# Evaluación del alumnado mediante la medida relativa de sus resultados: el método de la Campana de Gauss

**Francisco Javier Quirós Tomás**
Universidad de Sevilla
quiros@us.es

**Resumen**

Existen múltiples formas de evaluar a los alumnos. Pese a que se ha discutido mucho al respecto, el objetivo del presente trabajo es presentar un enfoque poco tratado: la calificación mediante métodos de medida relativa de los resultados versus el clásico empleo de métodos de evaluación de medida absoluta. Más concretamente, se presenta el método de la Campana de Gauss. Se discuten sus ventajas e inconvenientes, siendo su principal ventaja que permite reducir los problemas debidos a una baja fiabilidad al emplear formas paralelas en la evaluación de distintos grupos de alumnos.

**Palabras clave:** *Evaluación, fiabilidad, campana de Gauss, curva de la vitalidad.*

## 1. INTRODUCCIÓN

La evaluación del alumnado es una parte fundamental del proceso de formación del mismo, pues permite conocer en qué medida los conocimientos han sido adquiridos por los estudiantes y por tanto la utilidad del proceso formativo en sí mismo considerado. Por otra parte, sus esfuerzos formativos y los efectos de los mismos sobre sus conocimientos, habilidades, aptitudes y sobre todo capacidades deben ser reconocidos y poder ser acreditados ante terceros interesados, siendo para ello imprescindible la evaluación. Por todo ello es necesario un adecuado y fiable proceso de evaluación del aprendizaje de la materia objeto de la formación.

Desafortunadamente, reconocer la importancia de la evaluación del aprendizaje y su adecuada medición son dos cosas distintas. Idealmente, la medición debería tener en cuenta índices objetivos, personales y juicios de valor. Desafortunadamente, como indican Landy y Farr (1980) es difícil obtener incides objetivos y los datos personales también son difíciles de aplicar, motivo por el cual los evaluadores tienden a recurrir en gran medida a juicios de valor.

Existen múltiples sistemas de evaluación que se aplican en diferente medida por los docentes a la hora de calificar a sus alumnos. El ánimo del presente trabajo no es analizar el cómo evaluar (evaluación continua, evaluación final, tipo test, examen de desarrollo, etc.), ni la adquisición de conocimientos (el qué evaluar), sino dar a conocer un método para reducir la subjetividad asociada a todo sistema de evaluación incrementando la fiabilidad del proceso.

La subjetividad en la evaluación tiene su origen en la participación de evaluadores en el proceso de análisis del aprendizaje por los alumnos. Su participación hace que sus juicios de valor tiñan, en mayor o menor medida, la medición y la exactitud de los resultados del proceso evaluativo. Es muy común que aparezcan una serie de errores típicos como el efecto halo, el error de proyección, la existencia de estereotipos, los errores de limitación de escala, así como muchos otros que han sido analizados en profundidad por numerosos autores. De entre ellos sobresalen los desarrolladores de la Teoría de las Perspectivas (Prospect Theory)

Daniel Kahneman, que recibió el Premio Nobel de Economía el año 2002 por dichos estudios, y su coautor Amos Tversky. Estos autores han escrito obras como "Prospect theory: An analysis of decision under risk", citado en más de 45.000 trabajos posteriores (Kahneman & Tversky, 1979). La subjetividad en el caso concreto del empleo de métodos basados en rankings para medir la valoración del desempeño ha sido analizada por autores como Wherry & Bartlett (1982). Estos autores diferenciaron entre errores de medición y errores de recuerdo y analizaron la forma de reducirlos.

Existen múltiples formas para reducir la subjetividad asociada a la actividad del evaluador. Se basan en la reducción, en la medida de lo posible, de las ocasiones e intensidad con que éste debe emitir juicios durante el proceso evaluador. Para ello se puede actuar en diversas áreas, como el formato de las pruebas, su contenido o la calificación.

Como ejemplo del primer caso, el formato de la prueba, un método cada vez más extendido para reducir la subjetividad es el empleo del tipo test a la hora de examinar a los alumnos, frente a los clásicos exámenes de desarrollo donde el juicio de valor del evaluador es más importante y, por tanto, es más probable que aparezcan errores debidos a la subjetividad.

Otra forma de reducir la subjetividad, en este caso asociada al contenido de las pruebas, es la realización de exámenes amplios en los cuales se analice la práctica totalidad de los conocimientos que el estudiante debe adquirir en una materia concreta, evitando así el efecto suerte asociado al conocimiento parcial de la materia, de la que es un ejemplo típico la conducta de muchos estudiantes de "esto no me lo estudio porque seguro que no cae".

Una tercera forma de reducir la subjetividad, relacionada en este caso con la forma de calificar la adquisición de conocimientos y capacidades a lo largo de los estudios es el recurso a la medición relativa de los resultados, en contraposición a la absoluta. Esto permite aumentar la fiabilidad de formas paralelas y entre evaluadores. Existen múltiples métodos para ello, siendo uno el de la Campana de Gauss, que es el método que se analiza en el presente trabajo. Este método es conocido con otros nombres, entre los que destacan el de Curva de la Vitalidad, Stack Ranking o Rank and Yank, siendo ampliamente empleado en la evaluación de trabajadores por empresas de primer nivel.

## 2. LA FIABILIDAD DE FORMAS PARALELAS Y ENTRE EVALUADORES

La fiabilidad, según el diccionario de la Real Academia de la Lengua, en su segunda acepción, consiste en la “probabilidad de buen funcionamiento de algo”. Igualmente define como fiable aquel método “que ofrece seguridad o buenos resultados” o “creíble, fidedigno, sin error” en sus segunda y tercera acepciones respectivamente.

Las definiciones anteriormente citadas implican que el sistema de evaluación será fiable si las mediciones que suministra son consistentes, esto es, si proporcionan el mismo valor (calificación) para cada alumno, siempre y cuando las características o requisitos que está midiendo (conocimientos) no haya cambiado (Alcaide et al, 2011). Este es un requisito indispensable si se quiere ser objetivo a la hora de calificar al alumnado.

Dado que existen múltiples técnicas de evaluación y a la participación en el proceso de diversos evaluadores, en la práctica se puede hablar de la existencia de distintos tipos de fiabilidad en función de los procedimientos seguidos para su análisis (Quirós, 2015). Entre ellas podemos citar:

- Fiabilidad de reaplicación o fiabilidad test-retest
- Fiabilidad de juicio
- Fiabilidad entre evaluadores
- Fiabilidad de la medida de similitud
- Fiabilidad interna o fiabilidad de las agrupaciones
- Fiabilidad entre formas paralelas

De ellas, las que más nos interesan en el presente trabajo son la fiabilidad entre formas paralelas y la fiabilidad entre evaluadores, al ser las directamente afectadas por el empleo del método de la Campana de Gauss. El recurso a este método busca, básicamente, aumentar ambos tipos de fiabilidad, permitiendo con ello una mejor evaluación de los estudiantes.

El análisis de la fiabilidad entre formas paralelas consiste en estudiar la correlación entre resultados obtenidos en dos pruebas similares (paralelas) por un mismo grupo de sujetos, normalmente dejando entre una y otra un corto periodo para evitar cambios, pero suficientemente largo como para evitar la fatiga de los sujetos (Díaz et al, 2003). Para ello se requiere la realización de dos pruebas similares, es decir, usando la misma técnica pero con un contenido distinto, de forma que se mida la misma variable, obteniendo el coeficiente de equivalencia entre las mismas (Alcaide y González, 1997).

Es muy común, aunque es algo que todo profesor procura evitar dentro de lo posible, que por razones prácticas se diseñen dos pruebas diferentes para medir el aprendizaje de los alumnos. Entre los motivos se pueden citar la falta de espacio para celebrar la prueba a la vez por el conjunto de los estudiantes, la existencia de diversos grupos de alumnos evaluados en lugares y momentos diferentes como por ejemplo alumnos de grado y de doble grado en ciertas titulaciones, la insuficiencia de vigilantes para realizar una prueba única, la coincidencia de fechas entre pruebas de diversas asignaturas que obliguen a realizar una convocatoria extraordinaria o la simple sucesión de convocatorias en el tiempo, etc. En todos estos casos lo normal es que cada una de ellas conlleva sus propias pruebas, paralelas entre sí pero no iguales.

En estos casos, si se quiere que las evaluaciones obtenidas por dos alumnos que han sido sometidos a pruebas paralelas sean realmente comparables se debe exigir una elevada fiabilidad de las formas paralelas, algo que no siempre ocurre. El método de la Campana de Gauss ayuda a aumentar la fiabilidad en el caso de usar formas paralelas al evaluar a los estudiantes.

La fiabilidad entre evaluadores, por su parte, es importante cuando la evaluación del alumnado requiere del juicio de los evaluadores, participando varios de ellos en el proceso. En este caso, cada uno de ellos empleará sus propios juicios, siendo necesario que sigan unos criterios comunes de forma que los resultados no se vean afectados por cuál de ellos sea el que realice la evaluación de un alumno concreto. Esto se consigue mediante una alta fiabilidad entre evaluadores. Esta se mide mediante la correlación entre las valoraciones emitidas por cada uno de ellos respecto a los conocimientos mostrados por cada uno de los alumnos al realizar una misma y única prueba de evaluación (Alcaide y González, 1997).

Los problemas de falta de fiabilidad entre evaluadores suelen presentarse cuando cada profesor evalúa a distintos grupos de alumnos siguiendo su propio juicio al respecto. Esto es cada vez más común con la introducción de la obligación de recurrir a sistemas de evaluación continua, donde, pese a los intentos de coordinación en las asignaturas, es frecuente que cada profesor califique a sus propios alumnos, con el consiguiente aumento de la importancia de una elevada fiabilidad entre evaluadores. Ante unos mismos conocimientos, lo que un evaluador considera, por ejemplo, suficiente o notable en una escala de calificación puede ser muy diferente de lo que considere otro. La falta de fiabilidad entre evaluadores puede llevar a una falta de homogeneidad en las calificaciones llevadas a cabo por diversos evaluadores. Como en el caso anterior, el método de la Campana de Gauss puede ayudar a paliar este problema.

## 3. SISTEMAS DE EVALUACIÓN ABSOLUTOS Y RELATIVOS

A la hora de medir cualquier variable, en este caso el aprendizaje de los alumnos, se puede acudir a dos tipos de mediciones:

- Mediciones absolutas
- Mediciones relativas

En el primer caso, la medición absoluta, las cifras obtenidas son independientes para cada uno de los casos a analizar, siendo por tanto el resultado independiente del obtenido por los demás sujetos. Aplicándolo al caso de la evaluación de conocimientos, las notas o calificaciones obtenidas son independientes para cada uno de los alumnos, no guardando relación, por tanto, el resultado obtenido por cada uno de ellos con el de los demás, o lo que es lo mismo, cada evaluado obtiene su nota en función de sus propios méritos. Éste tipo de medición se basa en el establecimiento de un juicio absoluto sobre la formación obtenida por cada participante.

La medición se realiza dentro de una escala de valoración, de forma que cada medición se hará siguiendo unos criterios o estándares previamente establecidos, en función de los cuales se otorgará una valoración a cada sujeto de forma que se puedan comparar los valores obtenidos por los mismos. En el caso de la medición del aprendizaje en la universidad española se emplea una escala 0 - 10, estableciéndose los sistemas de evaluación en los correspondientes programas de las asignaturas, en función de los cuales, tras la oportuna celebración de las pruebas previstas, se otorgará la calificación a cada alumno según sus méritos, pudiendo compararse los resultados obtenidos por los diversos alumnos. Igualmente, está previsto legalmente que las calificaciones se agreguen en grupos más o menos homogéneos, que van desde el suspenso a la matrícula de honor.

En el segundo caso, el de la medición relativa, la valoración obtenida por cada uno de los sujetos no tiene un carácter absoluto, sino que se expresa como un valor en relación al conjunto de los mismos. Ello implica que la evaluación no es independiente para cada uno de ellos, como en el caso de la medición absoluta. Los resultados de un estudiante concreto dependen del desempeño del resto de los alumnos tanto como del correspondiente a sí mismo. Por tanto, no existe un estándar previo con el cual comparar a cada uno de los evaluados, sino que sus resultados vendrán dados por su posición relativa respeto al resto de los estudiantes.

Existen múltiples métodos de medición relativa, destacando dos grandes tipologías:

- Métodos de jerarquía
- Métodos de distribución forzosa

Los primeros se basan en establecer un orden entre los sujetos o jerarquía clasificatoria y los segundos en determinar por anticipado una serie de grupos más o menos homogéneos según los resultados relativos a obtener por los participantes, preestableciendo la cantidad o porcentaje de ellos que pertenecerán a cada grupo por anticipado. El método de la Campana de Gauss pertenece a ésta última tipología.

Por tanto existen dos formas claramente diferenciadas de clasificar. Para su mejor comprensión se empleará el ejemplo de la clasificación deportiva. Los métodos basados en medición absoluta son los empleados en el deporte para establecer las marcas (personales, del país, del mundo, olímpicas, etc.). Éstas dependen del valor absoluto alcanzado por el deportista en una prueba concreta. Por otra parte, siguiendo con el ejemplo deportivo, un método basado en la clasificación relativa sería la posición que un deportista alcanza respecto al resto de los participantes en una prueba concreta o en un campeonato (por ejemplo medallas de oro, plata y bronce, diploma olímpico, finalista o semifinalista en las olimpiadas). En este caso no es importante la marca obtenida, sino que esta sea superior o inferior a las de los competidores en la prueba correspondiente.

La medición absoluta es el sistema que impera en general en evaluación universitaria, pero los sistemas basados en la medición relativa tampoco le son ajenos. Así, en el caso de la Universidad de Sevilla, según el artículo 23.2 de la Normativa Reguladora de la Evaluación y Calificación de las Asignaturas: “El número de menciones de "matrícula de honor" que se pueden otorgar no podrá exceder del cinco por ciento de los alumnos incluidos en la misma acta oficial, salvo que éste sea inferior a 20, en cuyo caso se podrá conceder una sola". Esto no es más que una aplicación parcial de medición relativa mediante distribución forzosa. En este caso habría dos grupos donde clasificar a los alumnos, uno formado por los que obtengan matrícula de honor (porcentaje máximo del 5%) y un segundo grupo formado por los alumnos con el resto de calificaciones (porcentaje mínimo del 95%). Así, se puede obtener una Matrícula de Honor sin necesidad de obtener un 10 (la normativa requiere únicamente la obtención de sobresaliente) o no obtenerla habiendo conseguido las máximas clasificaciones posibles (más del 5% de alumnos con calificación de 10). El resultado final dependerá de la posición relativa del alumno respecto a sus compañeros.

Estos sistemas también se emplean en otros múltiples casos en el funcionamiento propio de la Universidad, como puede ser el caso de oposiciones, las notas de corte de selectividad para la elección de carrera, etc., donde los importante es la posición relativa respecto al conjunto de participantes y no la valoración absoluta.

## 4. EL MÉTODO DE LA CAMPANA DE GAUSS

El método de la Campana de Gauss es un método de distribución forzosa ampliamente utilizado en el campo de los Recursos Humanos con el fin de evaluar a los trabajadores. Recibe su nombre del matemático Johann Carl Friedrich Gauss. Éste es uno de los mejores matemáticos de la historia, habiendo quedado su nombre asociado a la distribución normal o, como también se la conoce, distribución de Gauss o distribución gaussiana. La representación gráfica de la misma da origen a su vez a la conocida como Campana de Gauss, que es la base del método que se va a analizar y al cual da nombre.

Este método, en su uso en la evaluación de los recursos humanos, también recibe otros nombres como curva de la vitalidad, stack ranking o rank and yank.

Antes de proceder a su análisis, en el epígrafe siguiente, se procederá a describir la distribución normal, con el fin de comprender los fundamentos del método.

*4.1 Distribución Normal.*

Una variable aleatoria continua X se dirá que sigue una distribución normal $N(\mu, \sigma^2)$ si su función de densidad es la siguiente:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Ésta distribución posee unas características que hacen interesante su empleo en el caso que tratamos. Siendo media μ y su sus varianza y deviación típica $\sigma^2$ y σ, respectivamente, de entre sus propiedades se pueden destacar las siguientes (Voda, 2009):

- Es simétrica respecto a su media.
- Media, moda y mediana son coincidentes.
- Su distribución de probabilidad entorno a la media es la siguiente:
    - El intervalo [μ-σ, μ+σ] comprende en torno al 68,26% de la distribución
    - El intervalo [μ-2σ, μ+2σ] comprende en torno al 95,44% de la distribución
    - El intervalo [μ-3σ, μ+3σ] comprende en torno al 99,74% de la distribución

Estas propiedades hacen que su representación gráfica adopte la forma recogida en la Figura 1. Como se puede observar, la distribución adopta aspecto campaniforme, origen de su nombre.

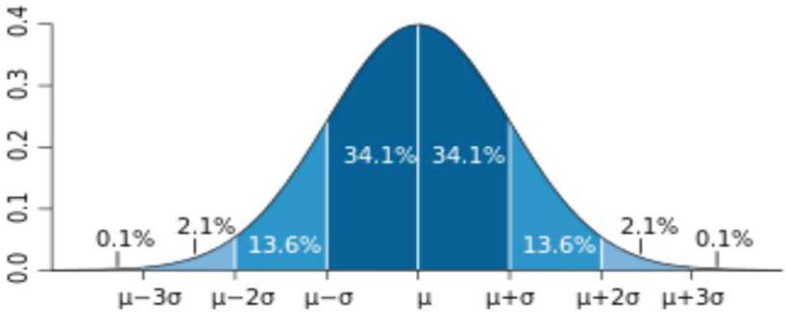



*Figura 1: Campana de Gauss. Fuente: Internet*

*4.2 Utilidad de la Función Normal: el aprendizaje y su evaluación.*

La importancia de esta distribución radica en que permite la modelización de múltiples fenómenos naturales, sociales y psicológicos, incluso si los mecanismos que subyacen a dichos fenómenos son desconocidos. Su empleo está muy extendido en numerosos campos científicos como biología (tamaño, peso), medicina (presión sanguínea, sintomatología clínica), física (intensidad de la luz láser), finanzas (rendimientos), economía (nivel de renta, consumo) o psicología (inteligencia, cociente intelectual).

Este es también el caso de la variable aprendizaje, que puede considerarse como una distribución normal. Existen múltiples modelos que tratan de explicar cómo se desarrolla el aprendizaje y la influencia de diversas variables sobre el mismo. Se puede citar a modo de ejemplo el "Modelo 3P" de J. Biggs (1987) con sus tres componentes principales Presagio, Proceso y Producto. Los diversos modelos recogen múltiples variables que condicionan en

mayor o menor medida el rendimiento académico de los estudiantes en general y de los universitarios en particular, resultando muy complicado dilucidar y ponderar la influencia específica de cada una de ellas (López et al., 2007). Schunk (1997) distingue en su libro "Teorías del aprendizaje" entre variables asociadas al medio y asociadas al estudiante. Entre las del primer tipo se encuentran, entre otras, la disposición y presentación de los estímulos, el modo de reforzar las respuestas o el historial de reforzamiento y entre las asociadas al estudiante el estadio de desarrollo del individuo y los pensamientos, creencias, actitudes o valores de los alumnos.

Ante tal cúmulo de variables y la falta de acuerdo a la hora de establecer la importancia relativa de cada una de ellas, la distribución normal permite modelar el aprendizaje. Esto es posible aun desconociendo los mecanismos que subyacen a dicha variable, si se asume que cada observación se obtiene como la suma de una serie de causas independientes. Con ello es posible describir el fenómeno, aunque no obtener una explicación del mismo.

Se puede partir de la base de que, independientemente de los factores o variables que afecten al rendimiento académico, las calificaciones tenderán a repartirse siguiendo una distribución normal. Esto implica que habrá un conjunto grande de alumnos con una notas parecidas a la media (el intervalo $[\mu-\sigma, \mu+\sigma]$ o media más-menos desviación estándar, que comprende en torno al 68,26% de la distribución) mientras que cuanto más nos alejemos de la nota media menos individuos nos encontraremos. Así, como se aprecia en la Figura 1, solo el 13,6% estarían a una distancia entre una y dos desviaciones típicas, tanto por la derecha como por la izquierda, un 2,1% a entre dos y tres desviaciones típicas y únicamente un 0,1% a más distancia a cada uno de los lados de la media.

### *4.3 Un ejemplo de aplicación exitosa del método de la Campana de Gauss: el caso General Electric y Jack Welch.*

La Campana de Gauss como medio de evaluación tiene un amplio recorrido histórico. Entre los casos más conocidos y exitosos de su aplicación a la evaluación se encuentra el caso de su empleo por la empresa General Electric en tiempos de su mítico CEO, Jack Welch, que fue elegido por la revista Fortune como el mejor manager del siglo XX (Fortune, 1999).

Welch, al que se le atribuye la invención de este modelo, lo aplicó durante muchos años en General Electric y sigue defendiendo su utilidad frente a las múltiples críticas que suscita. Así, por ejemplo, lo hizo el 14 de noviembre de 2013 en un artículo publicado por The Wall Street Journal (WSJ, 2013). Este modelo, que General Electric aplicaba a la evaluación de directivos, partía de la existencia de tres grupos de directivos a los que éstos eran adscritos en función de los resultados de su evaluación. Un primer grupo incluía a los mejores directivos, un segundo grupo era el formado por los directivos mediocres y el tercero y último era al que se asignaban aquellos con peor desempeño. Dichos grupos debían incluir un 20%, 70% y 10% de los directivos respectivamente, lo que lo convierte en ejemplo típico de modelo de distribución forzosa.

Los efectos de la clasificación en un grupo u otro eran terminantes. Los directivos que estaban en el primer grupo (excelente desempeño) ascendían y obtenían grandes bonus, los del segundo grupo (mediocres) no recibían ningún trato en especial, en tanto que los del tercer grupo (bajo desempeño) eran degradados o despedidos, tras darles la oportunidad de mejorar en un periodo de un año (El Confidencial, 2013).

Este método es empleado por muchas empresas en la actualidad. Por ejemplo, según una encuesta realizada por WorldatWork, en 2013 el 12% de las grandes corporaciones estadounidenses lo empleaba (WaW, 2013), este porcentaje ascendía al 29% según The Corporate Executive Board Company (TCBC, 2013) y al 60% para las empresas del Fortune 500 según en el año 2012 según el consultor especializado Dick Grote (2013). Pese a lo extendido de su uso, según CEB Inc., en 2015 un 6% de las empresas que lo usaban lo habrían abandonado (CEB, 2015). Este es el caso de Microsoft, que abandonó este método de evaluación de sus empleados tras el sonado fracaso de su implantación en el año 2011. Se estima que este fracaso fue una de las principales causas del despido de su CEO, Steve Ballmer en 2013.

*4.4 Aplicación de la Campana de Gauss a la evaluación del aprendizaje.*

La aplicación del método de la campana de Gauss a la evaluación del aprendizaje universitario pasaría por los siguientes pasos:

- Establecimiento de estándares de calificación
- Pre establecimiento de porcentajes de alumnos en función de los estándares
- Evaluación de los alumnos
- Distribución de los alumnos entre los diversos grupos establecidos al fijar los estándares, teniendo en cuenta los porcentajes prefijados y sus resultados académicos relativos.

El primer paso es el establecimiento de estándares de calificación. Para ello hay que decidir tanto su número como su amplitud (límites) y si esta va a ser homogénea o no. Dado que legalmente en el sistema educativo universitario español se emplea una calificación entre un mínimo de 0 y un máximo de 10, estos serían los límites máximo y mínimo a emplear para el establecimiento de los diversos grupos de clasificación. Dentro de dichos límites, el establecimiento de grupos podría hacerse de diversas maneras. a continuación se exponen algunas de ellas.

Una sería partir de la clásica división en cinco grupos según lo dispuesto en el Real Decreto 1125/2003, de 5 de septiembre, por el que se establece el sistema europeo de créditos y el sistema de calificaciones en las titulaciones universitarias, de carácter oficial y validez en todo el territorio nacional. Dichos grupos son: suspenso, aprobado, notable, sobresaliente y matrícula de honor. En este caso la amplitud de los mismos no sería homogénea, como se aprecia en la Tabla 1. La amplitud varía entre un máximo de 5 puntos en el caso del suspenso a un valor puntual en el caso de la matrícula de honor.

| **Calificación** | **Nota** |
|---|---|
| Suspenso | 0-4,9 |
| Aprobado | 5-6,9 |
| Notable | 7-8,9 |
| Sobresaliente | 9-? |
| Matrícula de Honor | ?-10 |

*Tabla 1: Grupos y calificaciones correspondientes. Fuente: Elaboración propia*

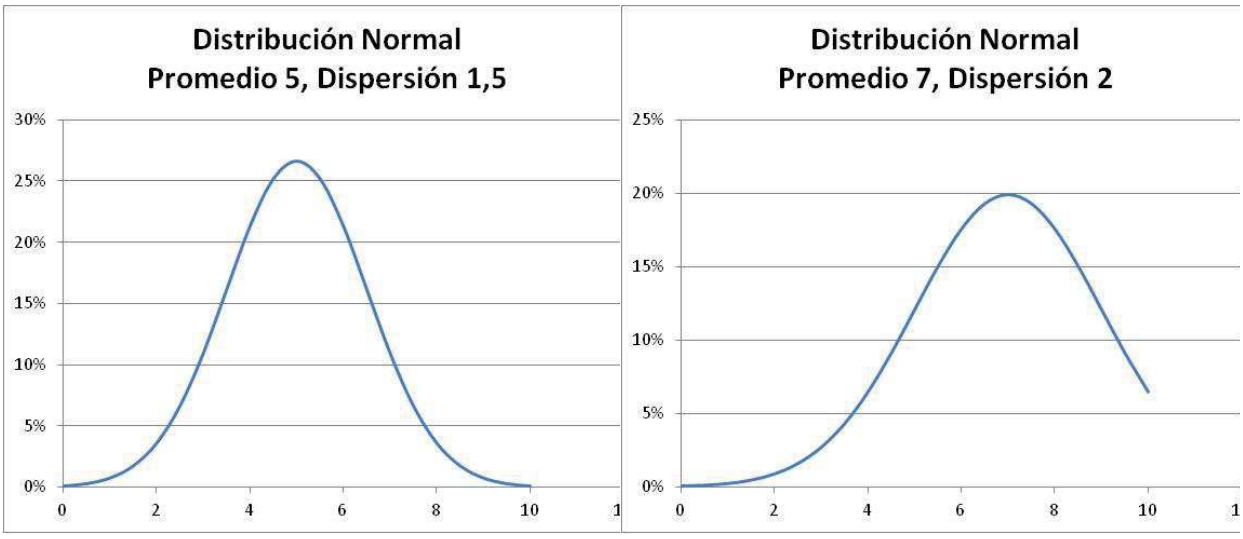
Naturalmente, el límite entre sobresaliente y matrícula de honor debería ser flexible para acomodarlo a la normativa vigente respecto a la obtención de esta última calificación ya comentada en el epígrafe 3.

Otra posible agrupación sería la que surge de aplicar la máxima variedad de notas recogida en la normativa. Dado que la nota en el expediente académico puede variar entre 0 y 10 con un decimal, podrían establecerse un máximo de 101 grupos, todos de igual amplitud (una décima de punto). Esta agrupación solo tendría sentido en el caso de que el número de alumnos fuera realmente muy numeroso.

Unas tercera y cuarta posibilidades, de entre las múltiples existentes, serían unas posiciones intermedias entre las dos anteriores. Podrían establecerse por ejemplo 11 grupos si cada uno de ellos se corresponde con un número entero (entre 0 y 10) o 21 si los grupos fueran cada medio punto (0, 0,5, 1, ... 9,5 y 10).

Una vez decididos los grupos la segunda fase consiste en preestablecer los porcentajes de alumnos que deben incluirse en cada uno de ellos. La proporción dependerá del número de grupos resultantes de la fase anterior, de la amplitud de cada uno de ellos y de la experiencia histórica de las evaluaciones llevadas a cabo en cursos anteriores, todo ello teniendo en cuenta que las clasificaciones adoptan una distribución normal. El sistema más objetivo, probablemente, sería analizar las notas obtenidas en cursos anteriores, establecer qué porcentaje medio de alumnos se encuentra en cada uno de los grupos prefijados y aplicar dichos porcentajes a los grupos correspondientes.

Las diferencias en el porcentaje de asignación entre asignaturas no sería un problema. Aunque en todos los casos las distribuciones de las calificaciones sean normales, podrían darse diferencias en función del empleo de distintas medias y desviaciones típicas a partir de la experiencia real de las diversas asignaturas y de la experiencia profesional de los profesores que aplicaran el método. Esto se entiende más fácilmente observando la Figura 2. En ella se encuentran representadas gráficamente diversas distribuciones de Gauss con distintas medias y desviaciones típicas. Todas ellas cumplen las propiedades de las distribuciones normales analizadas con anterioridad. Al variar la media la campana se desplaza hacia la derecha o hacia la izquierda, según esta sea mayor o menor. Igualmente, el empleo de una mayor o menor desviación típica hace que la amplitud de la campana, o dispersión respecto a la media, aumente o disminuya.



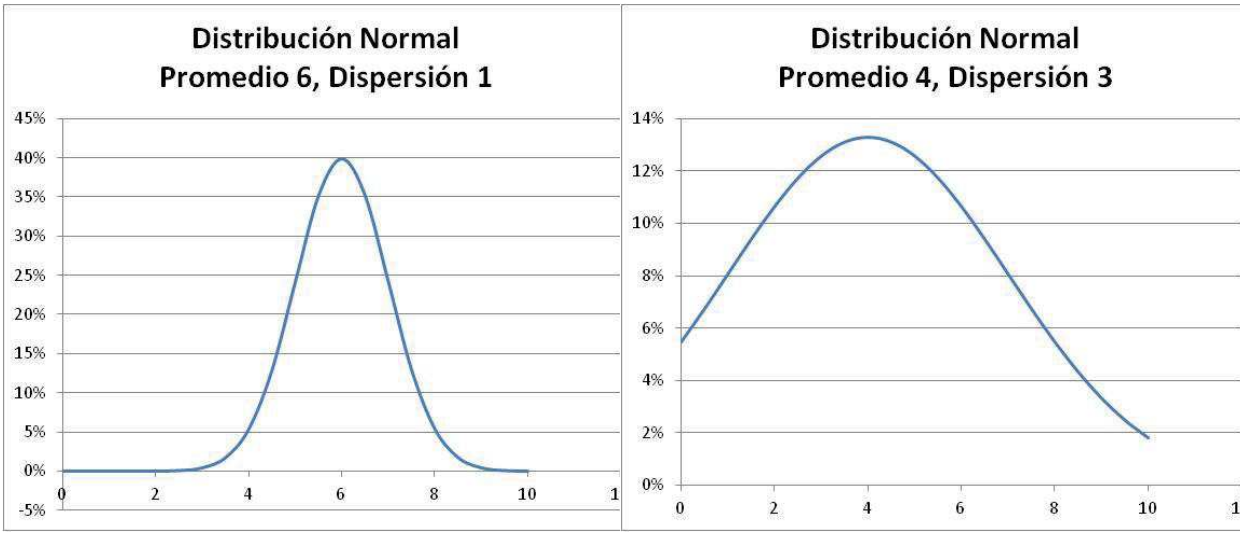



*Figura 2: Distribuciones normales con diferentes μ y σ. Fuente: Elaboración propia*

A modo de ejemplo, y partiendo de los grupos establecidos en la Tabla 1, podrían establecerse los porcentajes recogidos en la Tabla 2. Estos porcentajes estarían directamente relacionados con los valores de la media y desviación típica de la distribución de frecuencias que se estime oportuna para la asignatura en concreto.

| Calificación | Porcentaje de alumnos |
|---|---|
| Suspenso | 33% |
| Aprobado | 35% |
| Notable | 23% |
| Sobresaliente | 7% |
| Matrícula de Honor | 2% |
| Total | 100% |

*Tabla 2: Grupos y porcentajes de distribución forzosa. Fuente: Elaboración propia*

Una visión gráfica de los mismos sería la recogida en la Figura 3, donde dichos porcentajes se muestran mediante un gráfico circular de suspensos-aprobados con un subgráfico donde se incluyen los diversos subgrupos de aprobados (desde el aprobado estrictamente hablando hasta la matrícula de honor).

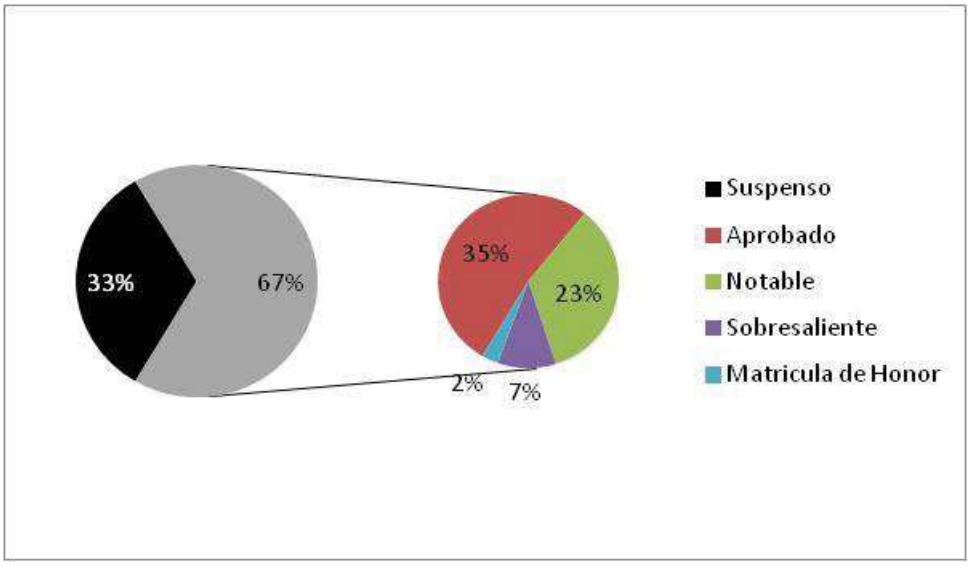



*Figura 3: Ejemplo de grupos y sus proporciones. Fuente: Elaboración propia*

El tercer paso consiste en realizar la evaluación de los alumnos por el mismo sistema que se ha empleado en cursos anteriores y puntuar a los alumnos en función de los criterios aplicados hasta entonces. Con ello se obtendrían unas notas que no serían las definitivas, sino orientativas y que se emplearían para la clasificación de los alumnos en los grupos preestablecidos, siendo esta la cuarta y última fase: la distribución de los alumnos en los diversos grupos establecidos al fijar los estándares teniendo en cuenta los porcentajes prefijados y sus resultados académicos.

Así, para un total de 300 alumnos, la distribución del ejemplo implicaría que los 99 alumnos con peores notas suspenderían, y 201 superarían la asignatura, independientemente de que el número suspensos en función de las notas fuera superior o inferior a 99. De igual modo, los 6 alumnos con mayores notas obtendrían matrícula de honor y los 21 siguientes sobresaliente, independientemente de la nota numérica resultante de la tercera fase.

| **Calificación** | **Porcentaje** | **Nº alumnos** |
|---|---|---|
| Suspenso | 33% | 99 |
| Aprobado | 35% | 105 |
| Notable | 23% | 69 |
| Sobresaliente | 7% | 21 |
| Matrícula de Honor | 2% | 6 |
| Total | 100% | 300 |

*Tabla 3: Calificaciones resultantes. Fuente: Elaboración propia*

Así, podría darse el caso de que, si la nota media de la evaluación ha sido muy elevada al haberse realizado una prueba "fácil", alumnos que obtuvieran más de un 5 en la misma suspendieran la evaluación o que, por el contrario, alumnos con una nota muy inferior a 5 la aprobaran en el caso de que la nota media hubiera sido muy baja debido a que la prueba hubiera resultado más difícil de lo común.

Como acción complementaria, dada la amplitud de los grupos establecidos, se podría usar la distribución jerárquica entre los alumnos de cada uno de los grupos obtenidos para mostrar las diferencias existentes entre ellos, mediante la asignación de una puntuación diferente entre los alumnos incluidos en cada grupo. Así, por ejemplo, en el caso del grupo de aprobados, existen un total de 20 posibles notas diferentes si la diferencia entre ellas es de un decimal (entre el 5,0 y el 6,9). Esto permitiría puntuar a los alumnos según su posición jerárquica dentro del grupo con diversas notas numéricas. En el caso que se está usando de ejemplo, al haber 105 aprobados, cada una de las notas numéricas debería serle asignada a 5 alumnos aproximadamente. Así, los cinco alumnos con peor evaluación dentro del grupo obtendrían un 5, los cinco siguientes un 5,1 y así sucesivamente. Todo ello independientemente de la nota obtenida en la prueba correspondiente, pues lo importante es el orden entre ellos y el valor absoluto de no dicha nota.

## 5. VENTAJAS E INCONVENIENTES DE LOS SISTEMAS DE EVALUACIÓN RELATIVOS

Este método tiene, como todos en general, sus ventajas e inconvenientes, que han sido profusamente tratados en la literatura científica por autores como Giumetti et al (2015), Obsborne & McCann (2004), Stewart et al (2010), Berger et al (2013), Blume et al (2013), Brown (2011), etc. En los dos subepígrafes siguientes se exponen algunas de las ventajas e inconvenientes más citados.

### *5.1 Ventajas.*

Existen evidentes ventajas, comunes a los diversos métodos de juicio relativo, frente a los métodos de juicio absoluto. Entre ellas destaca su utilidad para reducir los problemas derivados de una baja de fiabilidad tanto entre formas paralelas como entre evaluadores.

Es frecuente que cuando se realizan dos pruebas ambas no tengan la misma dificultad para el alumnado, de forma que los resultados varían en gran medida en función de a cuál de ellas se ha sometido el alumno, independientemente de que su nivel de conocimientos no varíe. Al aplicar el método de la Campana de Gauss éste inconveniente virtualmente desaparece. No importaría que un alumno se hubiera presentado a una u otra de las pruebas, toda vez que, al aplicar el método independientemente a la calificación obtenida en cada una de ellas, el porcentaje de alumnos con cada una de las calificaciones se mantendría uniforme entre ambas, independientemente de los resultados absolutos obtenidos.

Igual ocurre en el caso de la fiabilidad entre evaluadores al aplicar los grupos y porcentajes preestablecidos independientemente a los alumnos evaluados por cada uno de los profesores.

Una segunda ventaja a destacar es la mejora de la productividad de los sistemas de distribución forzada sobre aquellos que no lo son. Berger et al (2013) han encontrado que esta es superior entre un 6% y un 12%, salvo en caso de que se den conflictos competitivos entre los sujetos, siendo este uno de los inconvenientes del método, que se trata en el epígrafe siguiente.

Otra utilidad de estos métodos es su capacidad para atraer a las organizaciones que utilizan este tipo de sistemas de evaluación a sujetos con altas habilidades cognitivas, como indican Blume et al (2013).

Otras ventajas, menores en relación al objeto de este trabajo, atribuidas a los juicios relativos frente a los absolutos son el obligar a los evaluadores a establecer diferencias entre los distintos evaluados o el de facilitar la evaluación, dado que en general es más fácil para los seres humanos la evaluación comparativa que la absoluta.

### *5.2 Inconvenientes.*

Este método también tiene sus inconvenientes en comparación con los métodos basados en juicios absolutos. Entre ellos destacan los siguientes:

- No permite determinar la amplitud de las diferencias entre evaluados
- Si los grupos sometidos a evaluación no son homogéneos, las posiciones relativas pueden no ser comparables
- Pueden existir problemas si se evalúan diversas facetas a la hora de agruparlas en una calificación única
- Estimula conflictos competitivos entre los evaluados
- Percepción por los evaluados de ser menos justos que los basados en medidas absolutas
- Reduce el rendimiento de los sujetos con menor capacidad

El primer inconveniente citado es que este método no permite determinar la amplitud de las diferencias relativas entre evaluados. El motivo se encuentra en la base misma de la Campana de Gauss. Encuadra a los evaluados en grupos, según sus calificaciones relativas, sin tener en cuenta las diferencias absolutas entre ellos. Así, puede ocurrir que la diferencia entre el peor de los alumnos que ha obtenido notable y el mejor de los que han obtenido suficiente sea menor que la existente entre el primero y el último de los que han obtenido notable. Siguiendo con el ejemplo empleado con anterioridad, sería el caso de los alumnos clasificados jerárquicamente en los puestos 28º (mejor notable), 96º (peor notable) y 97º (mejor aprobado). Los alumnos 28º y 96º estarían en el mismo grupo (notable), cuando probablemente la distancia entre sus resultados sea mayor que la que haya entre el 96º y el 97º, que ha obtenido una clasificación de aprobado. Por otra parte, algo similar pasa en el caso de la medición absoluta, aunque en ese caso si se puede cuantificar la diferencia entre unos y otros.

El segundo inconveniente hace referencia a que si los grupos sometidos a evaluación no son homogéneos, las posiciones relativas pueden no ser comparables. Podemos partir de la base de que los grupos en convocatoria similares tienden a ser homogéneos, pero no tiene por qué ocurrir así en convocatorias diferentes. Se puede presuponer que los alumnos de la primera convocatoria de la asignatura tienden a ser mejores de media, en tanto que los de convocatorias posteriores (septiembre, diciembre) tienden a obtener calificaciones inferiores. El motivo es que los alumnos con mayor capacidad tienden a aprobar en primera convocatoria. Este inconveniente podría reducirse si se atribuyeran diferentes porcentajes a cada grupo preestablecido en función de la convocatoria, algo perfectamente factible.

Respecto al tercer inconveniente, la evaluación de diversas facetas a la hora de agruparlas en una calificación única, este problema es normal en todo proceso formativo. Es común que se realicen múltiples pruebas a los alumnos, midan las mismas o distintas facetas de la formación del alumnado, debiendo ponderarse los resultados obtenidos. Esto por ejemplo es común en procesos de evaluación continua, donde la calificación final suele ser resultado de la superación de diversas pruebas.

También se ha indicado que este sistema puede estimular conflictos competitivos entre los evaluados (Berger et al, 2013). Este inconveniente podría llegar a ser grave, sobre todo cuando últimamente las empresas y la sociedad en general demandan individuos con una elevada capacidad para el trabajo en equipo y la cooperación, siendo un inconveniente menor e incluso una ventaja el caso de empresas y sociedades más individualistas donde prime la competencia.

Otra característica de estos métodos es que, según autores como Roch et al (2007) y Schleicher et al (2009), los evaluados tienden a percibir estos métodos como menos justos que los basados en medidas absolutas, con los efectos que esto puede tener sobre la moral de los sujetos.

Por último, este método puede reducir el rendimiento de los sujetos con peores capacidades (Brown, 2011). Esta autora probado que en el caso de los torneos de golf, la participación o no de Tiger Woods (el mejor golfista del momento) afectaba a la puntuación del resto de los jugadores hasta en 0,8 golpes de media. Pese a ello, otros autores creen que en conjunto el rendimiento de los evaluados tiende a aumentar (Scullen et al, 2005).

## 6. DISCUSIÓN Y CONCLUSIONES

Con el presente trabajo no se pretende tanto la aplicación del método de la Campana de Gauss en la universidad como promover la reflexión sobre el hecho de que en muchos casos se produce una elevada inconsistencia a la hora de evaluar a los alumnos, debido a problemas de fiabilidad. Se propone por ello la necesidad de explorar de métodos que permitan reducir dicha inconsistencia. ¿A qué profesor no se le ha ido nunca la mano a la hora de poner una prueba, tanto en el sentido de que fuera demasiado difícil como en el de que fuera demasiado fácil? La baja fiabilidad hace que una parte más o menos sustancial de la nota obtenida no tenga su origen en la formación conseguida por el estudiante sino en la forma en que ha sido calificado.

El sistema propuesto, sin ser una panacea que todo lo cure como el Bálsamo de Fierabrás citado en el Quijote, es una propuesta sobre la que meditar para aliviar los problemas de fiabilidad que suelen aparecer a lo largo de la actividad evaluadora de un profesor universitario.

## REFERENCIAS

Alcaide M., González, M. y Florez, I. (2011). *Dirección de Recursos Humanos I: La función de personal en la empresa* 2011. Sevilla: Editorial Atril 97, S.L

Alcaide, M. y González, M. (1997) *Temas actuales de Dirección de Recursos Humanos*. Sevilla: Editorial Atril 97, S.L.

Berger, J., Harbring, C., & Sliwka, D. (2013). Performance appraisals and the impact of forced distribution—An experimental investigation. *Management Science*, *59*(1), 54-68.

Biggs, J. B. (1987). *Student approaches to learning and studying*. Hawt-horn, Vic.: Australian Council for Educational Research.

Blume, B. D., Rubin, R. S., & Baldwin, T. T. (2013). Who is attracted to an organisation using a forced distribution performance management system?. *Human Resource Management Journal*, *23*(4), 360-378.

Brown, J. (2011). Quitters never win: The (adverse) incentive effects of competing with superstars. *Journal of Political Economy*, *119*(5), 982-1013.

CEB Inc. (2015). Recuperado de *https://en.wikipedia.org/wiki/Vitality_curve*

Díaz, C., Batanero, C. & Cobo, B. (2003). Fiabilidad y generalizabilidad. *Aplicaciones en evaluación educativa*. Número 54, 3-21.

El confidencial (2013). Recuperado de http://www.elconfidencial.com/alma-corazon-vida/2013-08-27/la-curva-de-la-vitalidad-el-metodo-de-gestion-que-llevo-a-microsoft-al-fracaso_21219/.

Fortune Magazine (1999). recuperado de http://archive.fortune.com/magazines/fortune/fortune_archive/1999/11/22/269126/index.htm.

Giumetti, G. W., Schroeder, A. N., & Switzer III, F. S. (2015). Forced distribution rating systems: When does "rank and yank" lead to adverse impact?. *Journal of Applied Psychology*, *100*(1), 180.

Grote, D. (2013). Recuperado de https://online.wsj.com/news/articles/SB10001424052970203363504577186970064375222.

Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the econometric society*, 263-291.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, *87*(1), 72.

López, B. G., Pérez, C. P., Carbonell, B. S., Peris, F. S. I. & Ros, I. R. (2007). Actitudes ante el aprendizaje y rendimiento académico en los estudiantes universitarios. *Revista Iberoamericana de Educación*, *42*(1), 6.

Meliá, J. L. (2000). *Teoría de la fiabilidad y la validez*. Valencia: Cristóbal Serrano.

Normativa Reguladora de la Evaluación y Calificación de las Asignaturas de la Universidad de Sevilla. Recuperado de http://servicio.us.es/inspeccion/pdf/NORMATIVA%20REGUL.pdf.

Osborne, T., & McCann, L. A. (2004). Forced ranking and age-related employment discrimination. *Hum. Rts.*, *31*, 6.

Quirós, F. J. (2015). Análisis de las Tendencias en Gestión de los Recursos Humanos desde una Perspectiva Académica y Empresarial (Tesis doctoral no publicada). Universidad de Sevilla.

Real Academia de la Lengua Española (1992). *Diccionario de la Lengua Española, vol. I*. Madrid: Real Academia Española.

Real Decreto 1125/2003, de 5 de septiembre, por el que se establece el sistema europeo de créditos y el sistema de calificaciones en las titulaciones universitarias de carácter oficial y validez en todo el territorio nacional.

Roch, S. G., Sternburgh, A. M., & Caputo, P. M. (2007). Absolute vs relative performance rating formats: Implications for fairness and organizational justice. *International Journal of Selection and Assessment*, *15*(3), 302-316.

Schleicher, D. J., Bull, R. A., & Green, S. G. (2009). Rater reactions to forced distribution rating systems. *Journal of Management*, *35*(4), 899-927.

Schunk, D. H. (1997). *Teorías del aprendizaje*. Pearson educación.

Scullen, S. E., Bergey, P. K., & Aiman-Smith, L. (2005). Forced distribution rating systems and the improvement of workforce potential: A baseline simulation. *Personnel Psychology*, *58*(1), 1-32.

Stewart, S. M., Gruys, M. L., & Storm, M. (2010). Forced distribution performance evaluation systems: Advantages, disadvantages and keys to implementation. *Journal of Management & Organization*, *16*(1), 168-179.

The Corporate Executive Board Company (2013). Recuperado de http://www.washingtonpost.com/blogs/on-leadership/wp/2013/11/20/for-whom-the-bell-curve-tolls/.

Voda, V. G. (2009). Gauss' Bell Rings Forever!. *Economic Computation and Economic Cybernetics Studies and Research*, *43*, 237.

Wall Street Journal (2013). Recuperado de https://www.wsj.com/articles/8216rankandyank8217-that8217s-not-how-it8217s-done-1384473281.

Wherry, R. J., & Bartlett, C. J. (1982). The control of bias in ratings: A theory of rating. *Personnel Psychology*, *35*(3), 521-551.

Wherry, R. J., & Bartlett, C. J. (1982). The control of bias in ratings: A theory of rating. *Personnel Psychology*, *35*(3), 521-551.

WorldatWork (2013). Recuperado de http://www.computerworld.com/article/2486003/it-management/-stack-ranking--employee-eval-practice-falls-out-of-favor.html.