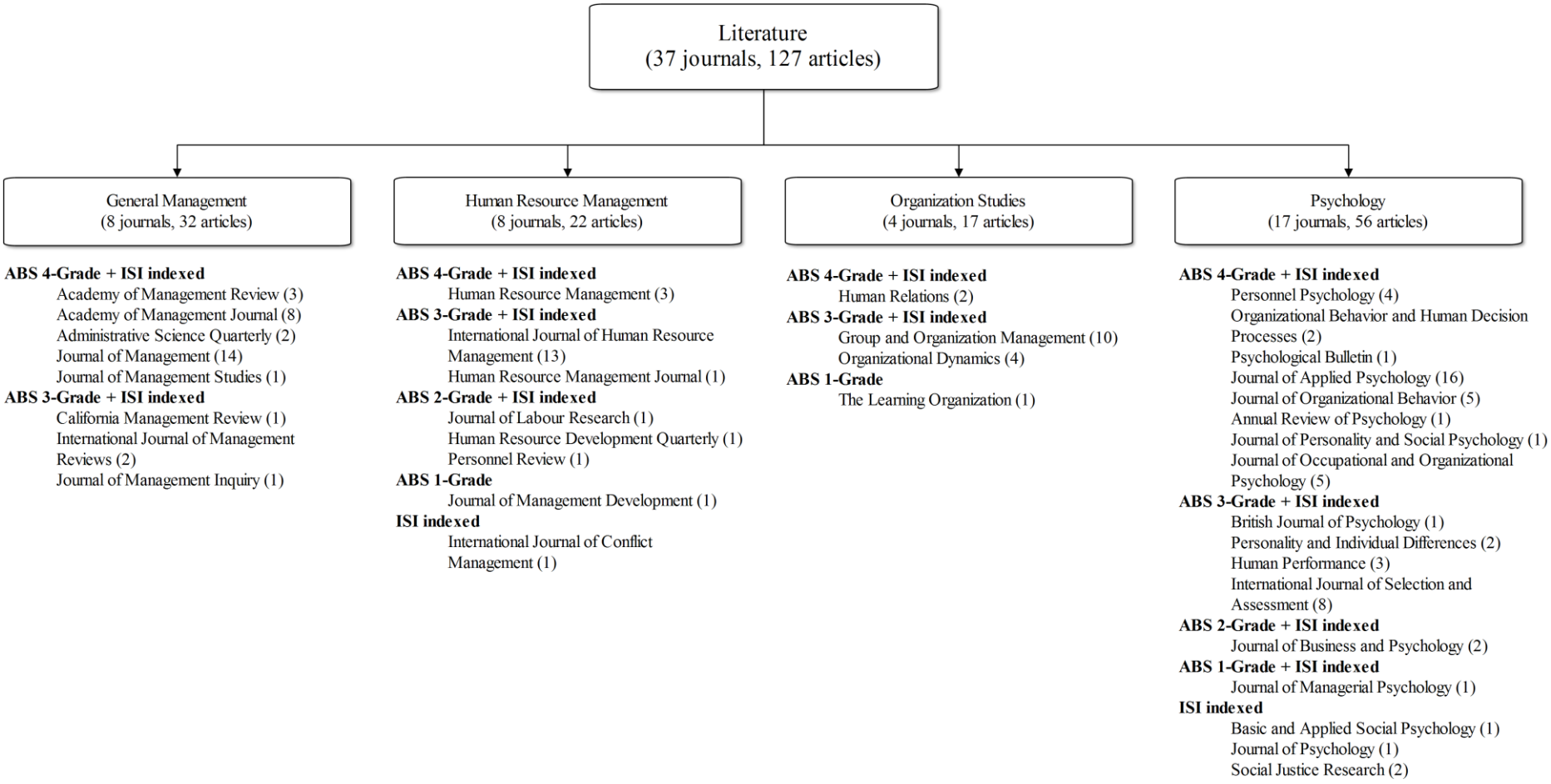
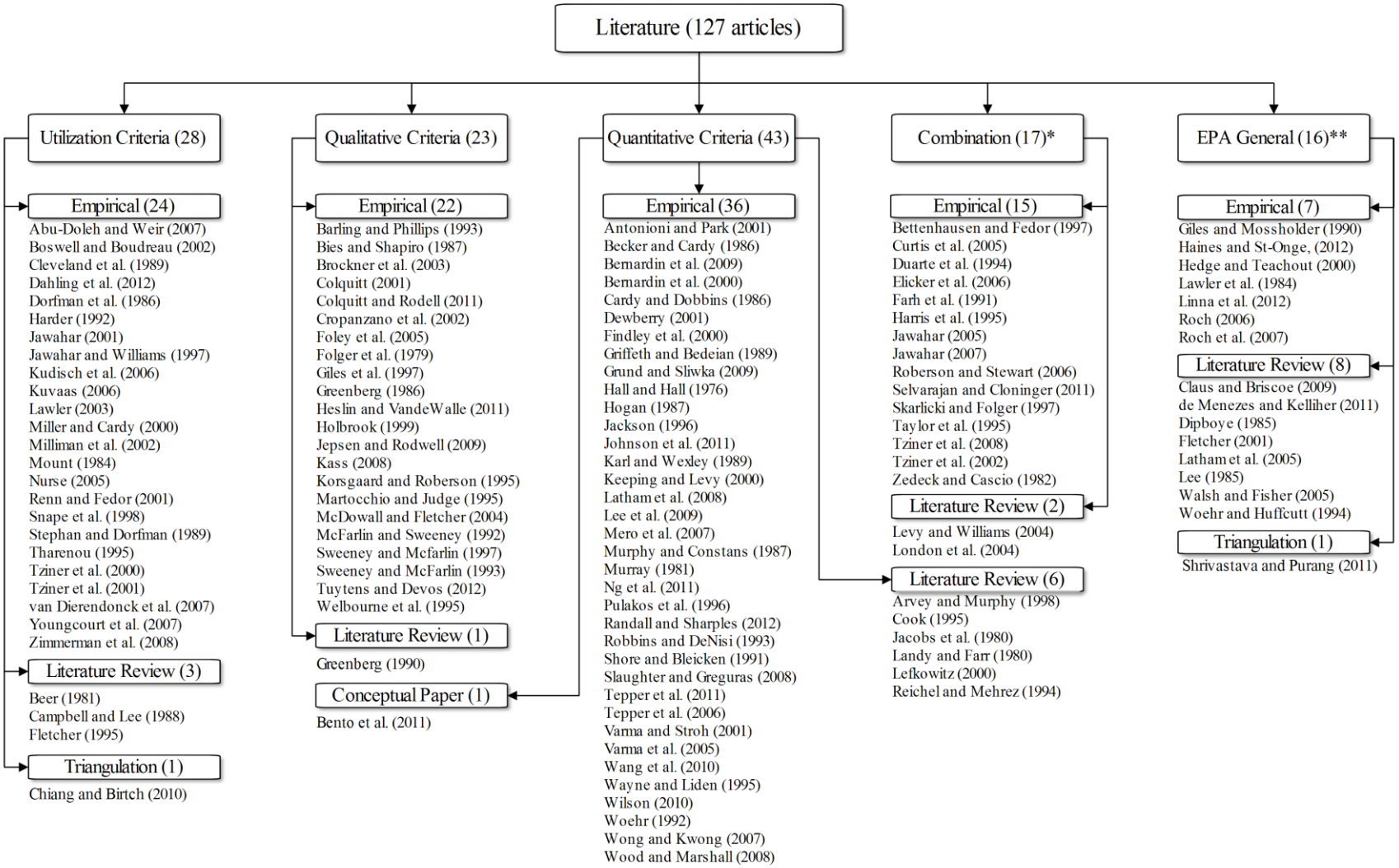


Figure 2. Content- and study type-wise articles (articles' relevance and quality)



* This category contains studies focusing on more than one measurement criteria.
** This category contains studies focusing on EPA in general.

Figure 3. Year of publishing-wise details (recentness)

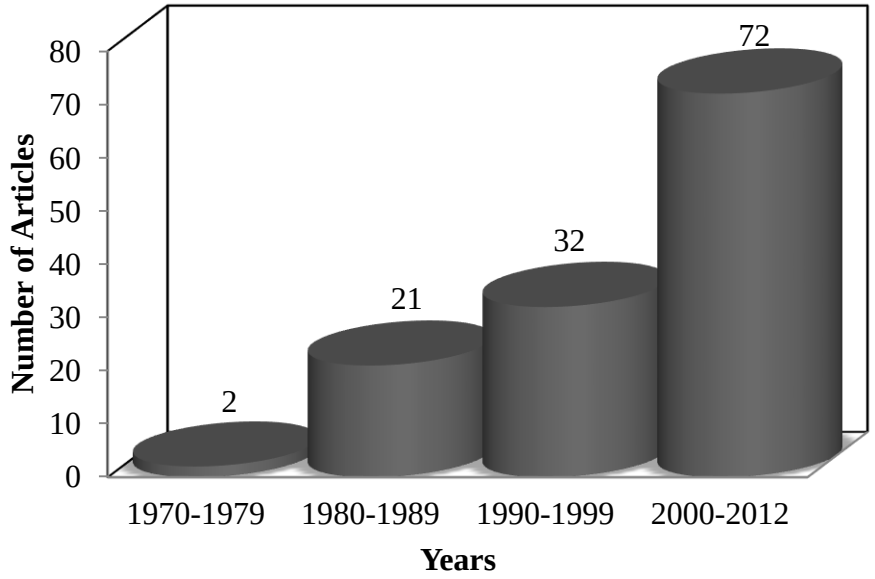
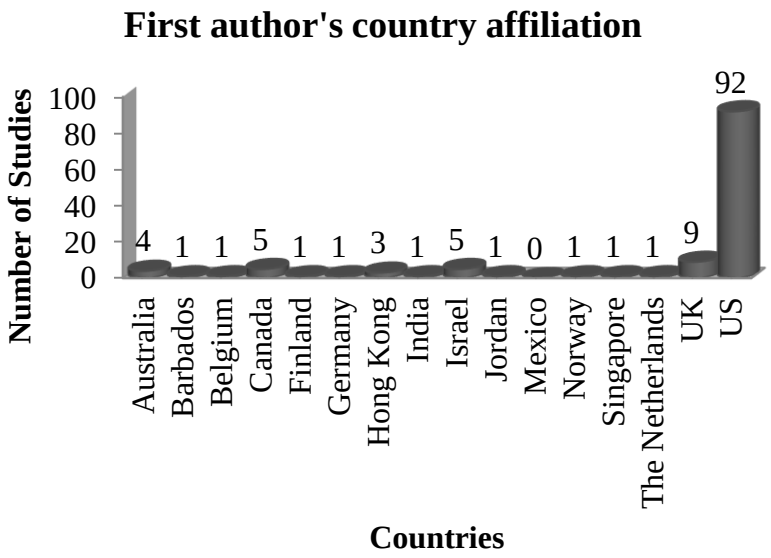
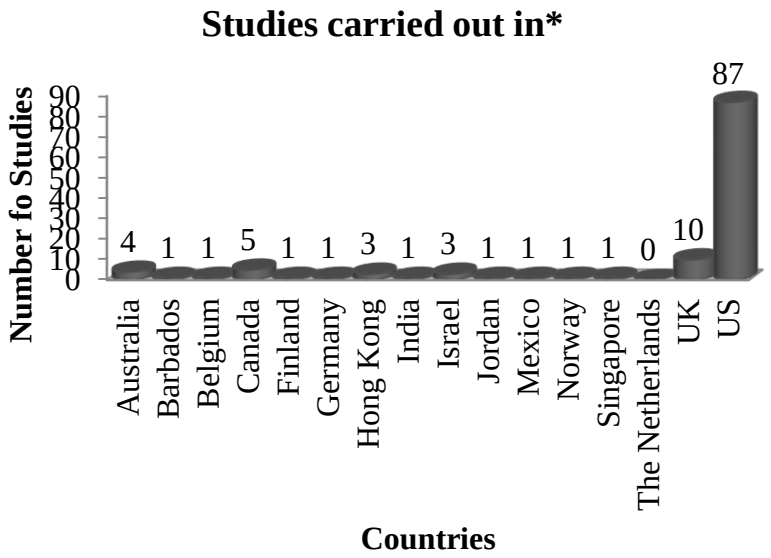


Figure 4. Principle authors' country affiliation and countries where studies were carried out





*The following studies are carried out in more than one country:
1. Tziner et al. (2001): US, Canada and Israel
2. Latham et al. (2008): Canada and Turkey
3. Snape et al. (1998): UK and Hong Kong
4. Varma et al. (2005): US and India
5. Milliman et al. (2002): Australia, Canada, Indonesia, Japan, Korea, Latin America, Mexico, People's Republic of China,

Figure 5. The integrated framework of EPA