Work Rules!

Wie Google die Art und Weise, wie wir arbeiten und leben, verändert

Bearbeitet von
Laszlo Bock, Meike Grow, Ute Mareik

1. Auflage 2016. Buch. VIII, 370 S. Gebunden
ISBN 978 3 8006 5093 4
Format (B x L): 14,1 x 22,4 cm
Gewicht: 672 g

Wir richteten Steuerungs- und beratende Gruppen ein; über einige Fragen ließen wir sogar öffentlich abstimmen.

Am Ende waren drei Dinge klar:

1. Einen Konsens zu erreichen war unmöglich. Es gab keine Daten, die klar für die eine oder die andere Lösung sprachen, daher wurde jeder zum Experten, und es gab Anhänger aller nur denkbaren Varianten. Unsere Mitarbeiter hatten eine ausgeprägte Meinung zu der Frage, ob fünf oder sechs Leistungskategorien besser seien. Eine Lösung zu finden, die alle zufriedenstellte, erwies sich als unmöglich – selbst bei Veränderungen an dem Verfahren, das bei Google am wenigsten beliebt war. Viele Leute mochten das aktuelle System nicht, doch jede andere Lösung gefiel ihnen noch weniger!
2. Die Leute nahmen das Thema Leistungsmanagement ernst. Beispielsweise befragten wir die Googler, wie wir unsere Leistungskategorien nennen sollten, und es wurden mehr als 4.200 Stimmen abgegeben. Der deutlichste Trend war ein Bedürfnis nach Ernsthaftigkeit und Klarheit, nicht nach launiger Verspieltheit.
3. Experimente waren unerlässlich. Es gab keine Erfahrungen von anderen, auf die wir zurückgreifen konnten, daher mussten wir eigene Vorstellungen entwickeln und mit den Leitern aller Abteilungen von Google zusammenarbeiten, um Ideen auszuprobieren. Bei YouTube versuchten sie, alle Mitarbeiter in eine Hierarchie zu bringen, vom schwächsten bis zum leistungsstärksten, ohne Rücksicht auf das Level, auf dem sie arbeiteten. Dabei merkten sie, dass einer der beiden effektivsten Mitarbeiter ein Angestellter im Mittelbau war, der daraufhin eine extragroße Aktienzuteilung erhielt. Zwar wurde die Belohnung für diesen speziellen Mitarbeiter nicht öffentlich gemacht, aber jeder wusste, dass es solche Umgruppierungen im Gehalt gab.[xliii] An anderer Stelle versuchten wir, mit nur fünf Leistungsniveaus auszukommen; in einigen Fällen bewerteten die Manager das um 20 % besser als den früheren Ansatz mit 41 Kategorien.

Ich kann nicht genug betonen, wie schwierig das alles für die Personalabteilung war. Bei unserer Arbeit geht es nicht um Leben oder Tod, aber die Leute schrien auf, sie weinten, sie waren kurz davor zu kündigen. Veränderungen wie die oben genannten sind bei Google eine Herkulesaufgabe, *weil* wir den Googlern so viel Freiraum geben, *weil* wir so datengesteuert sind und *weil* die Googler sich Gedanken über Fairness und den Umgang miteinander machen. Alle Gruppen,

[xliii] Das war eine deutliche Erinnerung an die Maxime von Alan Eustace, dass ein großartiger Ingenieur 300 Durchschnittsingenieure wert ist und dass die traditionellen Leistungs- und Zahlungssysteme dazu führen, dass die Leute mehr nach ihrem Platz in der Hierarchie bezahlt werden als nach dem Beitrag, den sie tatsächlich leisten.

mit denen wir sprachen, waren vom aktuellen System frustriert und alle Gruppen sperrten sich dagegen, etwas Neues zu machen. Allein innerhalb von YouTube gab es ein Dutzend verschiedene Ideen, welches neue Bewertungssystem ausprobiert werden sollte. Ich bin ohne Ende stolz auf das Durchhaltevermögen, das Verständnis und die Sorgfalt, mit der das Team der Personalabteilung sich durch diese Veränderungen hindurchgearbeitet hat. Und besonders dankbar bin ich den Teams, die mit uns zusammengearbeitet haben, um 15 Jahre Google-Tradition über den Haufen zu werfen und tatsächlich etwas Neues zu machen.

Ausgehend von unseren Experimenten gaben wir Anfang 2013 die quartalsweisen Bewertungen auf – zugunsten von Halbjahresbewertungen. Es gab ein wenig Gezeter, aber nichts Wildes. Damit hatten wir schon einmal die Hälfte der Zeit eingespart.

Ende 2013 führten wir für mehr als 6.200 Googler, also fast 15% der Firma, eine 5-Punkte-Skala ein: verbesserungswürdig, entspricht regelmäßig den Erwartungen, übertrifft die Erwartungen, übertrifft die Erwartungen deutlich und erstklassig. Das ähnelte den Bezeichnungen, die wir vorher hatten, aber mit weniger Abstufungen.

Wir hielten uns an eine Grundüberzeugung der Medizin: *Primum non nocere*. Vor allen Dingen keinen Schaden anrichten. Weil das der erste Probelauf für diese Veränderung war, bestand unser Ziel einfach darin, die gleiche Zufriedenheit, Fairness und Effizienz zu erreichen wie bei der alten Bewertungsskala. Wir dachten, sobald die anfängliche Skepsis überwunden wäre („Was erzählen Sie mir da? Ich habe keine 3,8 mehr? Ich habe hart dafür gearbeitet, die 3,8 zu erreichen!") und der Lernprozess eingesetzt hätte, würden wir Zeit sparen und müssten bei den Bewertungen nicht mehr über Zehntelpunkte nachdenken. Und die Manager wären gezwungen, tiefer schürfende Gespräche mit ihren Angestellten zu führen; sie könnten sich nicht hinter der Aussage verstecken: „Sie sind in diesem Quartal um 0,1 Punkte besser geworden. Gut gemacht, weiter so!"

Wir waren erleichtert festzustellen, dass der Verlust an „Präzision" uns nicht wehtat. Trotzdem wollten wir vergleichen, wie Googler, die nach der 5-Punkte-Skala bewertet wurden, sich im Verhältnis zu den Googlern fühlten, die immer noch 41 Punkte bekamen. Wir fragten:

- Haben wir die schwachen Leister richtig identifiziert?
- Haben wir die richtigen Leute zur Beförderung ausgewählt?
- Waren die Diskussionen sinnvoll?
- War das Verfahren fair?

Die allgemeine Einschätzung war: Das neue Verfahren ist nicht schlechter als das alte. Mag sein, dass sich das anhört wie ein Pyrrhussieg, aber ich war wirklich erleichtert. Manche Googler hatten sich Sorgen

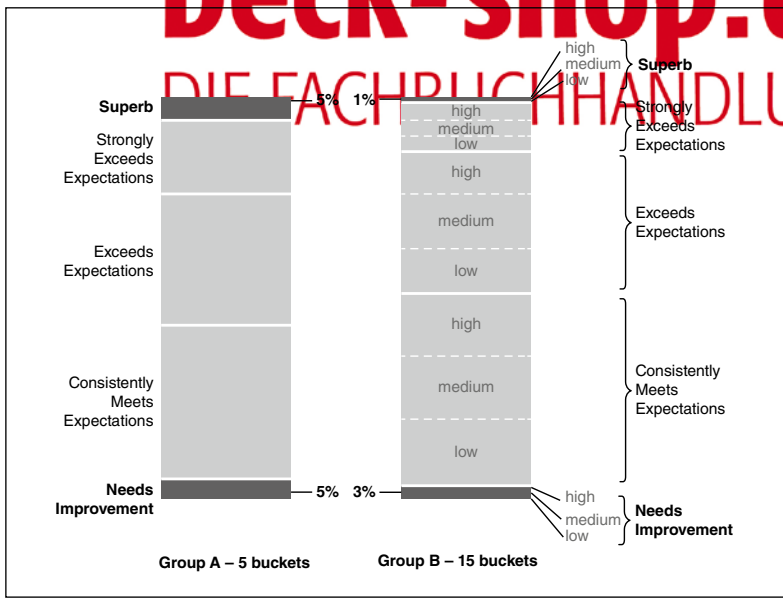
gemacht, dass der Verlust der präzisen 41-Punkte-Skala unsere Bewertungen weniger sinnig und aussagekräftig machen würde. Stattdessen enthüllten die Antworten der Googler auf die Umfrage, was wir schon lange vermutet hatten: Die 41 Punkte schufen lediglich eine Illusion der Genauigkeit.

Die meisten Googler gaben zu, dass es bei vielen Bewertungen nicht möglich war, den Unterschied zwischen 0,1 mehr oder weniger festzumachen. Beispielsweise gab es keine einhellige Meinung, was den Unterschied zwischen 3,1 und 3,2 ausmachte. Megan Huth, Mitarbeiterin in unserem „People and Innovation Lab" erklärt: „So war es möglich, dass die Bewertungen weder verlässlich noch stichhaltig waren. Die gleiche Person wurde bei gleicher Leistung mit 3,2 oder mit 3,3 eingestuft – je nach Bewerter und Kalibrierungsgruppe. Das bedeutet, die Einstufung war nicht verlässlich. Und wenn sie 3,3 bekam, obwohl sie tatsächlich 3,2 war, dann war die Einstufung auch nicht stichhaltig – sie gab einfach nicht die Realität wieder."

Die Bewertungen hätten, so Megan, die Angabe eines Fehlers beinhalten müssen. Wir hätten den Leuten sagen sollen: „Jim, deine Leistung liegt irgendwo zwischen 3,3 und 3,5." Aber in der Praxis fand das nicht statt. Die Manager nahmen die Zahl und wiesen ihr eine echte Bedeutung zu. Wenn jemand sich von 3,3 auf 3,5 verbesserte, musste das den Grund haben, dass er besser geworden war – auch wenn er tatsächlich einfach hätte weitermachen können wie bisher. Man stelle sich vor, wie viel schlimmer das noch war, wenn die Bewertung sank. Es wurde einem gesagt, das habe mit der eigenen Leistung zu tun, obwohl es sich in der Realität um einen Messfehler handelte.

Und dann passierte etwas Spannendes. Die 6.200 Googler waren auf acht verschiedene Gruppen verteilt. Davon beschlossen drei Gruppen, die zusammen mehr als 1.000 Leute umfassten, die fünf Kategorien weiter zu unterteilen. Ein Team fügte jeder Kategorie drei Unterkategorien hinzu; das heißt ein Star-Googler konnte als „hoch erstklassig", „mittel erstklassig" oder „niedrig erstklassig" eingestuft werden. Das folgende Diagramm zeigt, wie die Bewertungen am Ende verteilt waren. Die Unterkategorien habe ich in die fünf Hauptkategorien zurückgeführt; so kann man den Unterschied zwischen den beiden Ansätzen leicht erkennen. Gruppe A blieb bei den fünf Kategorien und Gruppe B hatte 15.

Gruppe B hatte mehr Leistungsstufen; die Hoffnung war, das würde zu einer besseren Differenzierung unter den Mitarbeitern führen. Tatsächlich war das Ergebnis jedoch weniger differenziert als in Gruppe A. 5 % aus Gruppe A waren „erstklassig", aber nur ein Prozent aus Gruppe B. Ich kann Ihnen verraten, dass die Teams insgesamt das gleiche Leistungsniveau hatten. Ihre Wertschöpfung war vergleichbar, und die Mitarbeiter in den Gruppen waren gleich gut. Einfach weil sie



Durchschnittliche Verteilung der Bewertungen von Gruppe A und Gruppe B.

mehr Kategorien zur Auswahl hatte, gelangte Gruppe B unbewusst, ungewollt und fälschlicherweise zu dem Schluss, dass sie praktisch keine Stars hatte. Obwohl das nicht beabsichtigt war, fielen in Gruppe B 80 % der Spitzenleute aus der Spitzenkategorie heraus.

Wenn Sie diesen Text lesen, haben wir bei Google überall die 5-Punkte-Skala eingeführt. Ende 2013 war sie noch ein Experiment, doch die Zeichen standen gut. Zunächst einmal erhielten die Angestellten durch das neue System konsequenteres Feedback, weil die undurchsichtigen Unterschiede zwischen Einstufungen wie 3,2 oder 3,3 verschwanden. Zweitens führte das neue System zu einer breiteren Leistungsverteilung. Bei weniger Leistungskategorien erhöhte sich die Wahrscheinlichkeit, dass die Manager auch auf die Extreme der Skala zurückgreifen. Die akademische Forschung zum Thema war nicht eindeutig, und das Feedback der Googler war neutral – trotzdem gelangten wir zu dem Schluss, dass fünf Kategorien besser sind als viele Kategorien, und zwar mindestens in zweifacher Hinsicht.

Mitte 2014 entdeckten wir weitere positive Resultate. Wir sind der Ansicht, dass verschiedene Aufgabenfelder unterschiedliche Möglichkeiten eröffnen, Einfluss auszuüben. Wenn Sie Ingenieur sind, können von Ihrem neuen Produkt 100 Leute profitieren oder eine Milliarde. Wenn sie ein Anwerber sind, können Sie sich abrackern, wie Sie wollen, Sie haben einfach nicht die Zeit, Einfluss auf eine Milliarde Menschen auszuüben. Als wir damit aufhörten, Empfehlungen zu geben, wie die

richtige Verteilung der Bewertungen aussehen sollte, stellten wir fest, dass sich vier charakteristische Bewertungsmuster herauskristallisierten, die tatsächlich die Leistungsmerkmale verschiedener Gruppen und Individuen besser reflektierten.

Wir stellten außerdem fest, dass die Manager doppelt so viele Mitarbeiter in die Extreme der Bewertungsskala einstuften. Damit wurden sie der effektiven Leistung ihrer Leute besser gerecht (lesen Sie in Kapitel 10, warum das so ist). Und da sich das Stigma verringerte, zur schlechtesten Kategorie zu gehören, fiel es manchen Managern leichter, direkte, einfühlsame Gespräche über mögliche Verbesserungen mit denjenigen Mitarbeitern zu führen, die sich abquälten.

Nach endlosen Diskussionen war es uns gelungen, ein Bewertungssystem, das ungenau und unwirtschaftlich war, durch ein ganz neues zu ersetzen – das einfacher und präziser war und den gleichen Zeitaufwand benötigte, um die Bewertungen zu kalibrieren. Ganz klar: Diskussionen gibt es immer noch! Aber wir arbeiten uns voran. Und wir können allmählich erkennen, dass die Leute sich immer mehr mit dem neuen System anfreunden und es schätzen lernen.

Ich teile Ihnen das alles wie für eine Betaversion mit: Genau wie wir Produkte ins Netz stellen, wenn es ganz so aussieht, als seien sie deutlich nützlicher als das, was es da draußen bereits gibt – aber bevor sie zu 100 % ausgefeilt und perfekt sind.

Dies vorausgeschickt, ist die Frage, wie viele Kategorien eine Bewertungsskala hat, nicht besonders wichtig, auch wenn die Googler sie erstaunlich leidenschaftlich diskutierten. Bieten Sie nicht 15 und mehr Stufen auf der Skala an, aber wenn Sie drei oder sechs verschiedene Bewertungsmöglichkeiten haben, lassen Sie das einfach so.

Fairness sicherstellen

Andererseits ist die Seele der Leistungsbewertung die Kalibrierung. Man kann wirklich sagen, ohne Kalibrierung wäre unser Bewertungsverfahren weniger fair, weniger zuverlässig und weniger effektiv. Ich glaube, dass die Kalibrierung der Grund dafür ist, warum die Einstellung der Googler zu ihrem Bewertungssystem doppelt so positiv war wie in anderen Firmen.

Worum geht es dabei also? Googles Bewertungssystem zeichnet sich (schon immer) dadurch aus, dass nicht der direkte Vorgesetzte entscheidet. Der Manager weist seinem Angestellten eine vorläufige Bewertung zu – sagen wir „übertrifft die Erwartungen“. Sie beruht auf erreichten Zielvorgaben (OKRs), wird aber auch durch andere Dinge beeinflusst, beispielsweise durch die Anzahl geführter Einstellungsgespräche oder

durch besondere Umstände, wie wirtschaftliche Veränderungen, die Auswirkungen auf das Anzeigenaufkommen hatten.[xliv] Ehe diese vorläufige Bewertung endgültig wird, setzen sich die Manager in Gruppen zusammen und gehen alle vorläufigen Bewertungen ihrer Angestellten gemeinsam durch – in einem Prozess, den wir Kalibrierung nennen.

Kalibrierung bedeutet einen Arbeitsschritt mehr. Aber sie ist entscheidend, um die Fairness sicherzustellen. Die Einschätzung eines bestimmten Managers wird mit den Einschätzungen von Managern verglichen, die ähnliche Gruppen leiten, und alle bewerten ihre Angestellten gemeinsam: Eine Gruppe von fünf bis zehn Managern trifft sich und projiziert 50 bis 1.000 Angestellte an die Wand, man spricht über den ein oder anderen und einigt sich auf eine faire Bewertung. So können wir den Druck abbauen, den manche Angestellte gern auf ihre Manager ausüben, Bewertungen aufzublähen. Und so ist sichergestellt, dass das Endergebnis eine gemeinsame Leistungseinschätzung widerspiegelt; einzelne Manager erwarten häufig doch ganz unterschiedliche Dinge von ihren Mitarbeitern und interpretieren Leistungsstandards auf ganz persönliche Art und Weise – genau wie in der Schule, wo manche Lehrer gute Noten gaben und andere einfach härter waren. Die Kalibrierung verringert die Voreingenommenheit, indem sie die Manager zwingt, ihre Entscheidungen zu rechtfertigen. Sie erhöht auch den Eindruck der Fairness unter den Angestellten.[114]

Was die Kalibrierung bei der Einschätzung von Mitarbeitern bewirkt, unterscheidet sich nicht so sehr von dem, was es bringt, wenn die Leute nach einem Einstellungsgespräch ihre Notizen vergleichen. Das Ziel ist dasselbe: individuelle Vorurteile möglichst auszuschließen. Selbst in einer kleinen Firma haben Sie bessere Ergebnisse und zufriedenere Angestellte, wenn die Einschätzungen auf einer Gruppendiskussion beruhen und nicht auf den Launen eines einzelnen Managers.

Doch auch mit Kalibrierung können Manager selbst in der Gruppe schlechte Entscheidungen fällen. Wenn wir andere bewerten, schleichen sich in den Entscheidungsprozess zahlreiche Fehler ein. Beispielsweise handelt es sich um den sogenannten Rezenzeffekt, wenn wir die jüngere Erfahrung mit einem Mitarbeiter zu stark gewichten, einfach weil wir uns noch gut daran erinnern. Wenn ich diese Woche eine großartige Begegnung mit einer Angestellten hatte und dann in die Kalibrierungssitzung gehe, wo über sie gesprochen wird, fällt meine Einschätzung vermutlich von vornherein zu gut aus, weil ich mich im Unterbewusstsein auf die jüngste, positive Interaktion stütze. Wir versuchen, dieses Problem aus der Welt zu schaffen, indem wir zu Beginn der meisten Kalibrierungssitzungen ein einzelnes Blatt verteilen, auf

[xliv] Das ist wichtig. Die OKRs *beeinflussen* die Bewertungen, sind aber nicht allein entscheidend.

dem die häufigsten Fehler beschrieben sind, die man bei der Einschätzung machen kann, und wie man sie vermeidet. Eine Version davon sehen Sie in der folgenden Abbildung.

Wir beginnen jede Kalibrierungssitzung, indem wir uns diese möglichen Fehler vor Augen führen. In den Kalibrierungssitzungen, denen ich beiwohne, ist mir aufgefallen, dass es ausreicht, den Managern diese Phänomene bewusst zu machen, und sei es nur für kurze Zeit. Viele Verzerrungen können so eliminiert werden. Mindestens ebenso wichtig ist es, dass auf diese Weise eine Sprache und eine kulturelle Norm entstehen, sich vor Kalibrierungsfehlern zu hüten. Es ist heute nicht unüblich, dass jemand in einer Kalibrierungssitzung dem Gespräch eine neue Richtung gibt, indem er sagt: „Moment mal. Das ist jetzt der Rezenzeffekt. Wir müssen uns die Leistung über den gesamten Zeitraum ansehen, nicht nur in der letzten Woche."

Tipps für eine evidenzbasierte Kalibrierung

Kognitive Verzerrung/ Gruppendynamik	Definition	Beispiel
Halo-Effekt	Der Gesamteindruck eines Menschen, der im Allgemeinen hervorragend/schrecklich ist, trübt das Urteil gegenüber neuen Erfahrungen, die in eine andere Richtung weisen	„Tom ist immer ...
Rezenzeffekt	Die Tendez, sich an die letzten Dinge zu erinnern, die jemand getan hat, und ihnen übermäßiges Gewicht einzuräumen	„Tom hat letztens ...
Attributionsfehler	Zu viel Aufmerksamkeit auf die „Fähigkeit" eines Menschen richten und nicht genug auf die Situation/den Kontext, die einen Einfluss auf die Leistung haben – oder umgekehrt	...
Tendenz zum Mittelmaß	„Auf Nummer sicher gehen", indem man mittlere Werte abgibt	...
Verzerrung durch Verfügbarkeit	Wenn man den Fehler macht, das, woran man sich gut erinnern kann, für das Häufigere zu halten	...

Auszug aus einem Handout, das vor der Kalibrierungssitzung bereitgestellt wird.
© Google, Inc.

Sie bemerken sicherlich, dass wir immer noch erhebliche Zeit in dieses Verfahren investieren – selbst nachdem wir die Häufigkeit unserer Bewertungen reduziert und die Skala vereinfacht haben, nach der wir unsere Mitarbeiter einstufen. Es braucht vielleicht zwischen zehn und 30 Minuten, Ihrer Gruppe vorläufige Bewertungen zuzuweisen, indem Sie Kästchen im Performance Management Tool abhaken. Aber eine Kalibrierungssitzung kann drei Stunden oder länger dauern. Dabei wird nicht unbedingt über jeden einzelnen Mitarbeiter gesprochen. Eine gewisse Zeit braucht es einfach, die Kalibratoren selbst zu kalibrieren,

indem wir unsere Einschätzung einzelner Mitarbeiter vergleichen, die verschiedenen Managern gut bekannt sind. Sie können dann als Maßstab oder Bezugsgröße dienen. Außerdem sehen sich die Kalibratoren die Verteilung der Bewertungen in den verschiedenen Gruppen an, nicht um zwangsläufig eine bestimmte Verteilung durchzusetzen, sondern um zu verstehen, warum in manchen Gruppen die Verteilung anders ist. Beispielsweise kann es sein, dass ein Team aus guten Gründen stärker ist als ein anderes. Der Großteil der Zeit dient dann dazu, Fälle zu diskutieren, die aus irgendwelchen Gründen aus dem Rahmen fallen, beispielsweise wenn die Leistung besonders schnell zu- oder abgenommen hat, wenn große Leistungsschwankungen vorliegen oder wenn es sich um Grenzfälle am Rand einer Kategorie handelt.

Viele Firmen geben ihre Bewertungssysteme ganz auf. Warum bleiben wir dabei?

Ich glaube, das hat mit Fairness zu tun.

Bewertungen sind Werkzeuge, Techniken zur Vereinfachung, die Managern helfen, Gehalts- und Beförderungsentscheidungen zu treffen. Als Angestellter möchte ich fair behandelt werden. Ich habe nichts dagegen, dass jemand mehr Gehalt bekommt als ich, wenn er dafür mehr leistet. Aber wenn wir die gleiche Arbeit tun und er bekommt deutlich mehr, macht mich das sehr unzufrieden. Ein gerechtes Bewertungssystem bedeutet, dass ich mir in dieser Hinsicht keine Gedanken machen muss. Es bedeutet auch, dass jemand, der außergewöhnliche Arbeit leistet, nicht nur von seinem Vorgesetzten gesehen wird, sondern außerdem von vielen anderen Managern – beim Kalibrierungstreffen, wo von allen gemeinsam ein einheitlicher Leistungsstandard für die gesamte Firma geschaffen und verkündet wird. Bewertungen machen es Mitarbeitern auch leichter, ihren Arbeitsplatz innerhalb der Firma zu wechseln. Als Manager kann ich mich darauf verlassen, dass jemand, der „die Erwartungen deutlich übertrifft", großartige Arbeit leistet, ob ihr letzter Job nun die Arbeit bei Chrome, bei Glass oder in der Vertriebsabteilung war. Als Angestellter kann ich das Vertrauen haben, dass die Leute wegen ihrer Verdienste befördert werden und nicht wegen taktischer Einflussnahmen. Bei einer kleinen Gruppe brauchen Sie diese Infrastruktur nicht – Sie kennen jeden Einzelnen. Aber wenn Sie viele Hundert Mitarbeiter haben, vertrauen die Angestellten lieber auf ein zuverlässiges System als auf einen einzelnen Manager. Nicht weil Manager grundsätzlich schlecht oder voreingenommen wären, sondern weil ein Bewertungsverfahren, zu dem eine Kalibrierung gehört, schlechte und voreingenommene Bewertungen aktiv ausmerzt.