If an appraisal cannot differentiate between performance levels, then training/development and the assignment of rewards cannot be accurately carried out.

The "Ideal" Appraisal System  [Most systems in practice fall short of the ideal.]
1. Does your supervisor follow set (standardized) procedures?  The system should provide clarity, depth and forewarning.
2. Are the results fair and accurate?  Perceptions of fairness and accuracy can be enhanced by allowing employee participation. Further, there should be some balance (strong and weak points). Case examples can be used to document the ratings. One should avoid surprises.
3. Does the appraisal consider special circumstances? (e.g. acts of God, supplier strikes, etc.)
4. Do you get feedback? Feedback is most effective when it is given near the time the event.
5. Does your supervisor suggest ways to improve performance?
6. Does the supervisor provide the resources to improve performance?
7. Does management care about how the appraisals are done? The appraisal should be timely. There should be some depth of detail. The appraisal should also capture the entire year.
Performance Appraisal Instruments

Before we review numerous appraisal methods, you should keep in mind two cautionary thoughts:
1) One should not blindly search for the perfect method and forget about coming to grips with the underlying social situation. For example, one should consider why managers inflate appraisals.
2) Perhaps we should spend less time on the format and more time on training the raters (Bernardin and Beatty, 1984:235)

**Graphic Scales** = The most popular method currently. It is more prevalent in lower level jobs. It is inexpensive to develop and use (Dessler, 2011: 187). The dimensions are very broad. Little useful feedback is given. A scale may be used with a range from "unsatisfactory" to "outstanding." A number of broad job dimensions are generally rated (broad generic categories): such as, directing, expression, knowledge, attitude, planning and organization.

Concerns: Leniency or Rating Inflation - For example, on a seven point scale, 80% of the ratings are normally 6 or 7 (Murphy and Cleveland, 1991:190).

Ambiguity = What do the terms mean? Is there too much room for interpretation? What do the rating categories mean? Raters standards vary just like teachers' grading standards vary.

Most experts feel you should stay away from undefined abstract "traits" (Dessler, 2011: 187) and focus your attention on past performance. Negative labels tend to stick. One should think in terms of observable behaviors. One should also recognize the context of the behavior.

**Behavioral Observation Scales (BOS)** = more specific rating scales and statements than graphic scales.

<table>
<thead>
<tr>
<th>Almost Never</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Almost Always</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-64%</td>
<td>65-74%</td>
<td>75-84%</td>
<td>85-94%</td>
<td>95-100%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Personally reviews all invoices for correct billing and merchandise received.
- Takes inventory monthly, checks discrepancies and promptly corrects any item requiring correction.
- When there is a conflict among employees, takes the time to sit down and discuss the causes and resolution.

Concerns = The added detail requires more time and money as far as construction is concerned. One could still rate everyone highly. Can anyone actually rate the percentage of the time one does things like this? Perhaps these instruments create a false sense of accuracy.

**Electronic Performance Monitoring (EPM)** = Electronically monitoring the employees’ computer terminals and telephones. The length of calls, number of calls, amount typed, and time taken on breaks may be computed. This increases stress and may decrease performance (Dessler, 2011: 181). An estimated 26 million US workers are subject to EPM (Dessler, 2005: 327).

**Essay Appraisals** = Narrative summaries that can potentially capture strengths, improvements, weaknesses, quantity and quality. Essays are inexpensive to develop, but expensive to use. No real
comparisons of employees are possible unless common dimensions are inserted and the raters trained. Manuals of common buzz words exist. Essentially, only the newcomers are deceived.

In-depth essays are extremely time consuming to apply with large numbers of employees. One Bus Line replaced a 100 point merit system with a new essay appraisal system. The supervisors did not welcome this, so 890 copies of a description of the "ideal driver" were turned in with the names of different people typed in on the top of the form, thereby subverting the effort.

There are a number of comparative methods aimed at making sure everyone is not rated highly.

**Ranking** = One can require an appraiser to order subordinates from lowest to highest. One problem with this method is there is no distance gauge, as in the diagram. Further, no feedback is given, so it is normally used in conjunction with other methods.

**Forced Distribution** = The raters are required to assign a specific proportion of their employees to predetermined categories (a quota system). About 25% of Fortune 500 companies use forced distribution (Dessler, 2006: 189). Once again, little information is generated for development purposes. Moreover, this system is rigid. What about differences in the quality of departments? What about differences in a group’s quality over time?

**Critical Incident Method** = The rater records incidents reflecting behavior that represents less than satisfactory or outstanding performance for both evaluation and feedback. These events should be critical elements of job performance. Supervisors prefer recording events daily, because it is easier to recall events. The specificity of the feedback is a big plus, but it is very time consuming.

Concerns = Employees often feel supervisors are more prone to note the negative. Employees may regard it as similar to a "little black book." The employees may also object because they face unequal challenges (for example, a customer complaint from Ms. Congenial vs. the Wicked Witch of the West). One might also be concerned about when the supervisors are watching. The uncertainty is troubling as is the unknown criteria. One difficulty is, absent some numerical ranking, it is hard to compare employees and make salary decisions (Dessler, 2011: 186).

**Checklists** = A set of behaviors, adjectives, or descriptive statements. The rater checks if it describes the person. The items are simply present/absent. Weightings can be attached for each job.

**Position:** PROGRAM AUDITOR  
**Job Dimension:** DATA ASSEMBLY/ORGANIZATION  
_____ 1. Unable to separate important from irrelevant information.
1. Omits important information from workpaper summaries.
2. Cross-indexes to facilitate referencing of reports.
3. Produces workpaper summaries, which can be readily transformed into a report.
4. Requires excessive amount of instruction to produce summaries.
5. General problems are identified, but not the degree of the extent of the problem. Some time and expense is involved in developing a good one, yet everyone could still be rated highly.

Forced Choice (Army) = Small groups of statements are bunched (3 to 5) that are related to a specific job. The rater selects the items that are most and least descriptive, which prevents rating inflation. It is costly to develop and the categories are not very specific compared to BARS.

Behaviorally Anchored Rating Scales (BARS) = A continuum of 5 to 9 points which has incidents/behaviors attached reflecting the employees relevant job dimensions. It has sometimes been described as a mixture of graphic scales and the critical incident method. The quality of the feedback should be high if the people who will apply the instrument develop it.

Concerns = It is process oriented, so end results are ignored. Are the incidents covered observable? It is time consuming to develop, easy to use. Everyone may be rated highly, so it may be hard to justify the added time and expense versus graphic scales. The actual scales are sometimes questionable. Some of the incidents seem like apples and oranges. A common situation should be posed, and the behavioral options should be mutually exclusive (i.e. you cannot do two of them).

Management by Objectives (MBO) = This is an end result orientation advocated by Peter Drucker. It is used in many upper level jobs. Some firms set organizational and departmental goals before specific superiors and subordinates start the process described below (Dessler, 2011: 180).

Step 1 = Objectives are jointly set. Employee participation should enhance motivation.
Step 2 = Coaching and progress reviews occur over the course of the appraisal period.
Step 3 = Actual results are compared at year end with the predetermined objectives. Merit pay is then applied.

Concerns = If the objectives are set too high, frustration will result. If they are too low, inappropriate rewards follow. The method fails if the employees just write up what management wants to hear (no buy in). Sometimes the employees understate what they can do (do not lead your superiors to expect too much) and report their progress with a lag (to provide security). To the motivated, it will seem like a lot of unnecessary paperwork. However, it forces the unmotivated to set targets. People often focus on what is easy to count (criterion deficiency). The end result orientation creates concerns about managers "milking" or "starving" an operation to receive short term rewards and recognition (hoping to outrun mistakes or locate a scapegoat).

Problems, Concerns, and Criticisms

* Psychologists use the term "relevance" to refer to the relationship between the performance rating and some measure of an ultimate criterion. "Deficiency" refers to variance in the ultimate criterion
not captured by the performance rating (factors left out). For example, MBO leaves out what is not easily quantified. "Contamination" refers to variance in the performance rating that is unrelated to the ultimate criterion. For instance, one should consider factors like the state of the economy, region, sex, race, as well as matters like opportunity bias (e.g. sales volume and region).

* Self Comparison = A supervisor might say "When I was your age, I ...." Since each supervisor’s past experiences are different, each is using a different ruler, which means low reliability. If this is combined with intolerance of ways that are different, the situation is made even worse.

* Proximity Errors = Does the employee the supervisor knows the best get rated differently? Does familiarity breed contempt? (i.e. The supervisor cannot get all of the growing pain errors out of his/her mind) Does the best known employee get the highest rating? Both are proximity errors.

* Halo/Horns Error = This is a somewhat different matter than in employment interviewing. This has been labeled as the biggest source of error (Bernardin and Beatty, 1984: 173). The rater only notes the positive aspects of the best employee and the negative aspects of the worst employee. Ideally, the rater would see the positive and negative aspects of both.

* Assessments require two types of data = In order to make pay decisions, you need broad comparative aggregate rankings. Frequently this is all that is generated. For development, you need the specific positives and negatives of each individual.

* Defensive reactions to negative parts of the appraisal (denial by the employees - a poor carpenter blames his tools). Most people believe they are above average and the truth hurts. In some instances, no one deserves top ratings. Will managers be honest if they know the message will not be well received? Directly attacking a person’s defenses is generally not a productive way to proceed, because it only heightens the intensity of the reaction (Dessler, 2011: 188).

Supervisors fear the following self-reinforcing cycle.

![Diagram of self-reinforcing cycle](image)

Managers often avert this by rating everyone highly (the sunshine foremen), or using "code words" to fool subordinates. The "sunflower effect" refers to the need to give superiors information that will not embarrass the appraiser ("None of my men is less than 100 percent or I would not keep him.").
* Some managers see appraisal and development as a secondary concern. Most organizations have no formal training on how to carry out the appraisal interviews. These are not inherent skills. Furthermore, the supervisors’ rewards are seldom contingent on the appraisal process (no measurement normally occurs). For example, only 22% in a survey of Fortune 100 companies reported evaluating the evaluator on appraisals (Heneman, 1992: 212).

* Are appraisals self-defeating? = TQM advocate Edward Deming argued it was the system that produced more of the variation in individual outcomes more than differences in motivation, skill or effort. If the variations in employee performance are beyond the control of the individual workers, employing a performance appraisal system that allocates rewards and promotions, as well as triggering development plans, could be quite counterproductive. This focus places emphasis on the individual parts rather than the relationship of the parts. Deming accordingly argued for the elimination of traditional performance appraisal. He believed in focusing instead on training, communication tools and supervision (Dessler, 2004: 194). Do you believe he is right or wrong?

* The Performance Management Approach = The polar opposite of this would be to identify the bottom 10% of your current employees, and tell them they will either improve in say 90 days or they will lose their jobs. At some companies, if they are still in the bottom 10% after 90 days, they can either resign and receive severance pay, or they will be fired (Dessler, 2011: 174). Sun Microsystems, Enron, Ford and GE have used some form of this approach. Is this appropriate?

Some Possible Improvements (These may have drawbacks)

(1) Setting Clear Goals = Subordinates can only be expected to manage their own performance effectively if they know what is expected of them. Many supervisors are lax in setting clear goals. Some organizations use the “SMART” method (Dessler, 2011: 170-72). Under this system goals should be specific, measurable, attainable, relevant, and timely (deadlines, milestones, etc.). Organizations such as General Electric, Kellogg’s, and Microsoft, have used this method.

Regardless of whether SMART is used, employees need to have clear expectations. Otherwise, it will be not be possible to later hold them responsible for poor performance.

(2) Multiple Raters = Utilizing multiple raters should reduce bias. Multiple perspectives are brought to bear. Views must be justified and documented. People see the person at different times. They also may have different areas of expertise. One might include superiors, peers or subordinates. A rater must have substantial exposure to be meaningful. Peers should be able to sympathize with the situation of the person being rated. Peers might be in competition for promotions and raises. This could produce unfairly low ratings. Alternatively, some peers may make a pact to give each other high ratings (Dessler, 2005: 332). Subordinates may cause superiors to be soft/buy love. Subordinates will tend to inflate ratings unless they are provided with anonymity (Dessler, 2011: 173). Low levels of performance can be enhanced via discussions with subordinates, but this makes it difficult to maintain anonymity. Perhaps subordinates can be used for development only. General Electric (G.E.) has explored the usage of 360-degree feedback. In addition to superiors, peers and subordinates, one might include internal and external customers. One would only use such a system where employees have important multiple contacts, and it is deemed worth the expense. GE eventually backed off this method.
Multiple evaluation methods = It is possible to use multiple evaluation methods to get around the limitations of the individual methods. For example, MBO and BARS, or graphic scales and essays. Obviously the added time and expense need to be noted. The "best" methods are the most time consuming and expensive (Bernardin and Beatty, 1984: 233).

Improving the feedback process = The pay interview and the development interview(s) should be separated so people will listen to advice. G.E. waits three months to talk about development. More frequent development interviews might be used (perhaps 3 to 4 per year). Recently, some people have endorsed “continuous appraisal and development.” How often can this be done? One can conclude with a no surprise summing up for clarity sake.

It is vital that the employee sign a copy of his/her appraisal. This documents that it has been received. If a dispute later arises, the organization may not be able to use unsigned appraisals as evidence with the EEOC, a court or in labor arbitration.

Trying to Reduce Employee Denial (Defensive Reactions) = The appraiser can ask the employee where he/she thinks he/she need help (self criticism). Under the 3 to 4 development meetings a year model this may hold promise. This cannot really be done near pay appraisal time. For example, an airline had both superiors and subordinates fill out an appraisal form and this heightened conflict as both the bosses and the subordinates defended their positions. Subordinates normally do rate themselves higher than supervisors or peers rate them (Dessler, 2011: 173).

What else can be done to overcome resistance to negative feedback? People can accept more from friends. It takes time to establish this type of rapport. One can be honest and start with a positive observation. The appraiser can provide both some balance of positive and some negative feedback in the appraisal. The problem can be stated in terms of work expectations, standards and timeframes, and avoid personalized condemnations. The employee should also have a chance to talk. The supervisor can offer to help with the negatives by providing suggestions, but the responsibility for development remains with the employee. The supervisor needs to get the employee to agree there is a problem. Further, an agreement needs to be reached on what needs to be done and by what date this should be accomplished (Dessler, 2011: 188). Do not try to correct too much at one time. One should set limited objectives, define a time frame and reward improvement. If the improvements are not forthcoming, then disciplinary steps up to and including discharge should be implemented.