# PRIMING AND PERFORMANCE RATING ACCURACY: NOTIFICATION OF RATING PURPOSE AND EXPOSURE TO COMPARATIVE EVALUATION STRATEGIES

by

CHRISTOPHER J. WAPLES

B.S., Nebraska Wesleyan University, 2006

A THESIS

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Psychological Sciences
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2013

Approved by:

Major Professor
Patrick A. Knight, Ph.D.

# Copyright

# Abstract

Despite the functional importance of performance appraisals in organizational settings, rating inaccuracies persist and have been a widely researched topic for decades. Contemporary efforts to explore the problem have turned to components of accuracy to foster a more detailed understanding of the influence of situational factors and individual biases. In particular, a great deal of research has examined the role of rating purpose (e.g., administrative, developmental) on subsequent accuracy, consistently revealing greater leniency for administrative ratings than for developmental ratings. On the basis of spreading activation theory, rating purpose was conceptualized as a priming event, and in combination with rating strategy priming, was expected to prompt predictable enhancements to specific components of accuracy. Participants for this experimental study were 160 undergraduate students. Participants were randomly assigned a rating purpose with "real-world" implications, and exposed to a strategy priming task designed to promote specific rating cognitions. Students viewed video-recorded competitive marching band performances, and rated them. Participants' ratings were compared to those made by experienced raters to compute accuracy estimates. Results were largely non-significant, but in the directions expected. Limitations and future research opportunities are discussed.

# Table of Contents

# List of Tables

# Chapter 1 - Introduction

Performance appraisals represent one of the most critical points in any successful organization. Though the interval may vary (i.e., annual, semi-annual), formal evaluations offer legal protection for business decision-making, highlight points of needed correction and opportunities for employee development, serve as a vehicle for providing well-supported performance feedback, and force supervisors and managers to seriously consider the contributions of their employees (DeNisi & Sonesh, 2010). Considering such important outcomes, strong appraisal systems often warrant considerable expense. Unfortunately, simply spending a great deal of time or money does not ensure that the system will work or continue to do so. Instead, research has shown that rater inaccuracy is a frequent problem at all levels of organizations, and in organizations of all sizes (Murphy & Cleveland, 1995; Sulsky & Balzer, 1988). For decades, researchers have sought solutions to this problem with somewhat limited success.

Although, historically, the bulk of research on performance-rating accuracy has focused on errors made by raters during the appraisal process (Landy & Farr, 1980; Murphy & Cleveland, 1995), problems with the conceptualization and use of error measures have prompted more recent examinations of target rating accuracy itself (Jelley & Goffin, 2001; Uggerslev & Sulsky, 2008). Refocusing on accuracy, and its components (Cronbach, 1955), has allowed researchers to more successfully assess the effectiveness of adjustments to rating processes and rater training protocols (Day & Sulsky, 1995; Schleicher & Day, 1998). Despite the empirical and practical improvements attained through rater training interventions, particularly those designed to establish a common frame-of-reference, inaccurate ratings persist. Accordingly,

exploration of alternative approaches as supplements to existing formal training efforts may be fruitful. It is with this goal in mind that the current project was conducted.

## Rating Accuracy

Assessment of rating accuracy in the appraisal process is not as simple as it may seem. Rating accuracy, with respect to multiple ratees, has four distinct components (Cronbach, 1955; Sulsky & Balzer, 1988): elevation, differential elevation, stereotype accuracy, and differential accuracy. Each is unique in terms of computation, and each provides evidence of wholly different organizational concerns.

Elevation (E) refers to raters' propensity, ignoring specific item and ratee differences, to rate high or low relative to true score estimates of performance. E can be computed using the following formula (Cronbach, 1955; Sulsky & Balzer, 1988):

$$E^2 = (\bar{x}.. - \bar{t}..)^2 \qquad (1)$$

where $\bar{x}_{..}$ is the average rating, and $\bar{t}_{..}$ is the average true score estimate. Given its calculation, if ratings are accurately made or vary as a function of non-systematic error, the value of E will be at or near zero. However, if raters are biased toward rating too high or low compared to true score estimates, E will increase, indicating a greater degree of inaccuracy. In an organizational context, a large value for E would represent the presence of either leniency or severity biases, suggesting a need for additional rater training efforts. Though high E-related inaccuracy may not result in erroneous promotion or salary decisions, it does have potential to negatively impact employee morale or influence decisions based upon direct comparison of work units (e.g., departments, work shifts).

Differential Elevation (DE) refers to raters' ability to correctly rank individual ratees, ignoring specific items or rating dimensions, relative to one another. To compute DE, the following formula can be used (Cronbach, 1955; Sulsky & Balzer, 1988):

$$DE^2 = \sigma_{\bar{x}_i}^2 + \sigma_{\bar{t}_i}^2 - 2\sigma_{\bar{x}_i}\sigma_{\bar{t}_i} r_{\bar{x}_i\bar{t}_i} \quad (2)$$

where $\sigma_{\bar{x}_i}^2$ is the variance of average ratings for all ratees across rating dimensions, $\sigma_{\bar{t}_i}^2$ is the variance of true score estimates for all ratees across rating dimensions, and $r_{\bar{x}_i\bar{t}_i}$ is the correlation coefficient between ratings and true score estimates. Conceptually, DE is most heavily influenced by the relationship between ratings and true scores ($r_{\bar{x}_i\bar{t}_i}$), such that a weak correlation between the two indicates that ratees are not consistently being rank ordered accurately. In the case of a large value for DE, raters are likely to require further training on either interpreting the rating scale correctly or consistently identifying relevant ratee behaviors. Because this type of accuracy addresses issues of rank order, DE becomes particularly important when making administrative decisions. A failure to accurately rank-order employees can lead to erroneous promotions, wage adjustments, and terminations, all of which have important legal implications.

Stereotype Accuracy (SA) refers to raters' capacity to accurately match true score average ratings for each item or rating dimension, ignoring ratee differences. Like DE, the computational formula is expressed in terms of rating and true score variance, and is expressed as follows(Cronbach, 1955; Sulsky & Balzer, 1988):

$$SA^2 = \sigma_{\bar{x}_{.j}}^2 + \sigma_{\bar{t}_{.j}}^2 - 2\sigma_{\bar{x}_{.j}}\sigma_{\bar{t}_{.j}} r_{\bar{x}_{.j}\bar{t}_{.j}} \quad (3)$$

where $\sigma_{\bar{x}_{.j}}^2$ is the variance of average ratings for specific dimensions across all ratees, $\sigma_{\bar{t}_{.j}}^2$ is the variance of average true score estimates for dimensions across ratees, and $r_{\bar{x}_{.j}\bar{t}_{.j}}$ is the

correlation between average ratings and average true score estimates. Provided the similarity in how values are obtained for both DE and SA, it is not surprising that SA is also largely determined by the relationship between ratings and true-score estimates. A large value for SA indicates that raters are failing to match the rank-order of true scores for each item or rating dimension across ratees. Such a value would represent a failure to identify which performance elements are relatively stronger or weaker than others for the entire collection of ratees. In an organizational setting, inaccuracies with regard to SA are particularly concerning from the perspective of training and development. Given the role of performance ratings in training needs assessment, SA inaccuracies can result in misdirected training efforts to address performance dimensions that require little correction or improvement.

Differential Accuracy (DA) refers to raters' ability to accurately assess individual ratees on each item or rating dimension. From a computational standpoint, this type of accuracy is the most complex, as it essentially requires calculation of DE for each rating dimension or facet, all of which are then averaged across items/dimensions. The following formulae can be used to obtain values for DA:

$$DA^2 = \frac{1}{kn}\sum_{ij}\sum\left[\left(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} - \bar{x}_{..}\right) - \left(t_{ij} - \bar{t}_{i.} - \bar{t}_{.j} - \bar{t}_{..}\right)\right]^2 \quad (4a)$$

or

$$DA^2 = {\sigma_a}^2 + {\sigma_b}^2 - 2\sigma_a\sigma_b r_{ab} \quad (4b)$$

where $x_{ij}$ is an individual ratee's rating on a single performance dimension or facet, $t_{ij}$ is that individual ratee's estimated true score for the dimension or facet, *a* is $\left(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} - \bar{x}_{..}\right)$, *b* is $\left(t_{ij} - \bar{t}_{i.} - \bar{t}_{.j} - \bar{t}_{..}\right)$, *n* is total number of ratees, and *k* is the total number of rating dimensions or facets. A high value for DA, then, may indicate that ratees were incorrectly rank-ordered for rating dimensions or facets. Practically, unlike inaccuracies with regard to DE or SA, large

values for DA may not result in individual employees being incorrectly classified for administrative purposes (e.g., promotions, wage adjustments), nor would it necessarily result in failures to identify departmental or organizational training needs. It may, however, result in failures to identify individual employees' training needs. Similarly, if an organization's appraisal system weights certain dimensions more heavily than others, DA's impact on administrative decision-making would become an important concern.

## Rating Purpose

In organizational settings, there are four general uses for performance appraisal information: 1) administrative decisions (e.g., promotions, raises, terminations); 2) employee development (e.g., training programs, feedback); 3) systems maintenance (e.g., validation of the appraisal instrumentation); and 4) documentation (Newman, Kinney, & Farr, 2004). The distinctive effects of administrative and development/research purposes on rating accuracy have received a great deal of attention from researchers (DeNisi, Cafferty, & Meglino, 1984; Jawahar & Williams, 1997; Newman, Kinney, & Farr, 2004).

Each of Cronbach's (1955) components of accuracy has different implications for these typical uses of appraisal information. Given that administrative uses of performance ratings focus primarily on comparing ratees, accuracy with regard to DE and DA becomes extremely important (Jelley & Goffin, 2001; Murphy & Cleveland, 1995). Accurate administrative decisions cannot be made on the basis of ratings that do not differentiate between ratees reliably. Though Sulsky and Balzer (1988) argued that accurate DE is not sufficient for administrative decision-making, it remains an important element for systematic decisions about the relative performance of multiple ratees. This is especially true for organizations that utilize variations of overall ratings. It can be expected that an appraisal system with reliable DE accuracy would be

better at correctly identifying ratees deserving of promotions and raises than would one without. Similarly, accurate DE can provide justification for less positive administrative outcomes: ratings-based terminations, lay-offs, and other undesirable consequences. Since DA is capable of providing the same information, albeit in a more specific manner, it contributes to these purposes in much the same way. However, in addition to allowing an appraisal system to accurately differentiate between employees in general, accuracy with regard to DA enables organizational leaders to target critical dimensions of performance and make decisions on the basis of immediate or strategic importance.

The impact of accuracy with regard to employee development is somewhat less straightforward. While a system with DE accuracy can provide some broad indications of which employees require development, it does not provide information about what sort of development an employee may need. SA, on the other hand, allows for determination of which dimensions are being performed optimally or suboptimally, but fails to indicate which employees are in particular need of development on those dimensions. Its value as an identifier of necessary group or departmental training is consistently recognized, but in instances where such training sessions have already taken place, that value is significantly diminished. Not surprisingly, however, previous literature is unanimous about the unmatched value of DA for developmental applications (Jelley & Goffin, 2001; London, Mone, & Scott, 2004; Murphy & Cleveland, 1995). DA is specifically concerned with raters' ability to correctly identify an individual employee's level on each rated performance dimension. This information is precisely that which is needed to make informed decisions about specific employee development.

## Priming

Since its development, priming has been defined in many ways. In what is, perhaps, the most inclusive conceptualization in the existing literature, Ratcliff and McKoon (1988) described priming as the extent to which an event "facilitates" a response in a future instance of performance or completion. Nevertheless, unaltered, this explanation seems to favor traditional, positive manifestations of priming effects, in which a prime-consistent response becomes more likely. It does not, however, address the existence of negative priming, in which an event may also actively *inhibit* future responses (Fox, 1995; May, Kane, & Hasher, 1995). Under conditions of negative priming, previously ignored information serves to suppress response time and accuracy when that information becomes a targeted cue. Failure to account for the potential inhibition of certain responses may limit the application of relevant theory to a range of priming phenomena, including in the current study. To correct for this deficiency, the following revised definition for priming will be used: *the extent to which an event facilitates or inhibits a response in a future instance of performance or completion*.

To truly understand priming, it is essential to understand the general organization of memory and the way in which it is accessed by the cognitive system. Historically, the conceptualization of cognitive architecture has been divided into varying levels of analysis (Lord & Maher, 1991). Traditionally, more macro-level, symbolic conceptualizations (i.e., sensory, short-term, long-term memories) have received the largest share of researcher attention (see Atkinson & Shiffrin, 1968). In recent decades, however, micro-level, connectionist models have grown in popularity. Connectionist models tend to rely on a neural metaphor that likens cognitive structure to that of a neural network, with conceptual nodes in memory being highly interconnected. Activation ("a momentary process based on an energy analogy that is closely

related to the idea of attention"; Lord & Maher, 1991, p. 20) of a given node allows for prompt, accurate retrieval of the information from memory. Due to the interconnectedness of the nodes themselves, strong activation of a single node prompts partial activation of other connected nodes.

Priming effects are loosely categorized by the mechanism through which they operate. Associative, semantic, and repetition priming each demonstrate functionally similar, but conceptually unique, effects. At their core, all priming effects are the result of learned relationships between concepts or events. The nature of those relationships, however, varies across the different approaches to priming. For both associative and semantic priming, the relationships can be expected to exist "naturally" across members of a population with a relatively homogenous educational and cultural history. Repetition priming, on the other hand, operates on specific relationships learned through systematic, repetitive pairing (frequently in experimental settings).

More specifically, associative priming exploits commonly-held associations between two events to prime a response to the second event (Fischler, 1977; Meyer & Schvaneveldt, 1971). For example, if presented the prime word "jump," followed by the target word "rope," participants tend to more quickly identify the target word as a real word than would participants first presented a word unrelated to the target (Fischler, 1977). Because of the way in which cognitive systems are organized, exposure to one concept frequently associated with a second expedites processing for the latter. If, however, the two concepts are infrequently or altogether unrelated, each concept must be processed independently.

Though comparable to associative, semantic priming operates by presenting events that lie within the same general cognitive categories, as opposed to events that are typically

associated with one another (Fischler, 1977; Lupker, 1984). As in the previous example, if a prime word "dog" were presented in advance of a target word "wolf," response time is likely to be faster than if an unrelated word were presented first. Although dogs and wolves are rarely paired in the vernacular, they are strongly associated in terms of cognitive classification. Consequently, the time required to process one when the other has been presented is decreased (Fischler, 1977).

Both associative and semantic priming tend to elicit relatively short-lived effects, with the latter sometimes lasting no longer than a few seconds (for a contrary perspective, see Becker, Moscovich, Behrmann, & Joordens, 1997). From a practical standpoint, such limited duration would seem to offer little value to solving realistic problems. Repetition priming effects, however, may offer more functional worth. These effects are driven by frequent exposure to events prior to the priming session. In essence, repeated pairing of previously unrelated events creates a lasting cognitive association between them (Forster & Davis, 1984; Logan, 1990). Though "wing" and "rock" are unlikely to be related cognitively, given sufficient repetition, such a relationship could be expected to form. There is some evidence that suggests that this form of priming has the potential to carry more long-term effects than the others (Kolers & Magee, 1978). One explanation is that the novelty of the manipulated association and the assessment setting may somehow create a stronger cognitive link than occurs for naturally associated events, specific to that setting.

Due to the minor differences between these types of priming effects, a number of theories have arisen as explanation of such effects (e.g., spreading activation theory – Collins & Loftus, 1975; two-process theory – Posner & Snyder, 1975; compound cue theory – Ratcliff & McKoon, 1988). For the most part, however, the theories are functionally equivalent and generally

complementary, and minor differences stressed in the literature tend to be largely semantic. On the basis of the literature reviewed, spreading activation theory is consistently regarded as the most broadly applicable, and one such application lies in the pursuit of enhancing the accuracy of performance ratings.

Spreading activation was initially proposed by Collins and Loftus (1975) to explain the operations of the cognitive system, and has been used to explain a number of cognitive phenomena. The theory was largely adapted and expanded from Quillian's (1967) theory of semantic networks, which lacked adequate translation from a computer's storage and retrieval mechanisms to psychological processes. Collins and Loftus translated the theory to the human mind, and proposed a number of corrections to account for then-recent research findings. Since its conception, the theory has remained fairly stable, and has been well-supported.

The most prominent theory in the priming literature, spreading activation theory, suggests that priming effects are a function of the cognitive activation of conceptual nodes which, in turn, partially activate adjacent nodes (Collins & Loftus, 1975). This activation continues to move along paths between conceptual nodes, growing weaker as the distance from the central concept increases. Given the widespread nature of the "Roses are red…" poetic framework in the vernacular, exposure to the word "red," and subsequent activation of its conceptual node, may therefore prompt somewhat weaker activation of the associated concept "rose." In turn, this partial activation of "rose" may prompt still weaker activation of the concept "violet," which may partially activate "blue," and so on. From a priming standpoint, use of "red" as a prime should then elicit an improvement in response time for the target word "rose", and a less impressive improvement if the target word were "violet". The process is, of course, more complex than it may seem at first glance. The word "red", as a color, may activate "blue"

through the channel described above, and/or through shared semantic categorization as prime colors. Subsequently, it is likely that "blue" would be more strongly activated than would "violet" having been stimulated by two activation paths (which are considered to be additive).

## Rater Priming, Rating Purpose, and Accuracy

The purpose of the present study is to examine the effectiveness of a simple priming manipulation with regard to enhancing accuracy on a subsequent experimental rating task, provided either an administrative or developmental rating purpose. In a practical setting, such an addition to existing performance appraisal processes would represent a cost- and time-efficient vehicle for increasing the likelihood of raters' application of desirable rating strategies.

Though an organization may have intended uses for ratings derived during the appraisal process, not all raters are aware of these intentions. Assigned no specific rating purpose, raters infer a purpose of the ratings from the situation itself to guide their evaluation of ratees, which has led to increased attention for techniques like frame-of-reference training (Uggerslev & Sulsky, 2008). These training programs are designed to increase accuracy by providing raters with a common perspective from which to rate employee performance on each dimension, and often include identification of the intended uses of the resulting ratings. However, research has shown that specific identification of rating purpose can have either a positive or a negative impact on accuracy (Greguras, Robie, Schleicher, & Goff, 2003).

Assignment of an administrative purpose has been found to lead to more lenient ratings, while a development or research purpose is often associated with comparatively severe ratings (Jawahar & Williams, 1997). While a practitioner's first response may be to simply conceal the true purpose of the appraisals (e.g., identifying all appraisal efforts as developmental to avoid leniency bias) such actions have serious ramifications for trust in management (Mayer & Davis,

1999). Because raters will, at some point, be making evaluations under the auspices of an administrative purpose, it is important to explore techniques to improve accuracy for such ratings.

Although it is not often presented as such, notification of rating purpose can be conceptualized as a priming event. By encouraging raters to make assessments for a specific purpose, trainers, supervisors, and researchers are effectively facilitating responses in rating behaviors. Consistent with spreading activation theory (Collins & Loftus, 1975), when a purpose is either explicitly identified or inferred, it is likely to activate associated memories and concepts. In the case of administrative ratings, for example, associations with potentially negative outcomes caused by an evaluative rating may become activated, discouraging the rater from making negative evaluations. DeNisi et al.'s (1984) implication that leniency may stem from raters' preferences to avoid the subsequent presentation of critical ratings to the ratee seems consistent with this notion. With a developmental purpose, however, positively-coded, goal-oriented outcomes (e.g., training, career enhancement) may be more likely to be activated (DeNisi et al., 1984). Although the results of negative developmental ratings may be aversive to a ratee, the consequences seem relatively less severe. Without such an immediate concern for negative outcomes, raters were expected to make an effort to identify correctible weaknesses, and avoid making lenient judgments.

*H1: Participants who are given a developmental rating purpose will provide performance ratings that exhibit a greater degree of accuracy with regard to elevation (E) than will participants given an administrative rating purpose.*

Nevertheless, raters who have been assigned the task of making ratings for an administrative purpose (i.e., to determine the winner) can be expected to conform to experimental instructions.

Further, when asked to make administrative ratings, activation of cognitive functions associated with distinguishing between ratees can be expected. Subsequently, it was expected that raters espousing an administrative purpose would provide accurate ratings with regard to the ratees' rank order.

*H2: Participants who are given an administrative rating purpose will provide performance ratings that exhibit a greater degree of accuracy with regard to differential elevation (DE) than will participants given a developmental rating purpose.*

Consideration of ratee performance from a developmental perspective can be expected to activate cognitive functions associated with identification of requisite areas for improvement. This enhancement of attention toward apparent deficiencies was expected to result in a greater degree of accuracy with regard to both SA and DA.

*H3a: Participants who are given a developmental rating purpose will provide performance ratings that exhibit a greater degree of accuracy with regard to differential accuracy (DA) than will participants given an administrative rating purpose.*

*H3b: Participants who are given a developmental rating purpose will provide performance ratings that exhibit a greater degree of accuracy with regard to stereotype accuracy (SA) than will participants given an administrative rating purpose.*

Spreading activation theory may also provide some explanation for strategy selection as an individual approaches an assigned task. Task feature identification will lead to increasing activation along certain cognitive pathways, until the task has been categorized. Once the task category has been identified, cognitive activation will spread to problem-solving strategies employed for previously-encountered, similar tasks. Research conducted by Earley and Perry (1987) provides evidence consistent with this description. In their study, participants primed

with planning strategies utilized similar strategies in a goal-setting/planning task that followed. It was therefore expected that priming a comparative evaluation strategy would promote the utilization of such an approach on a subsequent rating task, and in turn, enhance accuracy.

*H4: Participants who have been exposed to comparative rating strategy primes will provide ratings that are more accurate across all four of Cronbach's types of accuracy than will participants without such primes.*

While implementation of a comparative approach to performance ratings is expected to increase accuracy, particularly with regard to both DE and DA, findings reported by Jelley and Goffin (2001) indicated the contrary under some conditions. The authors primed participants with an instrument that required both global and comparative evaluations of each ratee. For subsequent ratings, primed participants did, in fact, show increases in DA when compared to an unprimed control group. In terms of DE however, they were significantly less accurate than that control group. To explain the unexpected result, the authors suggested that, having had the previous opportunity to make global judgments about the ratees, primed participants may have been more cognitively capable of moving beyond the global level into more specific behavioral distinctions, thus enhancing DA while detracting from DE.

Given Jelley and Goffin's (2001) findings and accompanying rationale, this study further sought to explore the impact of varying priming formats. Two distinct priming conditions were used to test the tenability of Jelley and Goffin's explanation for their pattern of results. The generic, non-target-specific prime (see Appendix D) was a stimulus that presented participants with a task requiring comparative judgments (systematically rank-ordering items from a "survival" team-building exercise) that were not immediately relevant to the actual ratings of interest. This stimulus was expected to enhance the likelihood that participants would employ a

comparative rating strategy when rating the target performances, without refocusing cognitive resources away from making global judgments about the target performances themselves. Exposure to a prime that required participants to make initial comparisons between the targets themselves (the target-specific prime; see Appendix C) was similarly expected to enhance the likelihood of employing a comparative rating strategy. However, having had the opportunity to make global judgments about the targets, as in Jelley and Goffin's study, was expected to reduce the relative accuracy of subsequent global comparisons (DE), while freeing cognitive resources for more accurate ratings of individual dimensions for each target performance (DA).

*H5a: Participants who have been exposed to a generic paired comparison prime will reflect a greater degree of accuracy with regard to differential elevation (DE) than will participants in other conditions.*

*H5b: Participants who have been exposed to a target-specific paired comparison prime will reflect a greater degree of accuracy with regard to differential accuracy (DA) than will participants in other conditions.*

A review of the literature revealed no previous studies that have examined the way in which primed effects of rating purpose and evaluation strategy interact. Spreading activation theory generally views priming effects as additive in nature (Balota & Paul, 1996). Consequently, an interaction effect seemed plausible. Where there is probable overlap between strategy and purpose primes (e.g., administrative purpose and generic comparative evaluation), it was expected that the impact of these priming events would be additive.

*H6: An interaction effect will exist between purpose and strategy priming, such that priming combinations expected to elicit similar accuracy enhancements will be more accurate than non-congruent priming combinations.*

# Chapter 2 – Method

## Participants

One hundred and sixty participants (31.3% Male, Mean Age = 19.26 years, 85.6% White/Caucasian) were recruited from undergraduate psychology courses at a large Midwestern university, and earned course credit as a function of their voluntary participation. *A priori* power analysis, conducted in *G*Power 3* using conservative estimates of effect size similar to those found in previous research (Jelley & Goffin, 2001), indicated that a total sample size of approximately 150 participants would allow for acceptable power when testing the hypotheses (~0.85; Faul, Erdfelder, Lang, & Buchner, 2007). Accordingly, the sample size was deemed sufficient for the purposes of this study.

Jawahar and Williams (1997) found that the effect of rating purpose on leniency – accuracy with regard to elevation – was moderated by "research setting, type of rater, type of appraisal stimulus, and direction/source of appraisal" (p. 921). As the authors noted, all of these moderators are stacked against researchers when using student samples. Although an organizational setting may have been more appropriate for this type of research, as is often the case, the study was conducted on a student sample to determine whether or not further examination of these hypotheses in an actual organization would be justifiable.

## Materials and Manipulations

### *Recorded Performances*

Participants were asked to view four videotaped marching band shows initially performed publicly at a regional marching competition. Each of the four recorded performances featured a different high school marching band of similar size (~120 individual band members). The four shows selected for inclusion in the study were chosen to promote variability in ratings across

dimensions on the basis of official judges' ratings at the conclusion of the competition. Although the use of group-oriented performances would seem to present an obstacle to ecological validity with regard to individual employee ratings, conceptualization of each rating element as a specific task within a greater job performance context actually creates a rating environment more closely mirroring organizational performance appraisal processes in organizations than many videotaped task performances used in previous research (Jelley & Goffin, 2001; Murphy, Balzer, Lockhart, & Eisenman, 1985). Inclusion of four rating targets is consistent with similar research conducted by Jelley & Goffin (2001) and will be sufficient for calculating the four types of accuracy being assessed.

### *True Score Estimates*

For determination of accuracy, true score estimates are a necessary component of this project. In accordance with Borman's (1977, as cited in Jelley & Goffin, 2001) approach to generation of true score estimates, experienced raters were used to evaluate the performance of the recorded marching bands. These experienced raters consisted of five graduate assistants and two undergraduates in marching band leadership roles from the Music department at the University from which participants were used. Each rater viewed the performances independently, with the opportunity and instructions to view each performance as many times as necessary to garner the information needed to provide a satisfactorily accurate rating on each rated dimension. As previous researchers have done, to avoid serial order effects, raters were given individual sheets for each dimension to be rated which could have been completed in any order.

To examine the ratings provided by experienced raters, both absolute agreement and consistency were examined. Using all provided ratings, two-way mixed-effects intraclass

correlation coefficients were computed for absolute agreement (*ICC* = 0.41) and consistency (*ICC* = 0.57). Within specific rating dimensions, observed ranges were unexpectedly high (*M* = 3.55, *SD* = 1.10). In order to reduce the potential effects of outlying ratings on true score estimates, averages of the experienced raters' ratings were calculated, having removed the highest and lowest rating for each dimension, reducing the observed ranges (*M* = 2.35, *SD* = 0.93).

### *Rated Dimensions*

In order to be consistent with typical competitive marching band rating dimensions, each performance was given specific ratings on the following five dimensions: Musical Performance, Marching, Percussion, Auxiliary, and General Effect (see Appendix B). The musical performance dimension consists of the rater's judgment of how well the performers *sounded*. Accurate evaluation of this dimension requires a rater to consider tone quality and clarity, as well as a general impression of the overall musicality of the performance as independent instrumental elements combine. When rating the marching dimension, raters evaluated how the performance *looked*, particularly with regard to performers' ability to remain in synchronized step and coordinated formation. The percussion ensemble of each taped presentation was independently evaluated, taking into account both musical and marching performance. For the percussion dimension, in particular, raters would have needed to consider the uniform movement and use of performers' equipment when making accurate ratings. The auxiliary dimension focuses upon the visual effect of non-musical performers (e.g., flag or rifle corps) in each performance. Movement synchronicity and precision represent the primary considerations of this rating dimension. Lastly, the evaluation of general effect requires raters to judge their overall impressions of both the musical and visual elements of each performance, making a judgment

that incorporates considerations of creativity, continuity, coordination, and subjective experience.

### *Rating Forms*

The rating forms intended for use by both experienced raters and participants were generated using both behavioral and conceptual anchors derived from the Bands of America Adjudication Guidelines (2012) for high school marching band competitions (see Appendices G & H). Each of the above-described dimensions was rated on a 10-point Likert-type scale with banded anchors allowing for some inherent subjective evaluation of performance elements. Prior to experimental use, these rating forms were submitted to the Director of Bands at the university from which the participants were obtained for review to ensure that anchors are appropriately described, and no objections were voiced.

### *Strategy Primes*

For this project, three distinct strategy primes were used (see Appendices C, E, & F). The first presented a target-specific paired comparisons task in which participants are asked to systematically compare the first song played by each of the four bands, placing the bands in rank order, without assigning any specific ratings. For the sake of consistency, the second presented a generic paired comparisons task that similarly forced participants to make systematic comparisons, between four objects from a survival scenario team-task (see Appendix D), in this case. Participants were asked to rank-order the items on the basis of their perceived importance to survival of the scenario's subjects. The third sheet presented a list of ten objects from the survival scenario, and asked the participants to evaluate whether the object in question was important or unimportant to survival. Participants presented with the third sheet were conceptualized as a control condition.

### *Purpose Manipulation*

To manipulate the assigned rating purpose, condition-specific verbal descriptions of the intent of the project were provided to participants. Participants assigned to the administrative condition were told:

> "…your ratings today will be used as one part of a decision to hire a visiting high school band director to lead a summer workshop hosted at the university. The director who is selected will receive $3,000 for his or her involvement with the workshop, which lasts for one week this upcoming summer. Additionally, in the past, being selected has led to additional consulting-type work on a more on-going basis, opening the opportunity for further compensation."

Those participants assigned to the developmental condition, however, were told:

> "Your ratings today, in combination with ratings made by University Bands' staff members, will be used to enhance the value of a summer workshop hosted here at the university. The directors of the four high school marching bands you will evaluate today have committed to participation in the workshop, which lasts for one week this upcoming summer. The ratings you make will help workshop coordinators specifically cater instruction for each band to focus upon those elements with the greatest potential for improvement."

## Procedure

After consenting to participation (see Appendix I), each participant was randomly assigned to a level on each of the independent variables (rating purpose & strategy primes) and provided with a corresponding packet that included: a letter in support of the cover story (Appendix A), a demographics questionnaire, instructions for evaluating marching band performances (Appendix B), strategy priming sheets (Appendix C, D/E, or D/F), and a rating form for each performance (Appendix G). Each participant was seated in a room with up to 11 other participants, and asked not to advance through the provided packet until instructed to do so.

Prior to presenting the videotaped performances, participants were asked to complete the demographic questionnaire on the front page of the provided packet. Once all session participants had completed the aforementioned questionnaire, the researcher instructed all

participants to carefully read the evaluation guidelines, and answered any questions participants voiced about the guidelines. The first video segment, which included one song performed by each of the target bands, was then shown. Participants were given approximately 30 seconds to reflect on the segment, after which they were asked to proceed to and complete the strategy priming task provided in their packet.

Upon completion, the specific rating scale anchors (Appendix H) were distributed to participants, followed by the presentation of the first of four target band performances. Each target performance was comprised of all remaining songs for that band (excluding the song included in the initial video segment). After each performance, participants were instructed to reflect upon the performance they had just seen, and then to make ratings based upon their evaluation of that performance. When all ratings had been made, the researcher presented the next band's performance, and the process was repeated for the remaining three video segments. Once all performances had been viewed and rated, packets were collected by the researcher, and participants were dismissed. To avoid the possibility of previous participants revealing the study's deceptive cover story, participants were debriefed by email at the conclusion of data collection (see Appendix J).

# Chapter 3 – Results

Preliminary analysis to examine the data's conformity with assumptions underlying the intended statistical procedures revealed significant violations of the assumption of normality. In all cells, significant positive skewness ratios existed for all four of the accuracy measures ($p < 0.01$). Given the squared nature of the dependent measures, such skewness is not uncommon. To improve the validity of any conclusions drawn from subsequent analyses, all dependent measures were transformed by applying a square root function (Tabachnik & Fidell, 2007).

Examination of the transformed variables revealed the transformation to have been effective, with skewness values falling to more acceptable levels (ranging from 0.49 to 1.08). From this point on, references to accuracy components specific to this study will refer to the transformed components. Tests of outliers on the accuracy components identified one participant who represented a significant outlier on multiple dependent variables. The participant was consequently excluded from all further analyses, resulting in a total sample size of 159 participants. Descriptive statistics, arranged by conditional cells, are provided in Table 1.

To verify the effectiveness of the deceptive cover stories with regard to establishment of the rating purpose, a series of manipulation check items were included at the conclusion of the experimental session. In response to the prompt, "Briefly explain the purpose for which the ratings obtained in this research will be used," 97% of participants gave open-ended responses that were consistent with the presented cover story. The remaining participants gave ambiguous responses, but no participants specifically indicated awareness of the deception inherent in the cover stories. Additionally, the participants generally considered their ratings to be "Somewhat Useful" to "Useful" ($M$ = 3.45, $SD$ = 0.73) in contributing to the purpose of the study. Taken in combination, these responses suggest that the Rating Purpose variable was successfully manipulated.

All formal hypotheses were tested using a 2 (Purpose) x 3 (Prime) between-subjects Multivariate Analysis of Variance (MANOVA). Cronbach's (1955) four types of rating accuracy were utilized as dependent variables. Results for the subsequent univariate ANOVAs are presented in Table 2.

Hypothesis 1, which predicted that participants assigned to a developmental purpose would make ratings displaying a greater degree of elevation-related accuracy than would those

participants assigned to an administrative purpose, was tested by the univariate main effect of purpose on elevation. Despite a mean difference in the expected direction, the analysis revealed no significant main effect, $F(1,153) = 3.55$, $p = 0.061$. Hypothesis 1 was not supported by the data.

Hypothesis 2, which predicted that participants assigned an administrative purpose would make ratings displaying greater differential elevation accuracy than would participants assigned a developmental purpose, was tested by the univariate main effect of purpose on differential elevation. The analysis revealed no significant main effect of purpose on differential elevation, $F(1,153) = 0.01$, $p = 0.85$. The direction of the non-significant difference between group means was, however, consistent with initial predictions. Hypothesis 2 was not supported by the data.

Hypotheses 3a and 3b predicted that participants assigned a developmental purpose would make ratings displaying greater levels of differential and stereotype accuracy, respectively, than would participants assigned an administrative purpose. The hypotheses were tested by the univariate main effects of purpose on differential and stereotype accuracy. These analyses revealed no significant main effect of purpose on either differential ($F(1,153) = 3.33$, $p = 0.070$) or stereotype accuracy ($F(1,153) = 0.08$, $p = 0.782$). For both types of accuracy, the direction of non-significant mean differences was consistent with predictions. Hypotheses 3a and 3b were not supported by the data.

Hypothesis 4, which predicted that participants exposed to comparative rating strategy primes would display greater accuracy across the four accuracy components than would those receiving a control task, was tested by the multivariate main effect of prime. Examination of group means revealed a pattern of mean differences consistent with the hypothesis. Statistical

analysis, however, revealed the multivariate main effect of prime to be non-significant, $\lambda = 0.97$, $F(8,300) = 0.59$, $p = 0.786$. Hypothesis 4 was not supported by the data.

Hypotheses 5a and 5b predicted more targeted effects of priming conditions, as a function of the primes' specificity and nature. In particular, Hypothesis 5a predicted that exposure to a generic paired-comparisons prime would prompt a greater degree of accuracy with regard to differential elevation than would other priming conditions. Group means were arrayed in the expected pattern. Analysis revealed the differences in differential elevation between priming groups to be non-significant, $F(2,153) = 0.21$, $p = 0.814$. Hypothesis 5b, on the other hand, predicted that exposure to a target-specific paired-comparisons prime would prompt a greater degree of differential accuracy than would exposure to other priming conditions. The pattern of group means was inconsistent with expectations, with the generic paired-comparisons group displaying greater accuracy than did the target-specific paired-comparisons group. Nevertheless, the main effect of prime on differential accuracy was not significant, $F(2,153) = 0.79$, $p = 0.457$. Neither Hypothesis 5a nor Hypothesis 5b were supported by the data.

Hypothesis 6 predicted a multivariate interaction between prime and purpose, such that cells with similar hypothesized influences upon accuracy would act to magnify one another, enhancing positive effects on accuracy, and exacerbating negative effects. The hypothesis was tested by the multivariate interaction between prime and purpose on the combination of accuracy components. Statistical analysis revealed the interaction effect to be non-significant, $\lambda = 0.95$, $F(8,300) = 1.05$, $p = 0.397$. Hypothesis 6 was not supported by the data.

A brief exploratory analysis was conducted to determine whether or not previous experience performing in marching bands would enhance participants' rating accuracy. A series of four *t*-tests were conducted to that end. The analysis revealed no significant mean differences

between participants with marching band experience and those without such experience for any of the four types of accuracy ($0.48 \leq t \leq 1.11$, $p > 0.05$).

# Chapter 4 – Discussion

The primary focus of this study was to examine the influence of various priming effects on the accuracy of performance ratings. To that end, both rating purpose and strategy exposure were conceptualized as priming events likely to activate specific cognitive mechanisms throughout the rating process. Tests of the hypotheses were largely unsupported by the observed pattern of results.

Despite the consistency of previous findings with regard to the impact of developmental and administrative rating purposes on elevation (DeNisi et al., 1984; Greguras et al., 2003; Jawahar & Williams, 1997), Hypothesis 1 was not formally supported by the data. Although the pattern of means was consistent with expectations, such that participants assigned to a developmental purpose ($M = 0.78$, $SD = 0.56$) were less prone to elevation-related inaccuracies than were participants assigned to the administration condition ($M = 0.97$, $SD = 0.63$), the magnitude of the difference was not sufficiently great to statistically confirm the hypothesis. In part, the small observed effect size was likely a function of the experimental setting itself. Meta-analytic research has indicated that leniency effects are dramatically smaller when using student raters in experimental settings than when in applied settings (Jawahar & Williams, 1997). Though the cover story was presented in a manner designed to magnify purpose effects beyond those typical of experimental settings, it does not appear to have successfully done so. Perhaps most notably, these results seem to present further evidence in support of the moderating influence of the rating environment upon the relationship between rating purpose and accuracy.

Participants who had been assigned an administrative rating purpose were expected to focus particular cognitive effort on distinguishing the performance of rating targets, and accordingly, to display more accuracy with regard to differential elevation than would participants who had been assigned a developmental purpose (Hypothesis 2). Assignment to a developmental rating purpose, on the other hand, was expected to activate cognitive processes associated with the identification of performance weaknesses with potential for improvement. The activation of such cognitive processes was hypothesized to enhance differential and stereotype accuracy for the target performances (Hypotheses 3a and 3b, respectively). The data, however, were not consistent with these expectations. Though small mean differences did exist in expected directions, they were generally so small as to be statistically and practically without value. The effect of rating purpose on differential accuracy may be an exception, with a small, but potentially informative, effect size ($\eta_p^2 = 0.02$). To the extent that rating purpose effects are suppressed in experimental settings (Jawahar & Williams, 1997), this small effect may very well represent a practically valuable effect in applied settings. Nevertheless, for this study, the nature of the rated performances, in combination with participants' limited experience with evaluation of such performances, may have overridden the activation of cognitive processes associated with correctly ranking the targets.

Hypotheses 4 and 5 were focused upon the impact of the strategy priming manipulation on the various types of accuracy. In particular, it was predicted that exposure to an algorithmic paired-comparisons strategy would prime subsequent utilization of systematic comparisons when rating target performances. Use of such an approach was expected to increase accuracy in general, and accordingly, Hypothesis 4 stated that participants exposed to pair comparisons primes (both task-specific and generic) would make more accurate ratings. As was the case for

the purpose-related predictions, the pattern of means was consistent with expectations, but mean differences were non-significant. In essence, Hypothesis 4 represented an examination of the effectiveness of priming strategy usage without formal instruction to do so. Though previous research has found strategy priming to influence strategy choice for subsequent tasks (Earley & Perry, 1987), the primes used in this study do not seem to have successfully prompted use of the intended strategies.

Hypotheses 5a and 5b were designed to test the viability of more targeted applications for rating strategy priming. Research conducted by Jelley and Goffin (2001) revealed that exposure to a target-specific prime resulted in unexpected differential elevation inaccuracy, but enhanced differential accuracy, for subsequent ratings. To address the issue, the authors speculated that participants, having made global performance considerations during the priming task, refocused their cognitive resources on making facet-specific evaluations. Accordingly, it was predicted that a generic paired-comparisons prime would impart the benefits of the algorithmic strategy without refocusing the participants' attention away from global judgments (Hypothesis 5a), while the target-specific paired-comparisons prime would operate similarly to the priming task used in Jelley and Goffin's study (i.e., detract from differential elevation, but enhance differential accuracy; Hypothesis 5b). Neither was supported by the data. The pattern of means was consistent with expectations for Hypothesis 5a, but not for Hypothesis 5b. For both differential elevation and differential accuracy, participants exposed to the generic paired-comparisons prime displayed the highest degree of accuracy.

The last of the formal predictions for this study focused on the interaction between the two priming manipulations – purpose and strategy. Because priming effects are generally regarded as cumulative (Balota & Paul, 1996; Ratcliff & McKoon, 1988), it was expected that

conditional intersections between primes expected to elicit similar accuracy effects would serve to enhance the likelihood of such effects occurring (Hypothesis 6). For example, a participant assigned a developmental rating purpose (+) and the target-specific pair-comparisons prime (+) was expected to display greater accuracy than a participant assigned an administrative rating purpose (-) and the generic evaluation prime (-). The pattern of results was not consistent with the hypothesis, however. From the perspective of spreading activation theory (Collins & Loftus, 1975), the absence of strategy priming effects may suggest that cognitive nodes associated with such strategies were underdeveloped or absent in the study's participants. If such were the case, priming those cognitive nodes would not have been feasible, and instead, the intended priming tasks may have simply represented an ineffective form of training.

## Limitations and Future Research Directions

The design of this study contains a number of inherent limitations which have the potential to limit the generalizability and magnitude of tested effects. As is often the case, however, these limitations likely represent fertile opportunities for further investigation of the mechanisms underlying accurate performance ratings.

First and foremost, this experimental study was conducted using university students enrolled in entry-level to mid-level psychology courses open to a wide variety of majors. While the breadth of experiences and backgrounds allows for a wider representation of the general public, student samples also carry a number of inherent similarities that prevent them from being easily generalized beyond university populations (e.g., age, education, environment). The employment of an experimental procedure using university students represents a noteworthy hindrance for this study, even beyond typical concerns regarding generalizability. For rating purpose manipulations, meta-analytic research has confirmed that the use of student samples can

reduce effect sizes by as much as 75% (Jawahar & Williams, 1997). To compound the issue, the same meta-analysis revealed that use of "paper people" or video-recorded task performances was associated with similar reductions in rating purpose-related effect sizes. Nevertheless, to justify research efforts in applied settings, it is important to first investigate and validate the occurrence of phenomena in more typical experimental settings. With consideration for the reduced effect sizes, the *a priori* power analysis was conducted with small estimates of effect size. To avoid underestimating realistic effect sizes, and to increase the likelihood of detecting legitimate effects, future research efforts targeting rating accuracy should be undertaken with special consideration for the viability utilizing applied samples.

The rating task itself also represents a likely limitation for this study. The decision to utilize recorded marching band competition films as the experimental stimuli was purposeful, as they: 1) allowed for consistent performance presentation, 2) were consistent with the perspective of a "booth" competition judge, 3) presented actual performances, instead of contrived task sequences involving scripted behavior, 4) represented a more realistic performance rating environment than micro-level task evaluation, and 5) provided an effective vehicle for "selling" the deceptive cover story. Having noted those benefits, however, the performances also bring some inherent difficulties to the rating environment. Given the wide-ranging educational backgrounds of the study's participants, their ability to effectively evaluate elements of marching band performances was likely inconsistent. Such inconsistencies prompt clear accuracy concerns, despite the availability of behaviorally-anchored rating scales. Additionally, the naturally subjective nature of the task content makes objective ratings challenging. Even the experienced raters from whom true score estimates were obtained returned highly variable responses. Last, the task of rating so many different elements in a single viewing represents an

unrealistic, and potentially overwhelming, rating circumstance, as in competition-judging, each element is rated by an individual judge.

Future research seeking to employ a task of this nature should examine the moderating influence of participants' experience with the content to be rated, to determine whether or not experienced participants possess a greater understanding of the requisite performance elements than do inexperienced participants. Use of a subsample consisting entirely of experienced marching band members seems the most feasible avenue for pursuing examination of such differences. Provided the concerningly high levels of disagreement among the experienced raters used to develop true score estimates, it may also be fruitful to explore more motivated expert or experienced raters as more reliable sources for true score estimation. The results of this study also suggest that a broader focus on the development of true score estimates likely warrants further investigation. For example, it may be that true score estimation would also be possible, given a well-constructed rating scale, by utilizing highly motivated, but naïve raters, as opposed to experienced raters prone to pre-existing biases. At the very least, a thorough review of Borman's (1977) guidelines is needed.

The primes used for the study also represent a probable limitation to the research design. Despite embodying the desired strategies, the capacity of the priming tasks to effectively activate strategy-relevant cognitions remains in doubt. Though some researchers have successfully primed strategy use (Earley & Perry, 1987), it may be more effective to integrate priming events as a sort of training refresher within the broader context of a rater training effort. Doing so would ensure that cognitive nodes consisting of comparison strategies exist, and potentially allow for their activation. Research examining the role of these low-impact reminders seems valuable to this body of literature. It is possible that the presentation order of the formal rating

process impeded successful priming by preventing back-to-back performance ratings. The serial nature of the experimental procedure may have prevented students from implementing a comparative strategy. Research on the potential moderating influence of serial versus simultaneous rating processes may be warranted to verify whether or not strategy priming can succeed in typical rating environments. Last, the use of a task-irrelevant scenario for two of the three priming tasks may have confounded the manipulation. The combination of a distraction from the task at hand and unique priming mechanisms may have unexpectedly influenced resultant accuracy. Inclusion of a brief, non-task-specific distractor task with the target-specific prime would clarify the results.

## Concluding Remarks

Although the hypotheses for this study were unsupported by the data, this research nevertheless provides some limited contribution to the discussion of performance rating accuracy. Certainly, the data is supportive of the existence of inherent problems underlying examination of performance ratings in experimental settings, using student samples. More interesting, however, were the inconsistencies at play during the development of true score estimates, and the corresponding implications regarding subjective evaluation of task performance. At the very least, such lack of agreement warrants consideration for the use of subject matter experts who may lack direct task experience.

# References


Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence (Ed.), *The psychology of learning and motivation: Advances in research and theory,* Vol. 2 (pp. 89-195). New York: Academic Press.

Balota, D. A. & Paul, S. T. (1996). Summation of activation: Evidence from multiple primes that converge and diverge within semantic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(4), 827-845.

Bands of America. (2012). *Bands of America official procedures and adjudication handbook.*

Becker, S., Moscovitch, M., Behrmann, M., & Joordens, S. (1997). Long-term semantic priming: A computational account and empirical evidence. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 1059-1082.

Collins, A. M. & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*(6), 407-428.

Cronbach, L. J. (1955). Processes affecting scores on 'understanding of others' and assuming similarity. *Psychological Bulletin, 52*(3), 177-193.

Day, D. V., & Sulsky, L. M. (1995). Effects of frame-of-reference training and information configuration on memory organization and rating accuracy. *Journal of Applied Psychology, 80*, 158-168.

DeNisi, A. S., Cafferty, T. P., & Meglino, B. M. (1984). A cognitive view of the performance appraisal process: A model and research propositions. *Organizational Behavior and Human Decision Processes, 33*, 360-396.

DeNisi, A. S., & Sonesh, S. (2010). The appraisal and management of performance at work. In S. Zedeck (Ed.), *Handbook of industrial and organizational psychology,* Vol. 2. Selecting and developing members for the organization (pp. 255-279). Washington: American Psychological Association.

Earley, P. C. & Perry, B. C. (1987). Work plan availability and performance: An assessment of task strategy priming on subsequent task completion. *Organizational Behavior and Human Decision Processes, 39*(3), 279-302.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191.

Fischler, I. (1977). Semantic facilitation without association in a lexical decision task. *Memory & Cognition, 5*, 335-339.

Forster, K. I., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 680-698.

Fox, E. (1995). Negative priming from ignored distractors in visual selection: A review. *Psychonomic Bulletin & Review, 2,* 145-173.

Greguras, G. J., Robie, C., Schleicher, D. J., & Goff, M. (2003). A field study of the effects of rating purpose on the quality of multisource ratings. *Personnel Psychology, 56*, 1-21.

Jawahar, I. M.& Williams, C. R. (1997). Where all the children are above average: The performance appraisal purpose effect. *Personnel Psychology, 50,* 905-925.

Jelley, R. B. & Goffin, R. D. (2001). Can performance-feedback accuracy be improved? Effects of rater priming and rating-scale format on rating accuracy. *Journal of Applied Psychology, 86*(1), 134-144.

Kolers, P. A., & Magee, L. E. (1978). Specificity of pattern-analyzing skills in reading. *Canadian Journal of Psychology, 32*, 43-51.

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

Logan, G. D. (1990). Repetition priming and automaticity: Common underlying mechanisms. *Cognitive Psychology, 22*, 1-35.

London, M., Mone, E. M., & Scott, J. C. (2004). Performance management and assessment: Methods for improved rater accuracy and employee goal setting. *Human Resource Management, 43*(4), 319-336.

Lord, R. G., & Maher, K. J. (1991). Cognitive theory in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology*, 2nd ed., Vol. 2 (pp. 1-62). Palo Alto, CA: Consulting Psychologists Press.

Lupker, S. J. (1984). Semantic priming without association. *Journal of Verbal Learning and Verbal Behavior, 23*, 709-733.

May, C. P., Kane, M. J., & Hasher, L. (1995). Determinants of negative priming. *Psychological Bulletin, 118*, 35-54.

Mayer, R. C. & Davis, J. H. (1999). The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of Applied Psychology, 84*(1), 123-136.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90,* 227-234.

Murphy, K. R., Balzer, W. K., Lockhart, M. C., & Eisenman, E. J. (1985). Effects of previous performance on evaluations of present performance. *Journal of Applied Psychology, 67,* 562-567.

Murphy, K. R. & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Thousand Oaks, CA: Sage.

Newman, D. A., Kinney, T. B., & Farr, J. L. (2004). Job performance ratings. In J.C. Thomas (Ed.), *Comprehensive handbook of psychological assessment,* Vol. 4: Industrial/organizational assessment, pp.956-1008. New York: Wiley.

Posner, M. I., & Snyder, C. R. R. (1975). Facilitation and inhibition in the processing of signals. In P. M. A. Rabbitt & S. Dornic (Eds.), *Attention and performance*. New York: Academic Press.

Quillian, M. R. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science, 12,* 410-430.

Ratcliff, R., & McKoon, G. (1988). A retrieval theory of priming in memory. *Psychological Review, 95*(3), 385-408.

Schleicher, D. J., & Day, D. V. (1998). A cognitive evaluation of frame-of-reference rater training: Content and process issues. *Organizational Behavior and Human Decision Processes, 73,* 76-101.

Sulsky, L. M. & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology, 73*(3), 497-506.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson.

Uggerslev, K. L. & Sulsky, L. M. (2008). Using frame-of-reference training to understand the implications of rater idiosyncrasy for rating accuracy. *Journal of Applied Psychology, 93*(3), 711-719.

## Table 1: Means and Standard Deviations (sorted by Cell Intersection)

| | | | Elevation | | Differential Elevation | | Stereotype Accuracy | | Differential Accuracy | |
|---|---|---|---|---|---|---|---|---|---|---|
| Purpose | Prime | N | M | SD | M | SD | M | SD | M | SD |
| Administrative | Task-specific Paired-comparisons | 26 | 0.95 | 0.65 | 0.97 | 0.46 | 0.45 | 0.21 | 2.16 | 1.13 |
| | Generic Paired-comparisons | 24 | 0.85 | 0.57 | 0.90 | 0.38 | 0.55 | 0.20 | 1.97 | 0.97 |
| | Generic Evaluation | 32 | 1.07 | 0.65 | 0.91 | 0.37 | 0.53 | 0.21 | 2.30 | 1.20 |
| | *Total* | *82* | *0.97* | *0.63* | *0.92* | *0.40* | *0.51* | *0.21* | *2.16* | *1.11* |
| Developmental | Task-specific Paired-comparisons | 25 | 0.71 | 0.60 | 0.87 | 0.34 | 0.51 | 0.23 | 1.72 | 1.04 |
| | Generic Paired-comparisons | 25 | 0.78 | 0.52 | 0.93 | 0.47 | 0.46 | 0.19 | 1.82 | 0.89 |
| | Generic Evaluation | 27 | 0.84 | 0.58 | 1.01 | 0.40 | 0.52 | 0.18 | 1.96 | 1.03 |
| | *Total* | *77* | *0.78* | *0.56* | *0.94* | *0.41* | *0.50* | *0.20* | *1.84* | *0.98* |

**Table 2: ANOVA Summary Table for Univariate Analyses**

| Source | Dependent Variable | Sum of Squares | df | MS | F | *p* | $\eta_p^2$ |
|---|---|---|---|---|---|---|---|
| Purpose | Elevation | 1.285 | 1 | 1.285 | 3.552 | 0.061 | 0.023 |
| | Differential Elevation | 0.006 | 1 | 0.006 | 0.039 | 0.845 | <0.001 |
| | Stereotype Accuracy | 0.003 | 1 | 0.003 | 0.077 | 0.782 | 0.001 |
| | Differential Accuracy | 3.724 | 1 | 3.724 | 3.333 | 0.070 | 0.021 |
| Prime | Elevation | 0.599 | 2 | 0.299 | 0.828 | 0.439 | 0.011 |
| | Differential Elevation | 0.067 | 2 | 0.034 | 0.206 | 0.814 | 0.003 |
| | Stereotype Accuracy | 0.058 | 2 | 0.029 | 0.687 | 0.505 | 0.009 |
| | Differential Accuracy | 1.758 | 2 | 0.879 | 0.787 | 0.457 | 0.010 |
| Purpose*Prime | Elevation | 0.239 | 2 | 0.119 | 0.330 | 0.719 | 0.004 |
| | Differential Elevation | 0.278 | 2 | 0.139 | 0.851 | 0.429 | 0.011 |
| | Stereotype Accuracy | 0.146 | 2 | 0.073 | 1.738 | 0.179 | 0.022 |
| | Differential Accuracy | 0.527 | 2 | 0.264 | 0.236 | 0.790 | 0.003 |
| Error | Elevation | 55.353 | 153 | 0.362 | | | |
| | Differential Elevation | 24.992 | 153 | 0.163 | | | |
| | Stereotype Accuracy | 6.447 | 153 | 0.042 | | | |
| | Differential Accuracy | 170.91 | 153 | 1.117 | | | |
| Total | Elevation | 57.631 | 158 | | | | |
| | Differential Elevation | 25.334 | 158 | | | | |
| | Stereotype Accuracy | 6.656 | 158 | | | | |
| | Differential Accuracy | 177.253 | 158 | | | | |

# Appendix A – Cover Letter for Experimental Packet

(Department Letterhead)

<<DATE>>

Dear Participant,

Thank you for your willing participation in this study. As you have been informed, this is a collaborative effort between the Psychology Department and the K-State University Bands. The project's purpose is two-fold: 1) members of the Psychology Department are interested in examining the differences between experimental and real-world rating scenarios, and 2) the K-State University Bands are interested in how typical audience members view marching band performances.

As undergraduate psychology students here at K-State, you (and your classmates) represent a wide range of majors and backgrounds. That variability is perfect for better understanding how audiences respond to marching band shows.

In combination with evaluations made in the Music Department, the ratings you make today will have a very real impact on decisions we will be making for our Summer Workshop for July 2013. The workshop is designed to improve high school marching bands and to offer detailed, constructive feedback to high school band directors. For 2013, four (4) moderately-sized bands will be attending, and the event will include direction from both K-State University Bands members/staff and an exceptional high school band director from Kansas.

Since your ratings will influence the success of the workshop, please evaluate the performances as accurately as possible, and make your ratings accordingly. Thank you in advance for your careful consideration.

Enjoy the shows,

Dr. Frank Tracz
Director of Bands
Kansas State University
226 McCain Auditorium

# Appendix B – Guidelines for Evaluation of Marching Performances

These guidelines have been prepared to give you an understanding of what you should be looking for, as you evaluate marching performances. There are two general questions to consider as you watch and listen to the performers: 1) How does the band *sound*? 2) How does the band *look*?

There is, of course, more to the process, but the above-listed questions underlie each of the elements you will be asked to rate individually. In all, there are five (5) ratings to make for each band: musical performance, visual performance, percussion, auxiliary, and general effect. For each rating, you will need to take into account a number of smaller elements.

**Musical Performance.** To evaluate each band's musical performance, you'll need to consider how well performers maintained tonal quality (e.g., Was the pitch even and accurate throughout, or were performers off-key?), how well the musical "phrases" were performed and integrated (e.g., Were there passages that seemed disjointed, out of place, or interrupted?), and whether or not the music was balanced across instrumental section (e.g., Did the trumpets play so loudly that other sections' were inaudible?).

**Visual Performance.** For marching shows, visual performance is largely a function of synchronicity. Accordingly, you will watch to watch for the extent to which the performers' movements are synchronized: Are performers staying in "step" (when taking a step, performers should be taking it at the same time, with the same foot)? Do members of a section lift/carry their instruments similarly? In looking at the entire scene, is the intended form of the band clear and well-presented? Are lines straight and/or evenly-spaced when it seems intended?

**Percussion.** As a specialized section, the percussion (drums, cymbals, etc.) usually receives a distinct rating from the rest of the band. To evaluate the percussion, you'll need to spend some of your viewing time focusing on the percussion section, and considering elements of both musical and visual performance: Is the rhythm consistent and accurate, or are there times when the music seems off-beat? Are visual elements performed by the percussion synchronized?

**Auxiliary.** Like the percussion, auxiliary units often receive their own ratings. The auxiliary consists of all the non-musical performers on the field (flag corps, dancers, etc.). Because these performers are not producing music, the primary consideration is that of visual performance. Are the visual elements well synchronized, or do they seem to be too early or too late? Are the auxiliary performers effectively integrated into the band's overall performance, or are they a distraction?

**General Effect.** Exactly what it sounds like, this is a something of an overall impression of the performance. Did the band *look* and *sound* good? Were they able to convey the intent behind the selected songs? Were you impressed by the performance? Was the performance cohesive and well-suited to the music?

# Appendix C – Target-Specific Paired Comparisons Prime

Please complete the following items with consideration for the performances you have just viewed by circling the appropriate response. When making comparisons between bands, please take into account the instructions provided for evaluating the performance of competitive marching bands. You may return to the instructions for evaluation and bands list for reference.

A = Band A
B = Band B
C = Band C
D = Band D

**1.** Which band performed their song more successfully?

| | | | | | |
|---|---|---|---|---|---|
| -A | or | B | -B | or | C |
| -A | or | C | -B | or | D |
| -A | or | D | -C | or | D |

**2.** Using the comparisons you have just completed please indicate each band's rank out of four (4). Remember to be consistent with your conclusions above (e.g., if you determined that A was more successful than B above, it should be ranked higher below).

| Employee | Rank |
|---|---|
| A | _____ |
| B | _____ |
| C | _____ |
| D | _____ |

# Appendix D – Scenario Description for Generic Primes


**The Situation**

For a year you and a group of friends have planned a canoe camping trip to the pristine wilderness of the Boundary Waters Canoe Area (BWCA). Finally, the big day arrived, and all of you met at Don Beland's Wilderness Canoe Trips base camp for the two-week, flatwater adventure. At Beland's, you pitched your tents, completed plans for your trip, and tried to get a good night's sleep. Unfortunately, the anticipation of the trip and the machine-gun-like sound of the rain on the tents caused most you to get very little sleep. The temperature was in the 40s, but weather in the Boundary Waters is uncertain, and you hoped it would change for the better.

Nonetheless, you all got up early, had a big north woods breakfast, and headed to the docks. There you loaded your canoes onto "towboats" for a short trip to New Found Lake, the start-off point of your adventure. Your destination was McEwan Lake, with your first overnight at Knife Lake. You planned the first two days to be hard paddles (16 miles the first day), but after that you all expected to fish and relax as you pleased.

By the time you were dropped off by your tow boats, the rain had stopped, but the sky remained various shades of slate grey. Your party shoved off in three 17' aluminum canoes. With considerable effort you made your first destination as planned, arriving tired and achy but still enthusiastic about the trip ahead. The tent set-up was kept simple, and dinner was freeze-dried food – tasteless but satisfying. Everyone turned in before 9PM so that they could be fresh for the following day's paddle.

The second day began just like the first day; the sky was gun-metal grey streaked with black and threatened more serious rain. In addition, the wind had started blowing harder during the night. The forecast for the day seemed ominous. Your paddle began after a quick breakfast of instant oatmeal. At mid-day, the group left the heavily traveled Knife Lake chain to head toward remote McEwan Lake.

By late afternoon the rains came, with the wind almost reaching gale force. You had reached the northern end of McEwan Lake. Heading to your campsite at the southern end, your group cut directly across the open water. Suddenly, the wind picked up and whipped the water into small whitecaps that threatened to engulf the heavily provisioned canoes.

"Hey, this wind is really strong," said one paddler. "Look at those waves," said another. "They could flip us!"

Just as you started to head the canoes toward shore from the center of the lake, they caught a few gusts of wind, causing them to swing broadside to the current and broach. You struggled to keep

your canoes upright, but two flipped over. The third canoe met the same fate as the two paddlers tried to come to the rescue of the others. All three canoes were now swamped!

The contents, which had not been secured, were dumped into deep water. Some of the items sank immediately, while others floated but were being carried off by the current. One paddler yelled, "Stay with the canoes! They'll float!" Another paddler shouted, "Grab whatever you can reach!" You were able to retrieve some of the equipment, but most of it disappeared. All of the paddles, which are designed to float, were out of reach. Despite the confusion, everyone remained calm. The water temperature was around a chilly 60F, yet in an hour you all managed to coax your swamped canoes to the shore.

**The Problem**

Your party has reached the west shore of McEwan Lake. This lake is in a particularly remote location; weeks may pass before another group is seen. Hiking out of the area would be very difficult and time-consuming because of the rugged terrain and lack of trails.

The shore you have reached is rocky, and there are no developed campsites to be found. Ironically, your only greeting is the lonely laugh of the loon, Minnesota's state bird. Growing almost to the edge of the rocky shore is a dense, coniferous forest, typical of this area. Timber wolves, moose, white-tailed deer, black bear, fox, and coyotes roam the woods.

Everyone is dressed in jeans, rain jackets, and hiking boots. All are wet. No one has anything in his/her pockets (except for a police whistle in one person's pocket) as all wallets, coins, and jewelry were checked at the outfitters for safe keeping. Two people have Swiss Army knives secured to their belts.

As your group emerges from its ordeal, all but one person seems in good shape. He appears unusually pale, and his skin is cold to the touch. His pupils are slightly dilated, and he is shivering violently. He also seems disoriented, unsure of what has happened.

You managed to salvage a small collection of items after the swamping. They are listed in on the next page of this packet. You must now decide what actions to take and how to use the salvaged items to aid your survival.

# Appendix E – Generic Paired Comparisons Prime

Please respond to the following questions regarding the scenario you have just read. When making comparisons between items on the basis of their importance, consider the specifics of the scenario and how the various items may be used to help you and your group survive the described circumstances. You may return to the scenario description for reference, if you choose.

A = Assembled aluminum cooking kit (contains plates, cups, & pots)
B = 1 sleeping bag (day-glow orange interior) with waterproof sack
C = Hiker's portable water filter
D = A small container of waterproof matches

**1.** Which item is more important to the survival of your group in the scenario?

| | | | | | |
|---|---|---|---|---|---|
| -A | or | B | -B | or | C |
| -A | or | C | -B | or | D |
| -A | or | D | -C | or | D |

**2.** Using the comparisons you have just completed please indicate each item's rank out of four (4). Remember to be consistent with your conclusions above (e.g., if you determined that A was more important than B above, it should be ranked higher below).

| Animal | Rank |
|---|---|
| A | _____ |
| B | _____ |
| C | _____ |
| D | _____ |

# Appendix F – Generic Evaluation Prime

Please respond to the following questions regarding the scenario you have just read. When making decisions about the importance of each item, consider the specifics of the scenario and how the various items may be used to help the group survive the described circumstances. You may return to the scenario description for reference, if you choose.

**For each of the following items recovered by you and your group in the scenario, please indicate whether you believe the item to be important or unimportant to the group's survival by circling the appropriate response:**

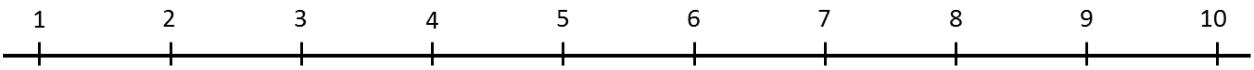
| | | | |
|---|---|---|---|
| 1. | "Muskol" insect repellent | Important | Unimportant |
| 2. | A small container of waterproof matches | Important | Unimportant |
| 3. | Four (4) floatable seat cushions | Important | Unimportant |
| 4. | A tackle box with assorted line, hooks, weights, lures, and snare wire | Important | Unimportant |
| 5. | A map of the Boundary Water Canoe Area (in plastic waterproof envelope) | Important | Unimportant |
| 6. | One (1) one-quart plastic water bottle, one-quarter full of Scotch whiskey | Important | Unimportant |
| 7. | Eight (8) Payday candy bars in a plastic bag | Important | Unimportant |
| 8. | One (1) sleeping bag (day-glow orange interior) in a waterproof sack | Important | Unimportant |
| 9. | One (1) hiker's portable water filter | Important | Unimportant |
| 10. | An assembled cooking kit for six (6) people (contains plates, cups, and cooking pots) | Important | Unimportant |

# Appendix G – Performance Rating Form

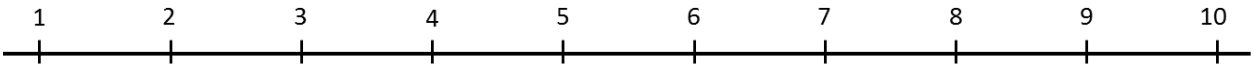
Performance Rating Form

*In combination with the rating anchors provided, please use the scales below to rate the band upon the following five (5) dimensions: musical performance, visual performance, percussion, auxiliary, and general effect. For each element,* ***circle one number*** *from the scale to indicate the band's performance.*

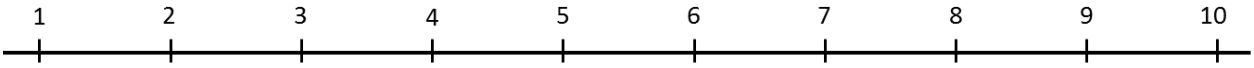
**Musical Performance.** How did the band *sound*? Consider: tonal quality, musical phrasing, musical balance across sections



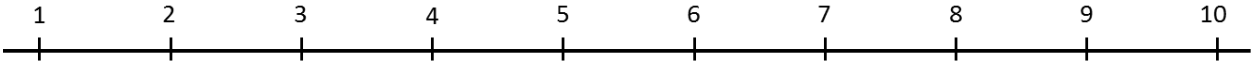

**Visual Performance.** How did the band *look*? Consider: synchronicity (staying in "step," instrument movement), form presentation, performer spacing, line straightness



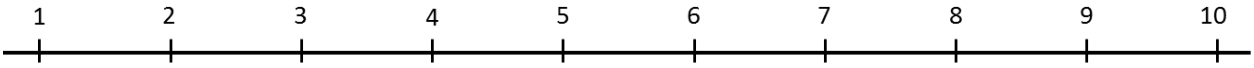

**Percussion.** How did the percussion section *look* and *sound*? Consider: rhythmic consistency, synchronicity (staying in "step," instrument movement), uniformity




**Auxiliary.** How well did the non-musical performers contribute to the performance, as a whole? Consider: synchronicity (equipment use, movements), integration into performance

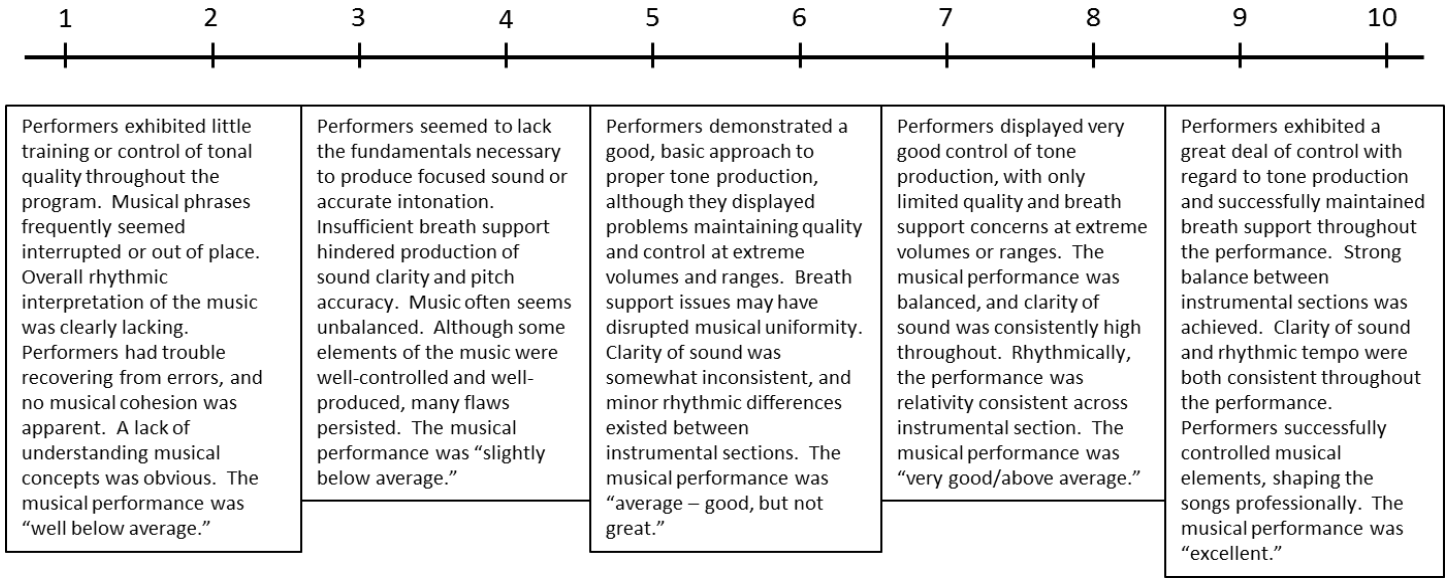



**General Effect.** Overall, how well did the band perform? Consider: integration of musical and visual elements, overall impression, cohesiveness

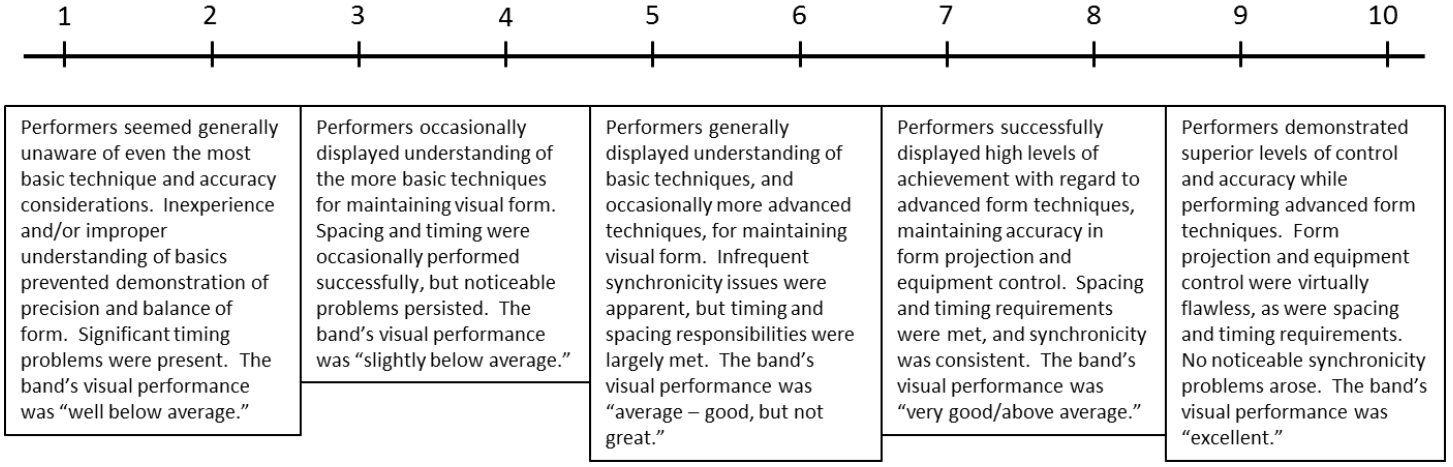


# Appendix H – Supplementary Anchors for Performance Rating Form

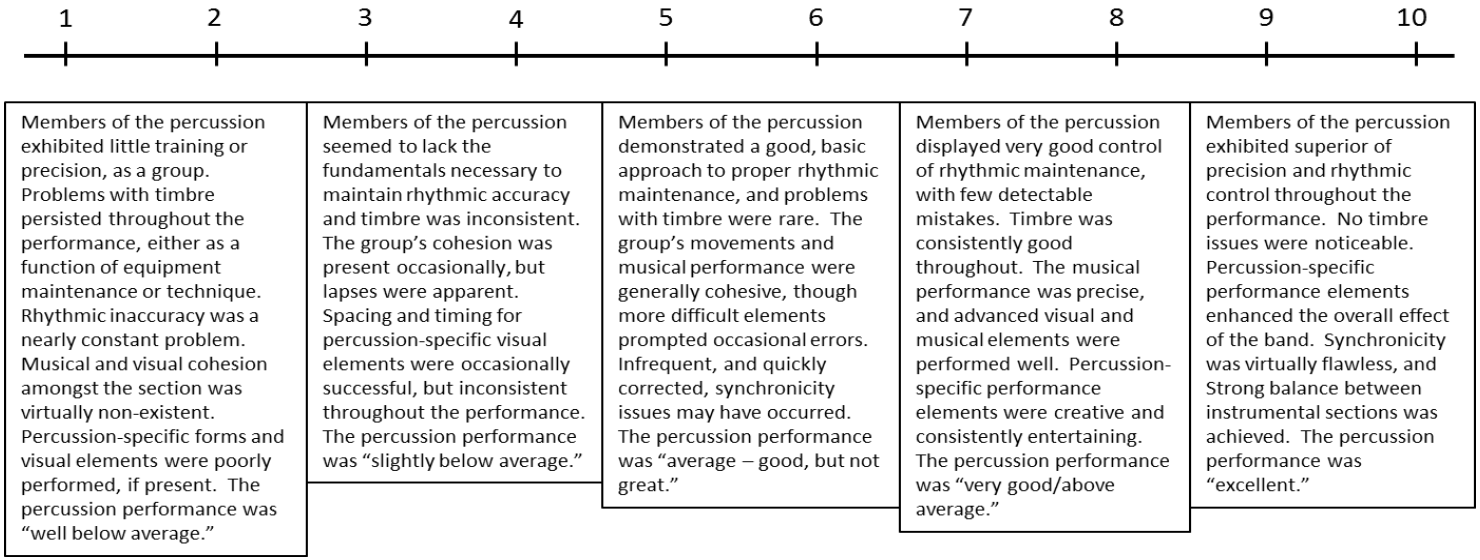
**Musical Performance.** How did the performance *sound*?



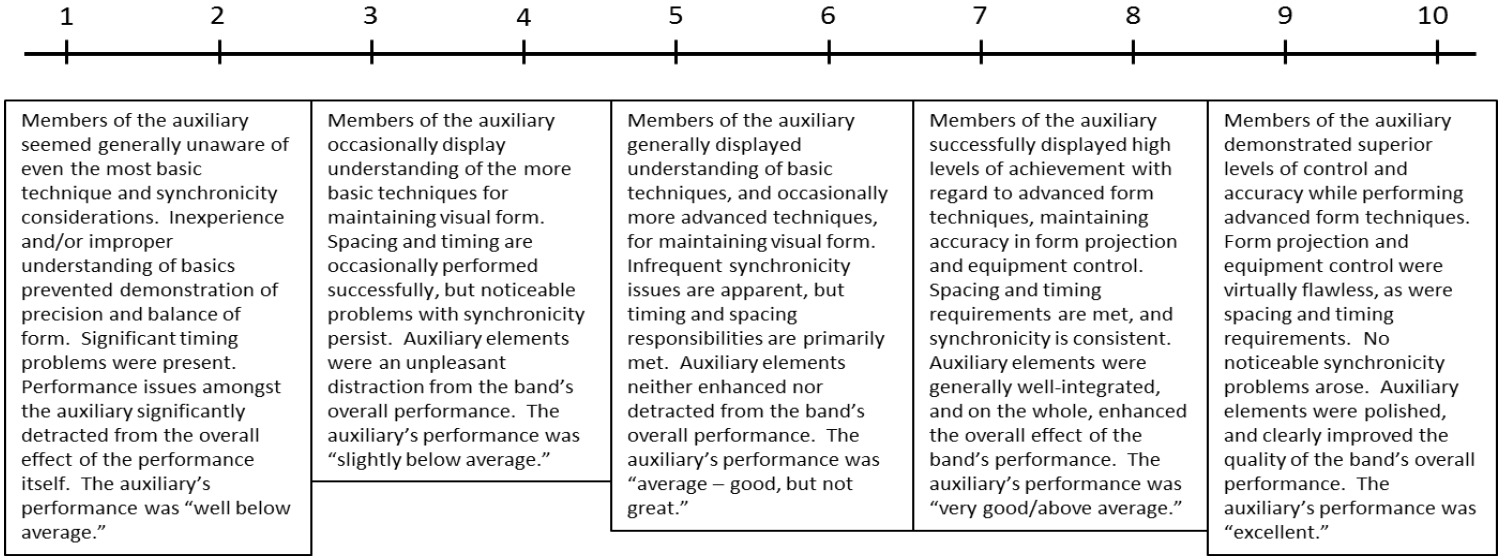

**Visual Performance.** How did the performance *look*?



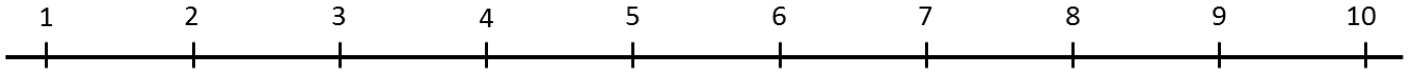
**Percussion.** How did the percussion section *look* and *sound*?




**Auxiliary.** How well did the non-musical performers contribute to the performance, as a whole?



**Overall Effect.** Overall, how well did the band perform?




| | | | | |
|---|---|---|---|---|
| A lack of understanding the most basic elements of overall effect makes evaluation difficult. The performers lacked unified effort and were unable to convey the program's intent and message. Musical and visual elements failed to complement one another and often appeared to be in conflict. The overall effect produced by the performers did not succeed. The overall performance was "well below average." | Performers occasionally displayed awareness of the fundamentals of aesthetically appealing overall effect. Unified effort to convey the music's intent and message was noticeable, but inconsistent. Energy level fluctuated throughout a seemingly uninspired performance. Despite some successful elements, others impeded the performance's success. The overall performance was "slightly below average." | Fundamental concepts of overall effect were clearly present, but generated only moderate levels of consistent aesthetic appeal. Performers displayed average effort in conveying the music's intent and message, and were generally successful. The performance itself was entertaining, but somewhat diminished by lapses in performer concentration or artistry. The overall performance was "average – good, but not great." | Performance of above-average quality, with cleared defined concepts of overall effect contributing to a sense of consistent aesthetic appeal. The performers displayed an above average effort in conveying the intent and message of the music, and demonstrated understanding of blending musical and visual elements. Some occasional breaks in performance continuity and climax were not maximally effective, but generally, the performance was constantly entertaining. The overall performance was "very good/above average." | Advanced concepts of overall effect were clearly understood and successfully developed throughout the performance, contributing to a strong sense of aesthetic appeal that was constantly entertaining. All musical and visual elements were well blended and delivered to maximum effect. Performers displayed significant effort and succeeded in conveying the music's intent and message. The overall performance was "excellent." |

# Appendix I – Informed Consent Form

**PROJECT TITLE:** Priming and Performance Rating Accuracy: Notification of Rating Purpose and Exposure to Comparative Evaluation Strategies

**PRINCIPAL INVESTIGATOR:** Patrick Knight, Ph.D.

**CO-INVESTIGATOR(S):** Chris Waples

**CONTACT NAME AND PHONE FOR ANY PROBLEMS/QUESTIONS:** Chris Waples, cwaples@ksu.edu or Dr. Patrick Knight, knight@ksu.edu

**IRB CHAIR CONTACT/PHONE INFORMATION*:***

- Rick Scheidt, Chair, Committee on Research Involving Human Subjects, 203 Fairchild Hall, Kansas State University, Manhattan, KS 66506, (785) 532-3224.

**PURPOSE OF THE RESEARCH:** To improve understanding of performance rating accuracy, with the intent to enhance accuracy for future appraisal processes.

**PROCEDURES OR METHODS TO BE USED:** You will be asked to watch a series of video segments and rate the performance of individuals you observe through that medium.

**LENGTH OF STUDY:** 50 minutes

**RISKS OR DISCOMFORTS ANTICIPATED:** None expected.

**BENEFITS ANTICIPATED:** A better understanding of the determinants of performance rating accuracy can improve the way performance appraisals are administered in the workplace.

**EXTENT OF CONFIDENTIALITY:** No identifying information is to be collected. Furthermore, all data will be kept in secure locations, both electronically and physically.

**TERMS OF PARTICIPATION: I understand this project is research, and that my participation is completely voluntary. I also understand that if I decide to participate in this study, I may withdraw my consent at any time, and stop participating at any time without explanation, penalty, or loss of benefits, or academic standing to which I may otherwise be entitled.**

**I verify that my signature below indicates that I have read and understand this consent form, and willingly agree to participate in this study under the terms described, and that my signature acknowledges that I have received a signed and dated copy of this consent form.**

**Participant Name:** _____

**Participant Signature:** _____ **Date:** _____

**Witness to Signature: (project staff)** _____ **Date:** _____

# Appendix J – Debriefing Information Provided to Participants

Thank you very much for your participation in our performance rating study (marching band evaluation) this semester! Your support of the research process is very important; without volunteers like you, much of what we know about psychology would remain unknown.

As part of this project, we chose to employ a deceptive cover story to encourage you to make ratings as carefully and accurately as possible. More specifically, you were told that the research was a joint effort between our department and the Kansas State University Bands, and that your ratings would directly contribute to a summer workshop being hosted by the Bands. Though we were indeed collaborating with the University Bands, they were largely involved for technical advisement. The summer workshop element was wholly deceptive. Ratings collected as part of the research process were not and will not be used for hiring staff or developing instruction.

The research you have just participated in is designed to investigate: 1) the effects of knowing what the purpose of your rating is before it is made, and 2) the effects of having been exposed to varying types of comparative evaluation strategies prior to making your rating. It is expected that changes in these elements will lead to very different impacts on rating accuracy.

Your ratings will now be compared to ratings made by individuals who have a great deal of involvement with competitive marching band preparation and evaluation to assess how accurate your ratings were. Differences between your ratings and the ratings of the "experts" will be examined to identify trends on the basis of the conditions to which you were assigned during the experiment, and will hopefully be used in the development of a performance rating process with the potential to enhance performance ratings in a variety of environments.

If you have any further questions about the study, or are interested in receiving information about its results upon completion, I encourage you to contact me (Chris Waples, cwaples@ksu.edu). If you have additional concerns about this project or the way it was administered, please refer to the informed consent form you were provided during the experimental session.

Thank you again! Your time is appreciated.