October, 2012

Rating Formats Revisited: Yes, They DO matter

C. Allen Gorman, Radford University

Available at: https://works.bepress.com/c-allen-gorman/38/
Rating Formats Revisited: Yes, They DO Matter!

C. Allen Gorman, PhD
Radford University
Overview

- Classic rating format research
- Halo error research
- Contemporary rating format research
- Frame-of-reference scales
- Conclusions of rating format research
- Future research on rating formats
Rating Format Research

- Landy & Farr (1980)
  - Interventions designed to improve rating formats are minimally successful
  - “Moratorium” on rating format research

- Rating format research fell out of favor in I/O

- Landy (2009)
  - Moratorium lifted
Conclusions regarding the lack of usefulness of rating format research are based almost entirely on the presence of psychometric "errors" in the ratings (DeNisi, 1996).

Rating "errors" are poor indicators of the quality of ratings (Fisicaro, 1988; Murphy, 2008; Nathan & Tippins, 1990).
Rating “errors”

- Errors were most frequently used criteria when evaluating performance ratings for most of the 20th century (Austin & Villanova, 1992)
  - Leniency
  - Severity
  - Central Tendency
  - Halo
Halo "Error"

- Thorndike (1920)
  - A rater’s favorable or unfavorable impression of a ratee leads the rater to rate all aspects of performance consistently with this overall impression

- Halo often confused with *logical error*
  - A rater’s tendency to rate similarly dimensions that he or she perceives as conceptually similar or logically related (Guilford, 1936)
Halo “Error” Research

- Relationship between “errors” and accuracy are weak and sometimes even positive (Becker & Cardy, 1986; Cooper, 1981; Murphy & Balzer, 1989)

- Halo can actually lead to higher levels of criterion-related validity in ability measures (Nathan & Tipps, 1990)
Attempts to remove halo have generally failed to control halo or increase the quality of ratings (Murphy, Jako, & Anhalt, 1993)

Problems with halo as a dependent measure (Balzer & Sulsky, 1992)

- No agreed upon conceptual definition
- Conceptual definitions are not related to operational definitions
- Halo measures are not strongly correlated with each other or rating validity or accuracy
Measuring Halo “Error”

- Small variances or standard deviations in ratings
- Large interdimension correlations
- Significant rater x ratee interaction term
- Dimensions load on a single factor
- Statistically controlling for overall rating
- Average rater interdimensional correlation exceeds average expert interdimensional correlation

- Which one do we choose?
Conclusions Regarding Halo

- All operational definitions are insufficient for diagnosing halo (Balzer & Sulsky, 1992)
- Halo “error” is based on erroneous assumption (Murphy, Jako, & Anhalt, 1993)
  - How do we know “true” levels of performance?
- Thorndike’s (1920) conceptual definition implies causality
  - None of the operational definitions model this
Why Research Rating Formats?

- Research on halo calls into question the conclusions of an entire body of rating format research dismissed by Landy & Farr (1980)

- Contemporary research suggests rating formats DO matter!
Rating Formats and Rating Validity

- Forced-choice formats resulted in higher validity coefficients than Likert rating scales (Bartram, 2007)
  - Multinational samples from 29 studies
- Computer adaptive rating scales evidenced higher reliability, validity, and accuracy than BARS or graphic rating scales (Borman, Buck, Hanson, Motowidlo, Stark, & Drasgow, 2001)
Absolute vs. Relative Methods

- Relative ratings were more accurate than absolute ratings *(Wagner & Goffin, 1997)*

- Relative format resulted in higher validity coefficient than absolute format *(Goffin, Gellatly, Paunonen, Jackson, & Meyer, 1996)*

- However, absolute rating formats were perceived as more fair than relative formats *(Roch, Sternburgh, & Caputo, 2007)*
Influence of Individual Differences

- Field independent raters provided more accurate ratings than field dependent raters using holistic formats (Hartel, 1993)
New Formats

- **Frame-of-reference (FOR) scales** (Hoffman, Gorman, Blair, Meriac, Overstreet, & Atchley, 2012)
  - Based on principles of FOR training
    - Create a common conceptualization of performance among raters (Gorman & Rentsch, 2009)
  - Presents dimension definitions and examples of positive and negative behaviors within each dimension
  - Rating formats rarely considered in 360-degree rating research
Example FOR Scales

APPENDIX A

Problem Solving
- Problem solving involves understanding problems and making appropriate decisions to resolve these problems. Effective problem solving entails gathering pertinent information, recognizing key issues, basing decisions on sound rationale, and considering the implications of one’s actions.
- Ineffective problem solving occurs when a manager does not attempt to gather relevant information, makes premature decisions, or confuses details of a given problem.
- At work, he/she
  1. Searches for additional information in order to identify the cause of problems. 1 2 3 4 5
  2. Considers multiple solutions to problems. 1 2 3 4 5
  3. Explicitly provides rationale for his/her decisions 1 2 3 4 5

Interpersonal Sensitivity
- Interpersonal sensitivity is defined as an individual’s concern for the feelings and needs of others. Effective interpersonal sensitivity occurs when a person works to build rapport with others, is attentive to others’ thoughts and feelings, and shows concerns for coworkers’ personal issues. Ineffective interpersonal sensitivity occurs when one is inattentive or alienates others.
- At work, he/she
  4. Treats others with dignity and respect 1 2 3 4 5
  5. Responds appropriately to the feelings of others 1 2 3 4 5
  6. Avoids interrupting others when they are speaking 1 2 3 4 5
FOR Scales Results

- **Study 1 (Field Study)**
  - 321 executives enrolled in MBA program
  - Resulted in cleaner factor structures, fewer inadmissible solutions, increased variance due to dimensions, decreased overlap among dimensions, and decreased error variance

- FOR scales potentially useful in 360 rating contexts

- **Study 2 (Lab Study)**
  - 151 undergraduate students
  - More accurate ratings than control condition
  - Rating accuracy results comparable to those of FOR training

- FOR scales potentially more practical and effective than full training programs
Current Research on FOR Scales

- Validity of FOR scale ratings
- FOR scales in administrative settings
- FOR scales for subordinate, peer, or client/customer ratings
- Fairness reactions to FOR scales
Conclusions

- Rating “errors” are poor indicators of rating quality
- Rating formats need to be evaluated using alternative dependent measures
- Research indicates there are substantive differences in the quality of ratings resulting from different rating formats
- Individual differences may moderate the effects of rating formats
Future Research Directions

- Individual differences and rating formats
- Rating formats in 360-degree contexts
- Combined effects of rating formats and rater training
- Rater and ratee reactions to various rating formats
- Equivalence of computer-based and paper-and-pencil rating formats