Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Estudios de Postgrado
Maestría en Estadística Aplicada

CONSTRUCCIÓN DE CURVAS DE FRECUENCIA PARA LA EVALUACIÓN DEL DESEMPEÑO Y LA APLICACIÓN DE LA METODOLOGÍA DE 9 CAJAS EN LA MEJORA DE LA ADMINISTRACIÓN DE TALENTO EN UNA INSTITUCIÓN SEMIAUTÓNOMA

Inga. María del Carmen Muñoz Pineda
Asesorado por la Mtra. Mayra Virginia Carvajal Castillo

Guatemala, noviembre de 2022

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

CONSTRUCCIÓN DE CURVAS DE FRECUENCIA PARA LA EVALUACIÓN DEL DESEMPEÑO Y LA APLICACIÓN DE LA METODOLOGÍA DE 9 CAJAS EN LA MEJORA DE LA ADMINISTRACIÓN DE TALENTO EN UNA INSTITUCIÓN SEMIAUTÓNOMA

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

INGA. MARÍA DEL CARMEN MUÑOZ PINEDA
ASESORADO POR LA MTRA. MAYRA VIRGINIA CARVAJAL CASTILLO

AL CONFERIRSELE EL TÍTULO DE

MAESTRA EN ESTADÍSTICA APLICADA

GUATEMALA, NOVIEMBRE DE 2022

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

NÓMINA DE JUNTA DIRECTIVA

DECANA	Inga. Aurelia Anabela Cordova Estrada
VOCAL I	Ing. José Francisco Gómez Rivera
VOCAL II	Ing. Mario Renato Escobedo Martínez
VOCAL III	Ing. José Milton de León Bran
VOCAL IV	Br. Kevin Vladimir Cruz Lorente
VOCAL V	Br. Fernando José Paz González
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

DECANA	Inga. Aurelia Anabela Cordova Estrada
EXAMINADOR	Mtro. Edwin Adalberto Bracamonte Orozco
EXAMINADOR	Dra. Aura Marina Rodríguez de Peña.
EXAMINADOR	Mtro. William Eduardo Fagiani Cruz
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

HONORABLE TRIBUNAL EXAMINADOR

En cumplimiento con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

CONSTRUCCIÓN DE CURVAS DE FRECUENCIA PARA LA EVALUACIÓN DEL DESEMPEÑO Y LA APLICACIÓN DE LA METODOLOGÍA DE 9 CAJAS EN LA MEJORA DE LA ADMINISTRACIÓN DE TALENTO EN UNA INSTITUCIÓN SEMIAUTÓNOMA

Tema que me fuera asignado por la Dirección de la Escuela de Estudios de Posgrado de la Facultad de Ingeniería, con fecha 6 de agosto de 2021.

Inga. María Del Carmen Muñoz Pineda

Decanato
Facultad de Ingeniería
24189101- 24189102
secretariadecanato@ingenieria.usac.edu.gt

LNG.DECANATO.OI.740.2022

La Decana de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer la aprobación por parte del Director de la Escuela de Estudios de Posgrado, al Trabajo de Graduación titulado: **CONSTRUCCIÓN DE CURVAS DE FRECUENCIA PARA LA EVALUACIÓN DEL DESEMPEÑO Y LA APLICACIÓN DE LA METODOLOGÍA DE 9 CAJAS EN LA MEJORA DE LA ADMINISTRACIÓN DE TALENTO EN UNA INSTITUCIÓN SEMIAUTÓNOMA,** presentado por: **María del Carmen Muñoz Pineda,** que pertenece al programa de Maestría en artes en Estadística aplicada después de haber culminado las revisiones previas bajo la responsabilidad de las instancias correspondientes, autoriza la impresión del mismo.

IMPRÍMASE:

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
DECANA
FACULTAD DE INGENIERÍA

Inga. Aurelia Anabela Cordova Estrada
Decana

Guatemala, noviembre de 2022

AACE/gaoc

Guatemala, noviembre de 2022

LNG.EEP.OI.740.2022

En mi calidad de Director de la Escuela de Estudios de Postgrado de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen del asesor, verificar la aprobación del Coordinador de Maestría y la aprobación del Área de Lingüística al trabajo de graduación titulado:

"CONSTRUCCIÓN DE CURVAS DE FRECUENCIA PARA LA EVALUACIÓN DEL DESEMPEÑO Y LA APLICACIÓN DE LA METODOLOGÍA DE 9 CAJAS EN LA MEJORA DE LA ADMINISTRACIÓN DE TALENTO EN UNA INSTITUCIÓN SEMIAUTÓNOMA"

presentado por **María del Carmen Muñoz Pineda** correspondiente al programa de **Maestría en artes en Estadística aplicada** ; apruebo y autorizo el mismo.

Atentamente,

"Id y Enseñad a Todos"

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
FACULTAD DE INGENIERÍA
DIRECCIÓN
ESCUELA DE POSTGRADO

Mtro. Ing. Edgar Darío Álvarez Cotí
Director
Escuela de Estudios de Postgrado
Facultad de Ingeniería

Guatemala 17 de noviembre 2021.

M.A. Edgar Darío Álvarez Cotí
Director
Escuela de Estudios de Postgrado
Presente

M.A. Ingeniero Álvarez Cotí:

Por este medio informo que he revisado y aprobado el Informe Final del trabajo de graduación titulado **"CONSTRUCCIÓN DE CURVAS DE FRECUENCIA PARA LA EVALUACIÓN DEL DESEMPEÑO Y LA APLICACIÓN DE LA METODOLOGÍA DE 9 CAJAS EN LA MEJORA DE LA ADMINISTRACIÓN DE TALENTO EN UNA INSTITUCIÓN SEMIAUTÓNOMA"** de la estudiante **María del Carmen Muñoz Pineda** quien se identifica con número de carné **8511810** del programa de Maestría en Estadística Aplicada.

Con base en la evaluación realizada hago constar que he evaluado la calidad, validez, pertinencia y coherencia de los resultados obtenidos en el trabajo presentado y según lo establecido en el *Normativo de Tesis y Trabajos de Graduación aprobado por Junta Directiva de la Facultad de Ingeniería Punto Sexto inciso 6.10 del Acta 04-2014 de sesión celebrada el 04 de febrero de 2014*. Por lo cual el trabajo evaluado cuenta con mi aprobación.

Agradeciendo su atención y deseándole éxitos en sus actividades profesionales me suscribo.

Atentamente

MSc. Ing. Edwin Adalberto Bracamonte Orozco
Coordinador
Maestría en Estadística Aplicada
Escuela de Estudios de Postgrado

Guatemala, 30 de julio de 2021.

M.A. Ing. Edgar Darío Álvarez Cotí

Director

Escuela de Estudios de Postgrado

Presente

Estimado M.A. Ing. Álvarez Cotí

Por este medio informo a usted, que he revisado y aprobado el Trabajo de Graduación y el Artículo Científico: "**CONSTRUCCIÓN DE CURVAS DE FRECUENCIA PARA LA EVALUACIÓN DEL DESEMPEÑO Y LA APLICACIÓN DE LA METODOLOGÍA DE 9 CAJAS EN LA MEJORA DE LA ADMINISTRACIÓN DE TALENTO EN UNA INSTITUCIÓN SEMIAUTÓNOMA**" de la estudiante **María del Carmen Muñoz Pineda** del programa de Maestría en **Estadística aplicada,** identificada con número de carné: **8511810.**

Agradeciendo su atención y deseándole éxitos en sus actividades profesionales me suscribo.

Mayra Virginia Carvajal Castillo
Ingeniera Industrial
Colegiado No. 15,165

Mtra. Ing. Mayra Virginia Carvajal

Colegiado No. 15,165

Asesor de Tesis

ACTO QUE DEDICO A:

Dios	Por haberme permitido realizar una de mis metas.
Mi mamá	Antonia Pineda de Muñoz, por siempre creer en mí y apoyarme incondicionalmente.
Mi papá	Francisco Muñoz Oliveros, por darme la vida y todo lo que necesito para tomarla.

AGRADECIMIENTOS A:

Universidad de San Carlos de Guatemala	Por la oportunidad de formarme profesionalmente en ella.
Mtra. Mayra Carvajal	Por asesorar mi trabajo de tesis.
Mi asesor	Mtro. Ing. Andrés García, por todos los conocimientos técnicos que me aportó para la realización de esta investigación.
Facultad de Ingeniería	En especial a la Escuela de Estudios de Postgrado, su claustro y dirección, por la oportunidad de acceder a un nivel superior de formación.
Dra. Aura Marina Rodríguez de Peña	Por el impulso y el acompañamiento para concluir este trabajo.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES

FIGURAS

TABLAS

LISTA DE SÍMBOLOS

Símbolo	Significado
α	Alfa de Cronbach
r_{oe}	Coeficiente de correlación de Pearson
r_{xx}	Coeficiente de correlación
χ^2	Distribución chi cuadrado
F_0	Distribución F
μ	Media
Σ	Sumatoria
∪	Unión
valor-p	Valor-p = $P(F>F_0)$
s_d^2	Varianza de diferencias
$s_{1\alpha}^2$	Varianza de mitades
σ_x^2	Varianza de x

GLOSARIO

ANOVA	Análisis de varianza.
Baremo	Conjunto de normas que establecen el conjunto de criterios para medir o evaluar los méritos, daños o aportes que presenta una persona o institución.
H_A	Hipótesis alternativa.
H_0	Hipótesis nula.
Metodología de 9 cajas	Metodología de clasificación de talento en categorías.

RESUMEN

En el presente trabajo de graduación se analizó, desde el punto de vista estadístico, un ejercicio de evaluación de desempeño realizado en una institución guatemalteca semiautónoma en diciembre del año 2019. Uno de sus objetivos está el conferir sustento estadístico al ejercicio y que con esa base los integrantes de la organización mejoren la percepción de imparcialidad y confiabilidad de este. Adicionalmente incrementar su fiabilidad por medio de metodología estadística y proponer una forma de clasificación de talento a través de referencias normativas que confirmen su imparcialidad e incrementar su credibilidad.

El estudio tiene un alcance descriptivo y correlacional y de diseño no experimental. Se analizó la base de datos que consta de 17,919 registros de evaluados y 1,123 evaluadores. Se determinó que el instrumento utilizado para evaluar es suficientemente confiable. También se identificó la existencia de tendencias extremas de evaluación y la necesidad de categorizar mediante un baremo, las clases de desempeño para asignarlas con mayor certeza en el eje horizontal del sistema de nueve cajas. El hallazgo principal fue que el nivel organizacional del evaluado tiene incidencia sobre la calificación que obtiene en la evaluación y que es necesario el uso de un baremo para clasificarla de manera más acertada.

PLANTEAMIENTO DEL PROBLEMA

- Contexto general

En una institución semiautónoma, con quince mil trabajadores, se realizó la primera evaluación de desempeño en noviembre del año 2019. Este ejercicio no dejó una buena impresión en algunos de los evaluados, ya que percibieron una inequidad en los siguientes aspectos: (1) asignación de calificaciones por parte de los evaluadores, (2) aspectos evaluados y (3) niveles de desempeño a los que fueron asignados. Actualmente no existe un análisis que pueda afirmar o negar que esta inconformidad es fundamentada y que lleve a una propuesta de mejora.

- Descripción del problema

La evaluación de desempeño presenta problemas en su ejecución debido a que la mayor parte del proceso puede ser subjetivo. En algunos casos las calificaciones se pueden asignar a través de una apreciación personal y/o los elementos que se califican solo pueden ser apreciados y por ello pueden ser afectados por juicios del evaluador. Esta subjetividad puede generarse desde el diseño del instrumento que se utiliza para evaluar, durante el proceso mismo de la calificación, hasta su clasificación en el tablero de 9 cajas.

El área de recursos humanos necesita respaldar el proceso de evaluación del desempeño. Una forma de dar este respaldo es brindar mayor objetividad al mismo por medio de la aplicación de métodos estadísticos para cuantificar la confiabilidad del instrumento, explicar a través de pruebas de hipótesis si existen

diferencias significativas sobre los punteos asignados entre evaluadores y finalmente construir un baremo para asignar los punteos a las 9 cajas de clasificación de talento.

- Formulación del problema

El conocer el planteamiento del problema nos dirige a la pregunta central de esta investigación: ¿Cómo reducir el efecto de imparcialidad que tienen las distintas variables que afectan la evaluación de desempeño en la asignación de calificaciones a los evaluados y su clasificación en niveles de la metodología de nueve cajas? De acuerdo con el análisis de investigaciones previas, para responder esta pregunta se deben contestar primero las siguientes interrogantes:

  - ¿Cómo dimensionar la fiabilidad y validez del instrumento de evaluación?

  - ¿Cuál es el efecto que tiene el evaluador sobre las calificaciones que asigna a los evaluados, que pueda generar imparcialidad al compararlo con la tendencia general del ejercicio?

  - ¿Cuál es la mejor forma de categorización de calificaciones de evaluación del desempeño que sea imparcial y utilice principios de distribución estadística?

OBJETIVOS

General

Incrementar la fiabilidad del ejercicio de evaluación del desempeño con el uso de una metodología estadística, proponer una forma de clasificación del talento a través de referencias normativas para que el ejercicio de evaluación sea imparcial y así reforzar su credibilidad.

Específicos

- Evaluar la fiabilidad y validez del instrumento aplicado para la evaluación del desempeño y utilizar distintos análisis de correlación.

- Detectar si existe parcialidad en la asignación de calificaciones de evaluación de parte de los evaluadores y si existen tendencias que afecten a los evaluados a través de una prueba de hipótesis. Si se detecta imparcialidad, recomendar formas de mejora de este punto.

- Proponer una forma de categorización del desempeño de los colaboradores de la organización para que se tome una decisión adecuada respecto de su talento. Se trabajará en la creación de baremos y se tomarán como base las categorías de desempeño que utiliza la metodología de 9 cajas.

RESUMEN DEL MARCO METODOLÓGICO

- Características del estudio

El enfoque del estudio realizado es mixto. El objeto de estudio es un proceso de evaluación de desempeño y por tanto los actores son personas que juegan el papel de evaluadores y evaluados. Todos ellos tienen comportamientos derivados de sus propios criterios y juicios dentro del proceso y es susceptible de ser analizado tanto el instrumento de evaluación, como la base de datos de las calificaciones de las personas evaluadas y sus evaluadores.

El alcance es descriptivo, correlacional y explicativo. Propone mejoras sobre la fiabilidad y validez cuantitativa de los procesos de evaluación de desempeño. Utiliza pruebas paramétricas para describir la fiabilidad actual del instrumento de evaluación, analiza la independencia y correlación de las calificaciones asignadas y el evaluador que las asigna, así como la construcción de un baremo para la clasificación de las calificaciones de los evaluados. Propone una mejora a la metodología de 9 cajas que actualmente se llena con criterios personales.

El diseño adoptado fue observacional, la investigación realizada es de tipo ex post facto. Primero se realizó la evaluación (diciembre del año 2019) y luego se analizaron las posibles mejoras de la asignación de calificaciones y su categorización en 9 cajas. Es una investigación donde no se modifica la evaluación del desempeño realizada y se proponen mejoras sobre su metodología.

- Unidades de análisis

La población en estudio es el grupo de trabajadores permanentes de una institución semiautónoma guatemalteca, definidos por su rol dentro del proceso de evaluación de desempeño como evaluadores y evaluados. Estas últimas son las subpoblaciones que se estudian respecto de las calificaciones asignadas.

Se analizó a la población respecto de la fiabilidad y validez del instrumento de evaluación y para los otros aspectos se trabajó con una muestra para poblaciones finitas. La muestra se seleccionó haciendo un muestreo probabilístico aleatorio simple.

Tabla I. **Definición de variables**

Variable	**Definición teórica**	**Definición operativa**
Código del evaluador	Identificador del evaluador	Variable nominal
Nivel del evaluador	Nivel organizacional del evaluador	Variable ordinal
Código del evaluado	Identificación del evaluado	Variable nominal
Nivel del evaluado	Nivel organizacional del evaluado	Variable ordinal
Dependencia	Área organizacional a la que pertenecen el evaluador y el evaluado	Variable nominal
Competencia	Competencia organizacional, existen varias y fueron calificadas con escala de likert	Variable ordinal

Continuación tabla I.

Variable	Definición teórica	Definición operativa
Fiabilidad	Es “la principal característica de los instrumentos de medida, es requerida para garantizar que las interpretaciones sean adecuadas a la realidad estudiada, se dice que un instrumento es fiable cuando mide algo con precisión independientemente de lo que este midiendo” (Pérez, García , Gil, y Galán, 2009) Variable cuantitativa discreta.	Índice de correlación de consistencia interna. Puede medirse de varias formas todas ellas arrojan un resultado porcentual por lo que su escala es de razón.
Calificación	Se refiere a un valor asignado a los resultados del desempeño de una persona respecto de los resultados esperados de la persona en un puesto. Variable cuantitativa discreta.	Variable discreta que puede asumir valores entre 1 a 100, por lo que su escala es de intervalo.

Fuente: elaboración propia**.**

- Fases del estudio

  - Fase 1. Revisión de literatura existente. Con la finalidad de contar con un marco de referencia y orientación de la investigación se realizó una revisión de libros o publicaciones relativas tanto a la evaluación del desempeño como de la metodología de 9 cajas, así como de la metodología estadística aplicable a su análisis.

  - Fase 2. Gestión de la información. En esta investigación no fue necesario hacer recolección de información, se hicieron las gestiones necesarias para obtener acceso a la información, se cuidó el anonimato de los participantes y se guardó la confidencialidad de la institución a la que pertenecen los datos del ejercicio de evaluación analizado.

- Fase 3. Análisis de la información. Se realizó el análisis de la información a través de los siguientes pasos:

  - Se codificó la base de datos para identificar las distintas variables y analizar los datos con una hoja electrónica Excel y el software estadístico Minitab y SPSS.

  - En la primera parte se hizo la evaluación de la fiabilidad y validez del instrumento aplicado para la evaluación del desempeño. Se utilizó el modelo Alfa de Cronbach y de Spearman Brown para medir la consistencia de los aspectos evaluados por el formulario a través de un documento en Excel. No se analizó la validez de contenido debido a que el instrumento fue diseñado por un experto reconocido a nivel nacional en temas de evaluación de desempeño.

  - En la segunda parte de análisis para dar respuesta a la pregunta sobre si existen tendencias extremas de asignación de calificaciones de parte de los evaluadores, primero se hizo un análisis gráfico de la población mediante un histograma, un QQ-plot y una gráfica de valores atípicos y se aplicó una prueba de Grubbs sobre la normalidad de la población. Para confirmar el análisis gráfico se planteó una prueba de hipótesis sobre la igualdad de medias de calificaciones por evaluador. No fue necesario hacer una prueba post hoc, debido a que las acciones que pueden realizarse para mejorar las tendencias de evaluación se deben ejecutar, solo cuando se sabe que existen diferencias.

- Finalmente se diseñó un baremo para categorizar a los trabajadores de acuerdo con las calificaciones que se les asignaron en la evaluación de desempeño en las tres categorías de la metodología de 9 cajas. Para hacer el baremo primero se analizaron los factores sociodemográficos que influyen en la calificación por medio de un análisis de espina de pescado. De acuerdo con el resultado del análisis se planteó la hipótesis sobre el factor que se encontró que afecta significativamente la calificación del desempeño. Para determinar las distintas agrupaciones del baremo se procedió a analizar las diferencias de medias utilizando Bonferroni y finalmente para las categorías en que se detectó que tenían diferencias significativas entre sí, se diseñó el baremo en deciles y estos deciles se asignaron a las categorías de la metodología de 9 cajas.

- Con base en los deciles del baremo, se elaboró una norma de equivalencia de deciles a categorías del eje horizontal de la metodología de 9 cajas.

- El procesamiento de datos, las gráficas y las pruebas fueron hechas utilizando la hoja de cálculo Excel y el software estadístico SPSS y Minitab.

- Fase 4. Interpretación de la información. Con base en los resultados se procedió a realizar un análisis detallado del proceso de evaluación y se contrastaron los resultados de este con otros estudios de referencia.

- Fase 5. Redacción de informe final. El informe final consta del resumen de interpretaciones que será la base de las recomendaciones de mejora y las que darán respaldo estadístico al proceso y se hicieron las recomendaciones pertinentes para la mejor forma de clasificar los resultados en el eje horizontal de la metodología de 9 cajas.

- Técnicas de análisis de información

    - A fin de organizar, analizar y presentar la información se usaron las siguientes técnicas:

    - Histograma de frecuencias: facilita el análisis gráfico del comportamiento de los datos agrupados.

    - QQ plot: muestra gráficamente si los datos se comportan siguiendo una distribución normal.

    - Gráfico de valores extremos, muestra gráficamente si hay valores extremos en la muestra o población analizada.

    - ANOVA, análisis de varianza utilizada para comprobar la existencia de varianzas entre las medias de calificaciones asignadas por evaluadores y también entre niveles organizacionales de las muestras analizadas.

    - Bonferroni, prueba post hoc utilizada para determinar las diferencias entre las calificaciones de niveles organizacionales y definir sus agrupaciones para la elaboración del baremo.

- Identificación de deciles, con el fin de identificar las posiciones que dividen las calificaciones por los niveles agrupados y construir el baremo de las calificaciones del ejercicio de evaluación.

INTRODUCCIÓN

La administración del talento en una organización es imprescindible, de sus colaboradores depende su éxito o fracaso. Las herramientas usualmente utilizadas para su administración son: (1) evaluación del desempeño y (2) determinación de potencial.

Según Kaufman (2009):

> La importancia de la medición del desempeño individual, el éxito organizacional y su aporte al desarrollo de la sociedad, se evidencian con la siguiente declaración: todas las organizaciones son medios para fines sociales. Definir y lograr el éxito organizacional continuo es posible. Se basa en tres elementos básicos: (1) una sociedad con "mentalidad" de valor agregado (su perspectiva y compromiso sobre su organización, personas y nuestro mundo compartido), (2) determinación y acuerdos compartidos sobre hacia dónde dirigirse y por qué (todos los que pueden o podrían ser afectados por los objetivos compartidos debe acordar propósitos, criterios y resultados), y (3) pragmático y básico, herramientas (medición del desempeño). (p. 5)

La mejora del sistema de evaluación del desempeño implica el análisis estadístico del ciclo completo de un ejercicio tipo para lograr reducir la incertidumbre y generar mayor confiabilidad sobre sus resultados. Para este análisis se usaron los descriptores estadísticos de las variables identificadas en el ejercicio, para detectar las que afectan significativamente al mismo. Así también análisis paramétricos que proporcionan orientación sobre el instrumento

utilizado para evaluar, las tendencias de asignación de calificaciones de los evaluadores y las variables sociodemográficas que tienen influencia sobre las calificaciones asignadas.

Esta tesis se dividió en cuatro capítulos. En el primero se presenta el marco referencial en el cual se desarrolla una revisión de la bibliografía que se consideró relevante acerca de los análisis estadísticos y su relación con el tema de evaluación. Se investigó no solo sobre evaluación del desempeño, sino sobre evaluación en general, para que los conocimientos pudieran conferir un marco que sustentara el tipo de análisis cuantitativo empleado. La bibliografía y ensayos que generalmente relacionan la estadística y la evaluación son educativos y psicométricos.

En el capítulo dos se encuentra el marco teórico en el cual se describen las técnicas estadísticas aplicadas para evaluar la confiabilidad del instrumento utilizado para la evaluación, la teoría sobre pruebas de hipótesis, la teoría de diferencias múltiples de medias, la teoría básica sobre diseño de baremos y la metodología de nueve cajas.

El capítulo tres presenta las pruebas e hipótesis realizadas en detalle. (1) prueba de fiabilidad del instrumento mediante el procedimiento de Spearman Brown y Alfa de Cronbach, (2) análisis gráficos de tendencias de evaluación, (3) prueba de hipótesis utilizada para probar la existencia de diferencias significativas entre las tendencias de evaluación de los evaluadores y (4) análisis que condujeron al diseño del baremo para la categorización de las calificaciones obtenidas por los evaluados acorde a los niveles organizacionales que presentaron diferencias significativas de medias, lo que permitió asignar los deciles del baremo a las categorías del eje horizontal de la metodología de 9 cajas.

Y en el capítulo cuatro se presenta la discusión de resultados, tanto interna como externa respecto de la investigación realizada.

1. MARCO REFERENCIAL

En esta sección se presentan los antecedentes revisados y organizados de tal manera que proporcionen la línea general de investigación del trabajo de tesis de graduación. Se tomaron como punto de partida los objetivos planteados y una coherencia de organización para facilitar su desarrollo. Se presentaron fuentes actualizadas cuya sustancia respalde el tema. Algunas de ellas cuentan con más de cinco años de publicación con el propósito de aportar a la reflexión sobre las respuestas que se espera encontrar en este trabajo y facilitar el desarrollo de sus objetivos.

El primer estudio que orienta la investigación es sobre sesgo de escalada del compromiso en la evaluación del desempeño, Leiva (2013), en este estudio se hace evidente la necesidad de hacer un análisis y propuesta de mejora del proceso de evaluación del desempeño y se menciona que la evaluación del desempeño es un proceso vital en la organización. Es un proceso vital, Ilgen et al (1979), debido a que es un elemento que sirve para consolidar la motivación y el desarrollo de los empleados, al mismo tiempo, Skarlicki y Folger (1997), afirman que es una fuente de frustración e insatisfacción ya que el proceso de la evaluación de desempeño puede llegar a considerarse imparcial, político o irrelevante. “Por esta razón es necesario que el proceso de evaluación del desempeño sea lo más transparente posible, para que los trabajadores consideren el proceso como justo y así genere valor para la compañía” (Leiva, 2016, p. 38).

En esta primera parte se exploró la forma apropiada de estudiar la evaluación del desempeño. Realizar el análisis con datos de una

evaluación anterior (ex post facto), es acertado, basado en la siguiente cita:

La expresión ex post facto significa después del hecho, haciendo alusión a que primero se produce el hecho y después se analizan las posibles causas y consecuencias, por lo que se trata de un tipo de investigación en donde no se modifica el fenómeno o situación objeto de análisis. (Hidrugo y Pucce, 2016, p. 77)

- La base sobre los criterios a evaluar (variables) se centró en las interrogantes que, Gorriti (2007) plantea en respuesta a la pregunta: ¿cuándo es válida una evaluación del desempeño?, debe llenar los siguientes criterios:

    - Relevancia: ¿es un comportamiento relevante para la organización, lo que se evalúa?, esto se puede determinar si lo que se mide es mal ejecutado u omitido tiene consecuencias trascendentes para ella.

    - Fiabilidad: ¿la medida utilizada es consistente o estable?, es decir distintos evaluadores tienen el mismo criterio al evaluar y por tanto evalúan de igual manera el mismo comportamiento del mismo trabajador para un mismo período de tiempo.

    - Discriminación: existe una clasificación que permite un ordenamiento en función de la calificación obtenida.

    - Practicidad: que exista claridad sobre lo que se mide y que esta permita credibilidad en el sistema de evaluación de desempeño. (p. 301)

Este artículo orienta la línea de investigación desde el inicio del proceso, pues la fiabilidad comienza con el diseño del instrumento de aplicación de la evaluación de desempeño. Las variables de fiabilidad y validez de contenido son entonces las primeras a evaluar y se encontraron algunas tesis y publicaciones que permiten definir la forma en que se abordará el tema. Estos trabajos provienen del área educativa y de evaluación psicométrica; sin embargo, persiguen el mismo fin, investigar la fiabilidad y validez de un instrumento de evaluación. Específicamente en estos libros se explican indicadores y coeficientes estadísticos que miden la fiabilidad, relevancia y otros aspectos relacionados a la confiabilidad de instrumentos de evaluación, tales como el Coeficiente Alfa de Cronbach que es un modelo de consistencia interna (a mayor valor de Alfa, mayor confiabilidad) basado en el promedio de las correlaciones entre los ítems y se refiere a que tan bien los ítems miden un simple constructo unidimensional. El mayor valor teórico de alfa es 1 y de acuerdo con el criterio ampliamente aceptado de George y Mallery (2003, p. 231)

- El valor 0.7 del coeficiente alfa es aceptable.

    - Al seguir esta línea de investigación sobre evaluación psicométrica y educación, en la cual el evaluado es similar al examinado, un instrumento de medición para evaluar ha de cumplir su objetivo adecuadamente, permitir al examinado demostrar su potencial real, sin que su género, idioma u otras características propias no relacionadas con lo que se mide interfieran con él.

    - Reducir el riesgo de prácticas que interfieran sobre la implementación del instrumento de medida o prueba harán que no se beneficie o perjudique a los trabajadores sometidos a dichas pruebas. Minimizar los riesgos de interferencias de medición hace

que una prueba sea imparcial (Arbazúa, Andrea, 2017). Ello nos conduce nuevamente a la teoría de la estadística aplicada a la educación, que confirma que el hacer la evaluación del instrumento es el primer paso para desarrollar una propuesta de mejora de la evaluación de desempeño: Fox y López (1981) afirman en su libro sobre investigación de prácticas de evaluación educativas, que solamente al tener garantías de que un instrumento es fiable deberíamos analizar sus demás características técnicas.

- Se puede medir la fiabilidad como consistencia interna (procedimiento de las mitades) porque las preguntas de la evaluación del desempeño se pueden separar en dos partes y se espera una coherencia entre ambas, si los resultados esperados son buenos, los comportamientos que llevan a ellos también lo serán.

- De las referencias buscadas se encontró en el libro Estadística aplicada a la educación de Pérez, García, Gil y Galán (2009), que podría evaluarse la validez del instrumento de evaluación de desempeño con varios procedimientos tales como el de Kuder Richardson. Este procedimiento, por ejemplo, puede obtener no solo el coeficiente de consistencia interna, sino también el de homogeneidad.

Igualmente se explorará el uso de una prueba de longitud, utilizando la ecuación general de *Spearman-Brown* que, de acuerdo con Brown (1999): “La ecuación estima la confiabilidad de una prueba a partir de la confiabilidad por mitades”. (p. 88)

Para verificar la validez de constructo se halló un estudio previo hecho por Gutiérrez, Cabreo y Estrada (2017), sobre la validez de las pruebas sobre la competencia digital de los alumnos en las universidades españolas, en el cual se explica que la validez se comprobó al consultar 17 expertos que dominan el tema y son reconocidos por ello, adicionalmente realizaron un análisis factorial exploratorio sobre los factores principales del tema, con rotación normalización varimax con Kaiser.

En la segunda parte de la investigación se estudia específicamente la influencia del evaluador y su estilo de calificación, se pretende detectar si es necesario mejorar algún aspecto de esta parte del proceso de evaluación de desempeño, siempre y cuando se compruebe si existe o no independencia o influencia entre el evaluador y las calificaciones que asigna. Para esta parte se buscaron estudios experimentales relativos a la dependencia de evaluación de desempeño y dependencia de variables. De los estudios investigados, uno dedica un capítulo a cómo seleccionar la mejor forma de encontrar esta asociación e indica que una forma de explicar la dependencia entre dos variables aleatorias, es utilizando la correlación, para lo cual debemos cuidar los sistemas de medición pues dependen de la escala de medidas usada. El autor hace énfasis en que los coeficientes deben de cuidarse: “La selección de un determinado coeficiente de correlación está en dependencia de la escala de medidas usadas, el tipo de problema a resolver y los objetivos propuestos” (Rojas, 2007, p. 10). Si se cuidan los aspectos antes mencionados, se permitirá hacer las comparaciones entre muestras o poblaciones de diferentes tamaños.

Sin embargo, también están las tablas de contingencia: “La determinación del tamaño de muestra en las de tablas de contingencias varía según sea el objetivo: a) Determinar probabilidades de incidencias. b) Docimar

independencias entres dos variables. c) Analizar la asociación entre las variables" (Rojas, 2007, p. 52).

La última parte del estudio es dedicado a la interpretación de las calificaciones que se obtienen luego de aplicar el instrumento de evaluación de desempeño y se vuelve a la investigación educativa y evaluación psicométrica, específicamente a la aplicación de baremos o normas. Esto se hace necesario cuando queremos categorizar los resultados y utilizar la metodología de 9 cajas, que viene a ser una forma de clasificación establecida en recursos humanos que carece de un baremo o norma para clasificación.

Martínez (2004), en su trabajo sobre Elaboración de baremos de calificación en Educación Física con la hoja de cálculo Excel 2000, explica que interpretar de manera correcta los resultados de una evaluación; es decir, de forma objetiva, nos conduce a interpretar de manera adecuada. Esta interpretación debe hacerse a través de la construcción de un baremo, pues el juicio personal no es suficiente para la clasificación de los resultados obtenidos en pruebas. Para reafirmar que la interpretación diagnóstica adecuada es necesaria a continuación se cita de la publicación Matrices Progresivas de Raven el Efecto Flynn y actualización de baremos: "Para que la interpretación diagnóstica sea correcta los baremos deben estar actualizados, es decir, el grupo de referencia del cual se obtienen las puntuaciones promedio con las que se compara el rendimiento de un sujeto, debe ser el adecuado" (Rossie et. al., 2014, p. 3).

Gamboa y Heredia (2017), indican que el Modelo 9 Cajas (Nine Box) es una de las herramientas más utilizadas para el desarrollo y crecimiento de trabajadores en las empresas, una herramienta importante para la gestión del talento. El modelo de 9 cajas se cree que se originó a finales de los años 60's en

la General Electric Company con el propósito de incentivar el desarrollo del potencial de sus empleados. Se basó en una herramienta parecida que fue creada por Boston Consulting Group de nombre Boston Box que mide el potencial de un producto o servicio. En su tesis concluyen que el papel que juega la evaluación de desempeño y la clasificación del talento derivada de ella son determinantes para el futuro exitoso de la carrera de un trabajador y de la empresa misma. Se evidencia en el párrafo anterior la relevancia del uso de una metodología en la clasificación del talento más allá de un juicio personal, que es lo que actualmente se utiliza, para la asignación de calificaciones resultante de la evaluación del desempeño en los cuadrantes del modelo de 9 cajas.

En resumen, el análisis de los antecedentes nos lleva a comparar la evaluación de desempeño a la evaluación educativa y psicométrica, dónde se encuentra la base teórica conceptual que nos conducirá a cumplir con el objetivo del presente estudio.

2. MARCO TEÓRICO

2.1. Análisis estadístico de la evaluación de desempeño

Para abordar el análisis de la evaluación del desempeño estadísticamente es imprescindible citar desde la teoría de evaluación de desempeño, los antecedentes teóricos sobre cómo hacerlo.

Según Gorriti (2007):

- ¿Cuándo es válida una evaluación de desempeño? los criterios para saber si una evaluación de desempeño es válida, se basan por acuerdo de varios autores que este juicio ha de hacerse acorde a los siguientes criterios:

    - Relevancia: ¿es un comportamiento relevante para la organización, lo que se evalúa?, esto se determina si lo que se mide es mal ejecutado u omitido tiene consecuencias trascendentes para ella.

    - Fiabilidad: ¿la medida utilizada es consistente o estable?, es decir distintos evaluadores tienen el mismo criterio al evaluar y por tanto evalúan de la misma forma el mismo comportamiento del mismo trabajador para un mismo período de tiempo.

    - Discriminación: existe una clasificación que permite un ordenamiento en función de la calificación obtenida.

- Practicidad: claridad de lo que se mide que permita credibilidad en el sistema de evaluación de desempeño. (p. 301)

La propuesta de validez de Gorriti Bontegui es la base tomada para definir las variables a estudiar fiabilidad, discriminación (catalogación de resultados) y practicidad (validez de la prueba o instrumento de evaluación). Para cada variable, se investigaron las mediciones apropiadas en libros y publicaciones que hablan de su aplicación en el campo de la evaluación en general. Debido a que es escaza la teoría y estudios que tengan un enfoque estadístico en el análisis de la evaluación de desempeño, la mayoría de la literatura citada es acerca de la aplicación y análisis de pruebas estadísticas para la evaluación de aprendizaje y aplicación de pruebas psicométricas. El contenido presentado a continuación proviene de fuentes de esta índole.

2.1.1. Fiabilidad

De acuerdo con varios autores los instrumentos tienen valor únicamente si son confiables. “Antes de que una prueba pueda utilizarse con cierta seguridad, debe obtenerse información acerca de su confiabilidad y validez por lo que a sus propósitos específicos concierne” (Aiken, 2003, p. 85).

Entre ellos se seleccionaron los que se utilizaron en este estudio.

- Procedimientos de medición para determinar fiabilidad: la fiabilidad puede medirse como estabilidad, esta primera forma de medirla es conocida como procedimiento de repetición o retest, puesto que busca la correlación por un mismo grupo de evaluados en dos ocasiones distintas sobre la misma prueba, se acostumbra, aunque no es regla dar de 20 a 25 días entre pruebas.

- Una segunda forma es medirla como equivalencia, utilizando para ello la aplicación de dos pruebas que miden lo mismo, ambas por tanto se correlacionan y se mide la correlación o equivalencia entre ambas.

- La tercera forma es como consistencia interna, esta mide la coherencia o consistencia en las respuestas que ofrecerán un único grupo de estudio a los distintos elementos que integran tal instrumento de evaluación. En este estudio por la base de datos disponible se medirá la fiabilidad como consistencia interna.

- Medición de la fiabilidad como consistencia interna: se utilizan pruebas paralelas para distintos fines, esta forma de evaluar es cara y difícil de realizar. Se han diseñado distintos procesos que no son realmente equivalentes, pero que se toman como válidos y que se enumeran a continuación:

- Procedimiento Spearman Brown: se describe en el libro de Aiken (2003) de la siguiente forma, utilizando el método de división por mitades:

- “Este enfoque simplificado de la consistencia interna una sola prueba se considera compuesta de dos partes (formas paralelas) que miden la misma cosa”. (p. 87), “Suponiendo que las dos mitades equivalentes tienen medias y varianzas iguales, por tanto, la confiabilidad de la prueba como un todo puede estimarse mediante la fórmula de Spearman Brown” (p. 88).

$$r_{11} = \frac{2r_{oe}}{1 + r_{oe}} \qquad \text{(Ec. 01)}$$

r_{oe} Se calcula mediante el coeficiente de correlación de Pearson entre las mitades, se nombra o a las puntuaciones de las preguntas impares (1ª. Mitad) y e a la suma de las preguntas pares (2ª. Mitad).

- Procedimiento de Rulon**:** este procedimiento está basado en la varianza de las diferencias. Los sujetos mantienen la misma puntuación en las mitades de la prueba deberían mantener una correlación perfecta, fiabilidad perfecta. Esto se expresa en la siguiente fórmula de cálculo:

$$r_{xx} = 1 - \frac{s_d^2}{s_t^2} \qquad \text{(Ec. 02)}$$

El cálculo de s_d^2 y s_t^2 se realiza según las siguientes ecuaciones:

$$S_d^2 = \frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1} \qquad \text{(Ec. 03)}$$

$$s_t^2 = \frac{\sum T^2 - \frac{(\Sigma T)^2}{n}}{n-1} \qquad \text{(Ec. 04)}$$

Donde d es la resta entre la primera y segunda mitad de los resultados, n es la mitad de los datos y T es la diferencia entre el cuadrado de la suma de las calificaciones de la primera y la segunda mitad.

Procedimiento de Guttman: este procedimiento es basado en la varianza de las dos mitades, de tal forma que a menor valor de las varianzas más elevada es la fiabilidad del instrumento de evaluación.

$$r_{xx} = 2\left(1 - \frac{s_{1^\alpha}^2 + s_{2\alpha}^2}{S_1^2}\right) \quad \text{(Ec. 05)}$$

Donde las varianzas para las mitades se calculan de la siguiente forma:

$$s_{1\propto}^2 = \frac{\sum X_1^2 \quad - \frac{(\Sigma X_1)^2}{n}}{n-1} \quad \text{(Ec. 06)}$$

$$s_{2\propto}^2 = \frac{\sum X_2^2 \quad - \frac{(\Sigma X_2)^2}{n}}{n-1} \quad \text{(Ec. 07)}$$

Siendo X la notación de las calificaciones en la primera y segunda mitad y n el total de evaluaciones (Pérez, García, Gil y Galán, 2009).

- Procedimiento de Kuder-Richardson: de acuerdo con Aiken (2003), la base de esta prueba consiste en que "una prueba puede dividirse de muchas formas diferentes en dos mitades que contengan igual número de reactivos, como cada forma de división en mitades puede dar por resultado un valor distinto de r_{11}, no queda claro qué estrategia de división producirá el mejor estimado global de confiabilidad. Esto se puede hacerse bajo el siguiente procedimiento abreviado que fue elaborado por Kuder Richardson (1937)". (Aiken, 2003, p. 88):

$$r_{11} = \frac{k[1 - \sum pi(1 - Pi), s^2]}{k-1} \quad \text{(Ec. 08)}$$

$$r_{\|} = \frac{k - \underline{x}\left(k - \underline{x}\right) / s^2}{k-1} \quad \text{(Ec. 09)}$$

Donde k es el total de preguntas de la prueba, $\underline{X}$ la media de las calificaciones de la prueba s^2 es la varianza de las calificaciones de la prueba y P_i.es el número que dan la respuesta acertada a la pregunta.

- Procedimiento Alfa de Cronbach (**α**): es el cálculo de un coeficiente que sirve como medida de homogeneidad.

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum s_t^2}{s_x^2}\right] \qquad \text{(Ec. 10)}$$

k Es el total de preguntas de la prueba, s_t^2 es la varianza de los resultados de la prueba, s_x^2 es la varianza de las calificaciones de la prueba. El coeficiente de la correlación entre dos pruebas de la misma longitud tomadas de una misma muestra (Brown, 1999).

2.1.2. Análisis de varianza

El análisis de varianza es un modelo utilizado para pruebas estadísticas, generalmente conocido como ANOVA. El fin de este análisis es definir si existen diferencias significativas en la media de 2 o más muestras (Pulido, 2018).

El ANOVA compara tratamientos en cuanto a medias poblacionales y sus varianzas.

La hipótesis por probar cuando se comparan varios tratamientos con el ANOVA es:

H_0: $\mu_1 = \mu_2 = \cdots \mu_n$

H_1: $\mu_i \neq \mu_j$ Para algún $i \neq j$

Para validar esta hipótesis y decidir si los tratamientos son iguales estadísticamente en cuanto a sus medias, al tener la alternativa de que dos de ellas sean diferentes, se debe obtener una muestra representativa de las mediciones de cada uno de los tratamientos y construir un estadístico de prueba para poder definir el resultado de la comparación.

Para ello se utilizan diseños completamente al azar y el ANOVA, cuyos principios de aplicación se explican a continuación:

- Supongamos que se tienen k tratamientos o poblaciones independientes, cada una con medias desconocidas $\mu_1, \mu_2, \ldots \mu_k$ y varianzas desconocidas $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$, las poblaciones pueden ser k métodos de producción, k evaluadores en el caso que estudiamos, entre otros; y sus respectivas medias, todos ellos con respecto a la variable de respuesta (Pulido, 2018).

- Se realiza un experimento completamente al azar para comparar las poblaciones, mediante la hipótesis inicial planteada de igualdad de medias.

Los datos estudiados se verían como en la siguiente tabla:

Tabla II. **Ejemplo de experimento**

Tratamientos				
Trat. 1	Trat. 2	Trat. 3	…	Trat. k
Y_{11}	Y_{21}	Y_{31}	…	Y_{k1}

Continuación tabla II.

.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Y_{1n_1}	Y_{2n_2}	Y_{3n_3}	…	Y_{kn_k}

Fuente: elaboración propia.

El elemento del cuadro X, Yij, en este cuadro es la j-ésima observación que se hizo en el tratamiento i; ni es el tamaño de la muestra o las repeticiones observadas por cada tratamiento. Cuando ni = n para todas es un diseño balanceado, el diseño puede no ser balanceado.

El número de tratamientos K es acorde al criterio del investigador y el n número de observaciones por tratamiento debe escogerse con base en la variabilidad que se espera observar en los datos y la diferencia mínima que se espera detectar. Con estas consideraciones se podrán describir como el modelo estadístico lineal descrito por Pulido (2018):

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \qquad \text{(Ec. 11)}$$

En el cual μ es el parámetro de escala común para todos los tratamientos o media global τ ; es un parámetro que mide el efecto del tratamiento i y ε_{ij} es el error atribuible a la medición Y_{ij} este diseño completamente al azar nos explica dos fuentes de variabilidad, el de los tratamientos y el error aleatorio. La media común μ no se considera una fuente de variación por ser una constante común a todos los tratamientos y sirve de fuente de referencia con respecto al cual se comparan las respuestas de las medias de los tratamientos, si existe una

diferencia muy distinta de la media global, es un indicador de que existe un efecto de dicho tratamiento. La diferencia que deben tener las medias entre sí para determinar si los tratamientos son diferentes nos lo dice el análisis de varianza (ANOVA) (Pulido, 2018).

El ANOVA para diseños completamente al azar se centra en la idea de separar la variación total entre las partes con la que contribuye cada fuente de variación en el experimento. Cuando es un diseño completamente al azar, se separan estas variaciones y la debida al error. Para probar la hipótesis del ANOVA, se descompone la variabilidad total de datos y la que corresponde al error, tal como se describe a continuación.

La primera medida de variabilidad total presente en las observaciones en el cuadro X es la suma total de cuadrados dados por Pulido (2018):

$$\sum_{i=1}^{k} \quad \sum_{j=1}^{n_j} \quad \left(Y_{ij} - \underline{Y}..\right)^2 = \sum_{i=1}^{k} \quad \sum_{j=1}^{n_j} \quad Y_{ij}^2 - \frac{Y^2..}{N} \qquad \text{(Ec. 12)}$$

Donde Y es la suma de los N=$\sum_{i=1}^{n_i} \quad n_i$ datos del experimento. Al sumar y restar dentro del paréntesis la media de tratamiento i, $\left(\underline{Y_{j.}}\right)$:

$$SC_T = \sum_{i=1}^{k} \quad \sum_{j=1}^{n_j} \quad \left[\left(Y_i - \underline{Y_{i.}}\right) + \left(Y_{i.} - \underline{Y}..\right)\right]^2 \qquad \text{(Ec. 13)}$$

El cuadrado de SCT se descompone en los siguientes componentes (Pulido, 2018):

$$SC_T = \sum_{i=1}^{k} \quad n_i \left(\underline{Y_{i.}} - \underline{Y}..\right)^2 + \sum_{i=1}^{k} \quad \sum_{j=1}^{n_j} \quad \left(Y_{ij} - \underline{Y_{i.}}\right)^2 \qquad \text{(Ec. 14)}$$

En el cual el primer componente es la suma de cuadrados del tratamiento (SCTrat) y el segundo la suma de cuadrados del error (SCE). Al observar con detalle estas sumas de cuadrados se aprecia que la SCTRAT mide la variación o diferencias entre tratamientos, puesto que si son muy diferentes entre sí la diferencia tenderá a ser grande en valor absoluto, mientras que la SCE mide la variación dentro de tratamientos. De forma abreviada esta descomposición será la siguiente (Pulido, 2018):

$$SCT = SCTRAT + SCE \qquad \text{(Ec. 15)}$$

Los grados de libertad vienen dados por el total de $N=\sum_{i=1}^{n_i} n_i$ observaciones, la SCT tienen N-1 grados de libertad. Hay k tratamiento o niveles del factor de interés, así que SCTrat tiene k-1 grados de libertad y SCE tiene N-k grados de libertad, que se resume en (Pulido, 2018):

$$N - 1 = (k - 1) + (N - k) \qquad \text{(Ec. 16)}$$

Las sumas de cuadrados divididas entre sus grados de libertad se llaman cuadrados medios, denotados por Pulido, (2018):

$$CMTrat = \frac{SC_{trat}}{k-1} \qquad \text{(Ec. 17)}$$

$$y \quad CME = \frac{SC_E}{k-1} \qquad \text{(Ec. 18)}$$

Podemos observar en estas ecuaciones, que cuando la hipótesis nula es verdadera, ambos cuadrados medios estiman la varianza σ^2, al basarse en esta premisa se construye el estadístico de prueba como sigue Pulido, (2018):

$$F0 = \frac{CM_{Trat}}{CM_E} \qquad \text{(Ec. 19)}$$

Este sigue una distribución F con (k-1) grados de libertad en el numerador y (N-k) en el denominador. De aquí se puede deducir que, si F0 es grande, se contradice la hipótesis de que no hay efectos de tratamientos; de lo contrario si F0 es pequeño se confirma la validez de H0. Por lo que para un nivel de significancia ∝ (alfa) prefijado:

Tabla III. **Reglas comprobación de hipótesis ANOVA**

Si	donde	¿Se rechaza H0?
$F_0 > F_{\alpha,k-1,N-k}$	$F_{\alpha,k-1,N-k}$ es el percentil (1-α) x100 de la distribución F	si
El valor-p <α	Donde el valor-p es el área bajo la distribución $F_{k-1,N-k}$ a la derecha del estadístico F0 lo que es igual a decir que el valor-p= P(F>F0)	si

Fuente: elaboración propia.

La información que se necesita para el cálculo del estadístico F0 hasta llegar al valor-p se escribe en la tabla de análisis de varianza (ANOVA).

Tabla IV. **ANOVA para diseños completamente al azar**

FV	SC	GL	CM	F0	Valor-p
Tratamientos	$SC_T = \sum_{i=1}^{k} \frac{Y_{j\cdot}^2}{n_i} - \frac{Y_{\cdot\cdot}^2}{N}$	k-1	$CM_{Trat} = \frac{SC_{Trat}}{k-1}$	$\frac{CM_{Trat}}{CM_T}$	P(F>F0)
Error	SCE = SCT - SCTRAT	N-k	$CM_T = \frac{SC_E}{k-1}$		
Total	$\sum_{i=1}^{k} \sum_{j=1}^{n_j} Y_{ij}^2 - \frac{Y^2\cdot\cdot}{N}$	N-1			

Fuente: elaboración propia.

2.1.3. Comparaciones múltiples de medias

Cuando se estudia el comportamiento de tratamientos de un factor mediante un ANOVA, solamente podemos llegar a saber si en general los tratamientos son diferentes entre sí. Una vez se acepta la existencia de diferencias entre los efectos de los tratamientos y si interesa conocer qué tratamientos tienen diferencias entre sí o cuales tienen mayores efectos sobre la variable de respuesta, se hace útil hacer comparaciones múltiples de medias. Las comparaciones múltiples dan respuesta sobre cuáles son las medias diferentes y estimar su grado de diferencia.

Existen distintas técnicas para hacer estas comparaciones múltiples de medias, que son llamadas contrastes para comparaciones múltiples debido a que su objetivo es comparar entre sí las medias de tratamientos o grupos de ellas.

El primer procedimiento básico es el análisis gráfico, cuando este se puede realizar. Puede considerarse luego el método por parejas de Fisher, esta técnica se denomina método de la diferencia mínima significativa (LSD, *Least Significant Difference*), es para diferencias entre parejas. Cuando el número posible de diferencias es alto, existen otros procedimientos para este fin, uno de los más conservadores es el de la desigualdad de Bonferroni (HSU, 1996).

2.1.3.1. Método de Bonferroni

Para este procedimiento debe fijarse un nivel de significancia α que se reparte entre las comparaciones que se toman en cuenta y se utiliza la desigualdad de Bonferroni (HSU, 1996).

$$Pr(\cup_{m=1}^{M} Am) \leq \sum_{m=1}^{M} \quad Pr\ Pr\ (A_m)\ . \qquad \text{(Ec. 20)}$$

Se hace la estimación por intervalos para las M=$(I\ 2\ )$ comparaciones posibles, cada una con un nivel de significancia de α* = α/M origina M intervalos de confianza que contiene cada uno las probables diferencias μi - μj con probabilidad 1-α*. Se llama Cm al intervalo m-ésimo se tiene que (HSU, 1996):

$$\text{Pr= [μ1m - μ2m} \in \text{Cm] = 1-α*, donde m= 1, 2, …, M} \qquad \text{(Ec. 21)}$$

Al aplicar la desigualdad de Bonferroni (HSU, 1996):

$$Pr(\cup_{m=1}^{M} Cm) = 1 - Pr\left(\cup_{m=1}^{M} \underline{C}m\right) \geq 1 - \sum_{m=1}^{M} \quad Pr\ Pr\ \left(\underline{C}m\right)\ = 1 - \sum_{i=1}^{M} \quad \propto^{*} \text{(Ec. 22)}$$

Por $\underline{C}m$ se identifica al complementario del intervalo Cm

De acuerdo con el resultado y para garantizar el nivel de significación α para el conjunto de las M comparaciones por parejas o nivel de confianza 1- α para el conjunto de intervalos se toma:

$$\alpha^{*} = \frac{\alpha}{M} \qquad (Ec. 23)$$

La probabilidad de que todos los intervalos Cm tengan la correspondiente diferencia de medias será el nivel de confianza, de esta manera los intervalos quedan de la siguiente forma (HSU, 1996):

$$\underline{y_1}m - \underline{y_2}m \pm t_{\pi}^{\alpha}\sqrt{\hat{s}_R^2\left(\frac{1}{n1m} + \frac{1}{n2m}\right)} \qquad \text{(Ec. 24)}$$

En la cual $\underline{y_1}m$, $\underline{y_2}m$ y n1m, n2m, son las medias y los tamaños correspondientes a la comparación m-ésima.

Se identifica por $\theta_m = \mu_{1m} - \mu_{2m}$ donde m=1, 2, … M, una de las M comparaciones lineales por parejas de medias, que queremos contrastar.

H0: $\theta_m = 0$ vs. H1: $\theta_m \neq 0$ => se rechaza H0 si $|\theta_m| = B_m$

Y se acepta H0 en el caso contrario en el cual (HSU, 1996):

$$Bm = t\frac{\alpha}{2M}\sqrt{\hat{S}_R^2\left(\frac{1}{n_{1m}} + \frac{1}{n_{2m}}\right)} \qquad \text{(Ec. 25)}$$

En el caso del modelo equilibrado los valores Bm coinciden y son denotados por BSD (HSU, 1996):

$$BSD = t\frac{\alpha}{2M}\sqrt{\hat{S}_R^2\left(\frac{1}{n} + \frac{1}{n}\right)} \qquad \text{(Ec. 26)}$$

Donde n es el número de observaciones de cada grupo y$t_{\propto/2M}$ el valor crítico de la distribución t, en donde la varianza residual tiene el mismo número de grados de la distribución, dejando una probabilidad α/2M a su derecha. Al ser un modelo no equilibrado los valores estadísticos de contraste vienen dados por la siguiente expresión (HSU, 1996):

$$Bm = t\frac{0.025}{10}; 21\sqrt{4.67\left(\frac{1}{n_{1m}} + \frac{1}{n_{2m}}\right)} \text{ donde m=1, 2, …,10} \qquad \text{(Ec. 27)}$$

2.2. Interpretación de calificaciones del desempeño

La categorización de los resultados de la evaluación depende principalmente de la validez y fiabilidad que proporciona el instrumento de medición y de la interpretación de los valores que arrojan esos instrumentos. “La interpretación significativa de las calificaciones de las pruebas requiere tanto medios para expresar esas calificaciones (o sea, una escala) como datos de validez que indiquen lo que mide la a prueba” (Brown, 1999, p. 216).

El desarrollo de escalas y formas de calificación deben ser puramente estadísticos, esto hace necesario transformar las calificaciones brutas a otra escala, se utilizan para hacerlo distintas formas, cada una de ellas con sus ventajas y limitaciones. Existen tres formas de hacerlo, la primera es basada en la interpretación por comparación con otras personas, que son llamadas normas de grupo, son relativas a un grupo de comparación; la segunda está basada en calificaciones expresadas en grados de destreza o dominio de habilidades, en este caso se hace la comparación sobre el contenido que evalúa la prueba y la tercera es sobre calificaciones relacionadas con un resultado global, que básicamente califica el mismo sobre un criterio externo, estas tienen la ventaja de combinar la validez con los datos normativos.

En esta tesis se estudian normas de grupos, puesto que se hace la interpretación sobre una muestra o grupo estándar de referencia. Generalmente hay cierto número de grupos normativos posibles y lo primero es definir cuáles son los posibles grupos normativos. Estos grupos serán las variables que influyen en los resultados de la prueba. Si los grupos son para pruebas de capacidades generales o características de personalidad, los grupos están compuestos por personas de la misma edad o niveles educativos y si pensamos en términos generales, las características sociodemográficas pueden constituir estos grupos

normativos. Estos pueden ser, edad, sexo, grado educativo, educación, zona geográfica, posición socioeconómica, entre otros. Los grupos normativos son las variables que estadísticamente pueden tener influencia sobre los resultados finales de la prueba y deben de ser caracterizados o descritos perfectamente para ubicar el grupo fácilmente, esto debido a que de esta caracterización pueden surgir subgrupos que harán que existan normas separadas para cada uno de ellos por su rendimiento en las pruebas. Identificar al grupo normativo es sumamente importante para definir normas de calificaciones, por lo que la determinación del tamaño de la muestra, el procedimiento de muestreo y selección aleatoria es fundamental para la realización de normas. Es de tomar en cuenta la validez de las normas, por los cambios en las características de la población, por lo que debería estar constantemente validándose las características de los grupos normativos a fin de mantener la validez de la norma definida. Los principios básicos de las normas generalmente aceptados son:

- Que los subgrupos sean bien definidos.

- Usar normas de subgrupos separados cuando las poblaciones tengan características notoriamente diferentes (utilizar para ello análisis estadísticos que así lo confirmen).

- Indicar los datos normativos útiles, cuatro, usar normativos específicos para la población estudiada.

- Publicar los datos normativos a los usuarios de las pruebas.

- Utilizar los datos normativos para interpretar las calificaciones.

A continuación, se explican las cualidades de las distintas formas de categorización como tabla de especificaciones, tabla empírica de expectativas, los baremos y los pasos a seguir para su elaboración. Todas las anteriores son tablas usadas con el fin de selección ocupacional y son aplicables a la evaluación de desempeño. (Aiken, 2003)

Al final se incluye la metodología de 9 cajas que es utilizada en recursos humanos para clasificar el talento de las organizaciones.

2.2.1. Tabla de especificaciones

La tabla de especificaciones proporciona características y aspectos propios de la prueba que permiten describir el constructo. Para algunos autores esta es una condición que deben cumplir pruebas de alta calidad técnica. (Jiménez y Eliana, 2013)

Una alternativa para la construcción de la tabla de especificaciones es el modelo de *Rasch*, que está enfocado a la psicometría válida e interpreta instrumentos de medición de las ciencias del comportamiento. (Cerdas y Montero, 2017)

Puede utilizarse también el análisis factorial exploratorio que es una técnica muy utilizada para el desarrollo, validación y adaptación de instrumentos de medida psicológicos. Para ello se tienen cuatro pasos principales son:

- Paso uno: determinar el tipo de datos y la matriz de asociación
- Paso dos: estimación de factores
- Paso tres: determinar el número de factores a retener
- Paso cuatro: luego usar la rotación y asignación de ítems

(Lloret et al., 2014)

2.2.2. Tabla empírica de expectativas

Según Aiken (2003), Expone que “con propósitos de selección no es esencial determinar la correlación prueba-criterio ni la ecuación de regresión que vincula el desempeño en la variable de criterio con las calificaciones”. (p. 101)

Las tablas elaboradas de esta forma pueden usarse utilizando métodos correlaciónales, pero pueden no necesitar este método, solo utilizar frecuencias y porcentajes.

2.2.3. Baremos

Según la RAE un baremo es un cuadro gradual establecido convencionalmente para evaluar los méritos personales, la solvencia de empresas entre otros. O los daños derivados de accidentes o enfermedades. Sus elementos o componentes son:

- La puntuación más baja y alta posible
- Regla de medida
- Propio contenido
- Población de destino

Los baremos han de ser representativos de la población estudiada, generalmente se usan muestreos debido a las limitaciones económicas y de tiempo.

2.2.4. Construcción de baremos o normas

La construcción de baremos o normas sigue los métodos de evaluación de instrumentos de medida y su correcto uso. En este caso específicamente se aplica a los instrumentos de evaluación de desempeño y se propone como una metodología de estandarización de normas de interpretación.

Previo a la obtención de las normas, se hace necesario determinar los factores sociodemográficos de la población que inciden en las calificaciones de la población. Esto idealmente debería hacerse sobre la población; sin embargo, esto se hace imposible la mayoría de las veces por limitaciones de tiempo e inversión en los estudios, por lo que se recomienda obtener una muestra representativa de la población y determinar las variables sociodemográficas que pudiesen incidir sobre el punto final obtenido en la evaluación. Una vez determinadas estas variables sociodemográficas se puede elegir el estadígrafo o modelo para poder determinar estadísticamente si las variables seleccionadas inciden significativamente sobre los resultados de la prueba estudiada, estos métodos pueden ser desde el más sencillo al más complejo, depende de cada caso:

- Comparación de medias independientes mediante la prueba de t de *student*
- Test para datos relacionados, t de *student*
- Análisis de varianza de un factor, para más de dos grupos
- Análisis de varianza de múltiples factores

- Pruebas no paramétricas

Análisis factorial, para evaluar si existen grupos de variables sociodemográficas que inciden sobre la calificación final y que dimensionalmente sean distintos entre sí es utilizado este tipo de análisis, que se separa análisis unidimensional, bi o multidimensional.

Luego de comprobar que existe diferencia entre medias, se procede a hacer un análisis de diferencias entre grupos, a través de la comparación múltiple de medias, cuya teoría se explica en el numeral 2.1.3.1. Con base en las diferencias de las medias entre grupos y su significancia se define cuantas tablas tendrá el baremo y se hace uno por cada grupo que tenga diferencias significativas. La prueba post hoc a utilizar, depende del experimentador.

Finalmente se procede a elaborar los baremos por grupo y se relacionan las calificaciones con las normas. Generalmente se utilizan cuatro clases diferentes de calificaciones relacionadas con las normas. Estas calificaciones deben usarse de acuerdo con su fundamento y cálculo para que de esa manera se pueda saber sus ventajas y limitaciones.

Percentiles: este es el método más utilizado, son definidos como el rango que separa el porcentaje de personas que obtienen calificaciones menores. Ejemplo: un rango percentil 65 indica que el 65 % de las personas del grupo normativo obtuvieron calificaciones más bajas; es decir, el rango percentiliar indica la clasificación relativa de la persona en porcentajes.

Los rangos percentiliares se obtienen al determinar la proporción de personas con calificaciones menores al rango calculado dentro del grupo normativo. La facilidad de interpretarlos es la mayor ventaja del uso de percentiles

porque es una escala ordinal, sus desventajas son que tienen una distribución rectangular en que la distribución se aproxima a una curva normal por lo que cuando hay pequeñas diferencias de calificaciones brutas cerca de la media existen grandes diferencias percentiliares, y en los extremos, al contrario.

- Para poder calcular estos rangos percentiliares se sigue el siguiente procedimiento:

    - Preparar la distribución de frecuencias de las calificaciones

    - Calcular la frecuencia acumulada (FA) al límite inferior de cada calificación

    - Calcular la frecuencia acumulada en el punto medio del intervalo de las calificaciones (FApm)

    - Determinar la proporción acumulada (PA) = (FApm)/(FA)

    - El cálculo del rango percentiliar (RP) se hace multiplicando PA por cien

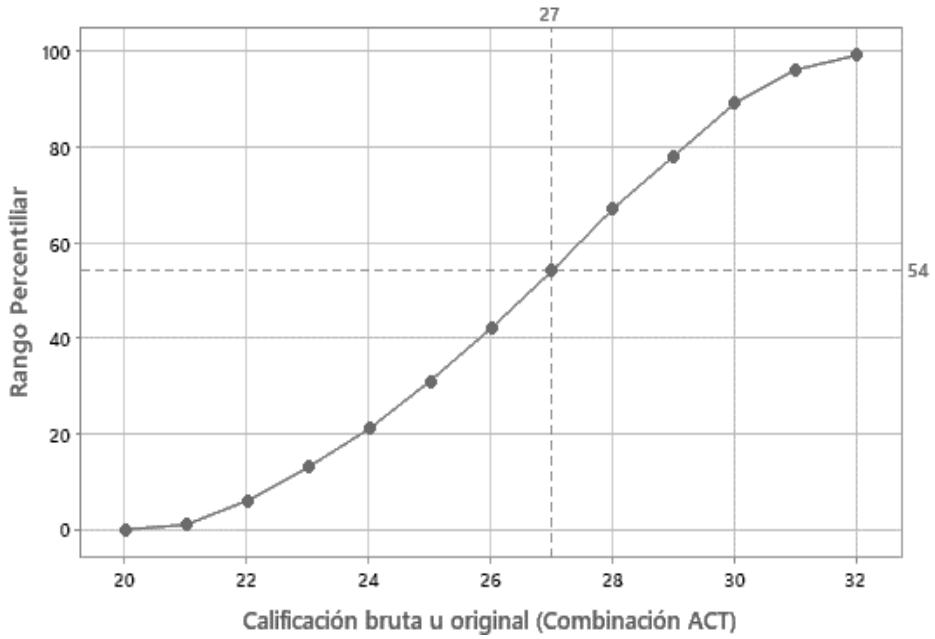
Ejemplo:

Tabla V. **Cálculos de percentiles**

x	f	FA	FApm	PA	RP
32	4	173	175	0.989	99
31	7	166	169.5	0.958	96
30	17	149	157.5	0.89	89
29	22	137	138	0.78	78
28	18	109	118	0.667	67
27	23	81	95	0.537	54
26	15	66	73.5	0.415	42
25	22	44	55	0.311	31
24	14	30	37	0.209	21
23	14	16	23	0.13	13
22	12	4	10	0.056	6
21	3	1	2.5	0.014	1
20	1	0	0.5	0.003	<1

Fuente: elaboración propia, realizado con Adaptación de Brown (1999).

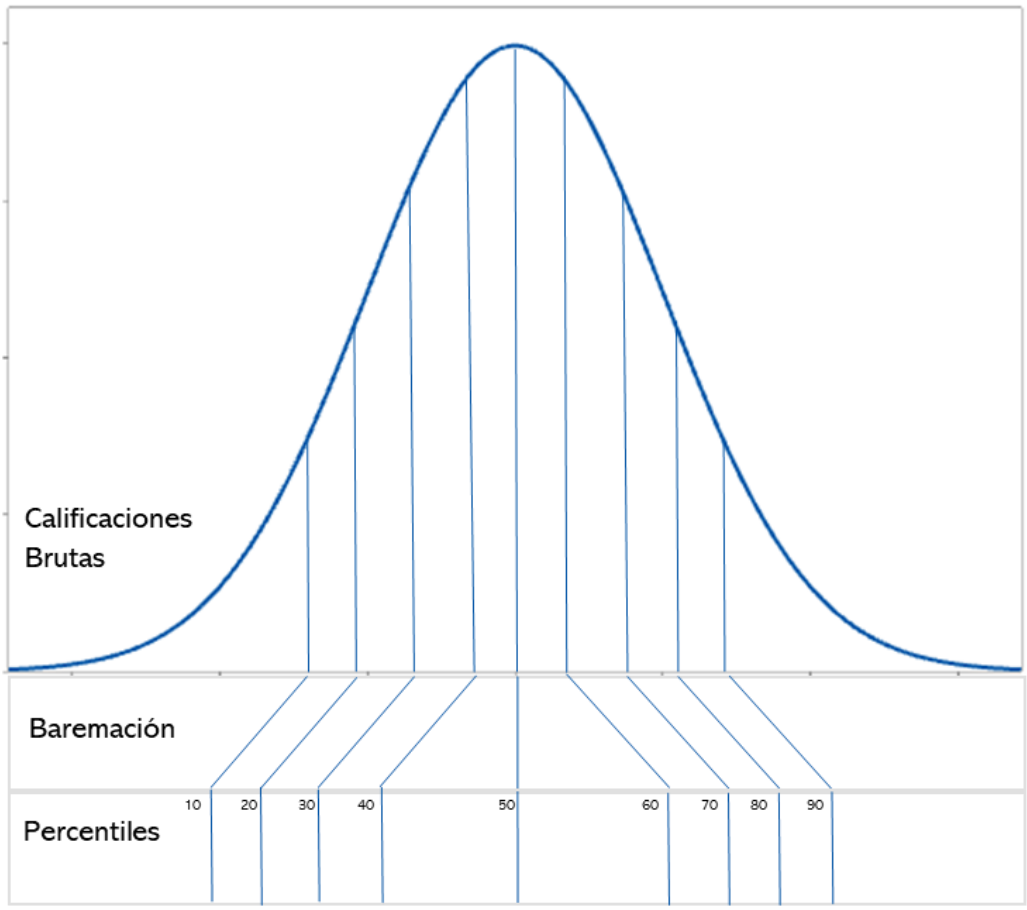
Figura 1. **Calificación bruta vs. Rango percentiliar**



Fuente: elaboración propia, con datos obtenidos de Brown (1999), *Principios de la medición en psicología y educación.*

En la gráfica se puede ver para la calificación bruta de 27 el rango percentiliar es 54 %.

Figura 2. **Relación entre las distribuciones brutas y percentiles**



Fuente: elaboración propia, con datos obtenidos de Brown (1999), *Principios de la medición en psicología y educación.*

- Deciles: los deciles dividen la distribución de calificaciones en 10 partes iguales, una de las divisiones posibles es en décimas, de hecho, son la empleada con mayor frecuencia. Estas dividen en diez segmentos la distribución, comenzando de 10 hasta 90 (10, 20, 30,90) cada etapa contiene el 10 % de las calificaciones.

- Calificación estándar: transformar calificaciones brutas a estándar, implica usar una escala con unidades de tamaño igual a las calificaciones originales. La calificación estándar (z) es la desviación que tiene una calificación bruta media:

$$z = \frac{x - \underline{x}}{s} \qquad \text{(Ec. 28)}$$

Debido a que la unidad básica de la escala es la desviación estándar, se dice que estas calificaciones son estándar. Las propiedades de las calificaciones estándar se expresan en una escala con media cero y desviación estándar 1. El valor absoluto de z es la distancia entre la calificación bruta y la media de la distribución, su signo identifica si la calificación está sobre (signo positivo) o debajo de la media (signo negativo), la transformación a estándar es lineal.

La distribución de la calificación será la misma con o sin transformación. Con el fin de no tener decimales y valores negativos las calificaciones z se transforman usando la ecuación:

$$z = A + \underline{B}z \qquad \text{(Ec. 29)}$$

z = calificación estándar transformada
A y B = constantes

Esta transformación es válida puesto que la suma o la multiplicación de o por una constante mantienen la proporción de la escala.

- Calificaciones estándar normalizadas: la transformación en este caso se hace utilizando una tabla de áreas de la curva normal (ver apéndice A).

Esta transformación se puede hacer debido a que en una distribución normal hay una relación especificable entre las áreas situadas a lo largo de la curva (proporción entre dos puntos) y las calificaciones estándar. Este método solo se puede usar si la distribución de la calificación se acerca a la normalidad.

- Para este cálculo se deben de tener las proporciones acumuladas (PA), en la tabla de zonas de la curva normal se encuentra la z comparable a esta PA, dependiendo si son por encima de la mediana (PA>0.500) se usa la columna de zona de proporción mayor y para las que son menores (PA<0.500) se usa la columna de zona de proporción menor, una vez obtenidos los valores de la tabla se utiliza la transformación siguiente, suponiendo que se utilizará una escala con media de cincuenta y desviación estándar de diez:

$$z' = 50 + 10z'z \qquad \text{(Ec. 30)}$$

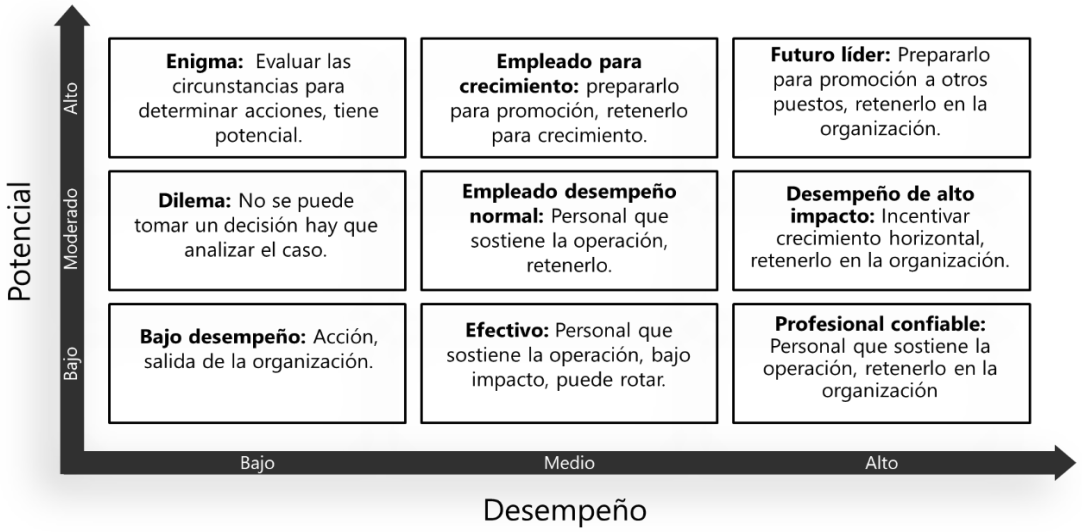
La calificación estándar y la estándar normalizadas son útiles cuando se necesita hacer algunas operaciones posteriores a la estandarización o se quieren comparar calificaciones entre varias pruebas o escalas y tiene la desventaja de no interpretarse fácilmente para las personas no familiarizadas con el concepto de desviación estándar.

2.2.5. El modelo de 9 cajas

Es un modelo utilizado en recursos humanos para la clasificación del talento de los empleados de una organización. Su origen es la matriz diseñada para uso en mercadotecnia del Boston *Consulting Group* en el cual se tienen

cuatro cuadrantes y dos ejes, en el eje horizontal se tiene la cuota de mercado de un producto y en el eje vertical su demanda, se dividen ambas en dos categorías baja y alta y de acuerdo con la posición en los cuatro cuadrantes divididos de esa forma se determina el potencial de un producto. Recursos humanos adaptó este concepto para clasificar al personal y determinar las categorías de este y las acciones a tomar acerca de su desarrollo y futuro dentro de las organizaciones. Recursos humanos usa 3 categorías en cada eje (bajo, medio y alto), por lo que existen 9 cuadrantes o cajas como se le conoce normalmente. En el eje de las X se coloca normalmente la calificación asignada a la evaluación del desempeño y en el eje de las Y el potencial de desarrollo de la persona, de esta forma se obtienen las siguientes categorías:

Figura 3. **Modelo de 9 cajas**



Fuente: elaboración diseño propia, realizado con PowerPoint

La definición de desempeño bajo, moderado y alto dependen del criterio de recursos humanos y de la dirección de la organización, no existe metodología para su categorización.

3. PRESENTACIÓN DE RESULTADOS

De acuerdo con los objetivos propuestos se presentan los siguientes resultados:

3.1. Objetivo General. Incrementar la fiabilidad del ejercicio de evaluación del desempeño con el uso de una metodología estadística, proponer una forma de evaluación de talento a través de referencias normativas para que el ejercicio de evaluación sea imparcial y así reforzar su credibilidad

Se presentan en los siguientes numerales los resultados parciales del objetivo general, resultados que hacen concluir que se logró mediante el uso de los estadísticos apropiados para cada objetivo específico, proporcionar resultados que mejoran la fiabilidad del ejercicio de evaluación de desempeño estudiado y en el último numeral se elaboró el baremo que permitió presentar mediante una ilustración descriptiva y una tabla, la correspondencia entre los deciles del baremo de calificaciones de evaluación de desempeño y las clasificaciones del talento de los colaboradores del sistema de nueve cajas.

3.2. Objetivo 1. Evaluar la fiabilidad y validez del instrumento aplicado para la evaluación del desempeño, utilizando el modelo alfa de Cronbach y distintos análisis de correlación

Para evaluar la fiabilidad y validez del instrumento aplicado para la evaluación del desempeño, se utilizó el modelo alfa de Cronbach para medir la consistencia de los aspectos evaluados por el formulario, para 17,919 casos que

fueron calificados por 1,123 evaluadores. Los aspectos evaluados a los empleados fueron:

- Conocimiento sobre el rol de la institución
- Resolución de problemas de trabajo
- Ejecutar su trabajo y rendición de cuentas
- Liderazgo
- Trabajo en equipo
- Comunicación e influencia
- Servicio
- Administración de talento
- Modelar conductas esperadas

Estos comportamientos no aplican a todos los empleados. Liderazgo, comunicación e influencia y administración de talento, solo pueden evaluarse a las personas que tienen personal a cargo. Debido a que dichos aspectos distorsionan los resultados, no fueron tomados en cuenta.

Las mediciones utilizadas fueron el Alfa de Cronbach y se tomaron los seis aspectos restantes. Cada uno de estos aspectos fue calificado de 1 a 5, siendo uno el mínimo y el máximo cinco. A continuación, se presenta el cálculo de la suma de estos ítems y sus varianzas poblacionales.

Tabla VI. **Resumen de cálculos base para medidas de confiabilidad y validez Alfa de Cronbach**

	Conocimiento sobre el rol de la institución	Resolución de problemas de trabajo	Ejecutar su trabajo y rendición de cuentas	Trabajo en equipo	Servicio	Modelar conductas esperadas
Varianza poblacional	0.275	0.234	0.217	0.204	0.227	0.221
Sumatoria de calificaciones	78531.620	78577.700	79200.980	79560.83	79717.690	79514.540

Fuente: elaboración propia.

3.2.1. Cálculo de Alfa de Cronbach (α)

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum s_t^2}{s_x^2}\right] \qquad \text{(Ec. 30)}$$

k = Es el total de preguntas de la prueba, en este caso de aspectos evaluados, son seis

s_t^2 = Es la varianza de los resultados de la prueba, suma de la varianza poblacional de los seis aspectos evaluados igual a 1.38

s_x^2 = es la varianza de las calificaciones de la prueba. La varianza es la suma total de los seis aspectos evaluados para cada individuo es igual a 6.76

$$\alpha = \frac{6}{6-1}\left[1 - \frac{1.38}{6.76}\right] = 0.955$$

La medida de homogeneidad es de 0.955

3.2.2. Cálculo de Spearman Brown

$$r_{11}=(2r_{oe})/(1+r_{oe}) \qquad \text{(Ec. 31)}$$

r_{oe} se calcula mediante el coeficiente de correlación de Pearson entre las mitades. Las mitades evaluadas fueron las preguntas pares e impares, se sumaron los resultados de la calificación de las preguntas para la población y entre ellas se calculó el coeficiente de correlación a través de Excel. Las sumas resultantes y los cálculos fueron los siguientes:

Tabla VII. **Resumen de cálculos base para medidas de confiabilidad y validez Spearman Brown**

Aspecto evaluado	Conocimientos sobre el rol de la institución	Ejecutar su trabajo y rendición de cuenta	Servicio
Σ *de calificaciones primera mitad*	78531.62	79200.98	79717.69
Σ *de calificaciones segunda mitad*	0.234	79560.83	79514.54

Fuente: elaboración propia.

Se tomó a las puntuaciones de los aspectos impares como la primera mitad y a la suma de las preguntas pares como la segunda mitad, de tal forma que el cálculo de Spearman Brown para la evaluación del desempeño quedó de la siguiente forma:

- Cálculo de coeficiente de correlación de Pearson en el que se usó Excel:

COEF.DE.CORREL=($vector\ de\ \Sigma\ de\ calificaciones\ primera\ mitad$; $vector\ de\ \Sigma\ de\ calificaciones\ segunda\ mitad$) = 0.812

- Cálculo de Spearman Brown:

$$r_{11} = \frac{2 * 0.812}{1 + 0.812} = 0.937$$

- Interpretación de resultados:

Tabla VIII. **Criterio George y Mallery para interpretación de alfa**

Coeficiente de alfa	Interpretación
$0 \leq r_{xy} < 0.49$	Inaceptable
$0.5 \leq r_{xy} < 0.59$	Pobre
$0.6 \leq r_{xy} < 0.69$	Cuestionable
$0.7 \leq r_{xy} < 0.79$	Aceptable
$0.8 \leq r_{xy} < 0.89$	Buena
$0.9 \leq r_{xy} < 0.95$	Excelente

Fuente: elaboración propia, con datos obtenidos de George y Mallery (2003). *SPSS for Windows step by step: a simple guide and reference*.

De acuerdo con la clasificación presentada la interpretación de la fiabilidad y validez del instrumento de evaluación es muy alta; es decir, es muy confiable por lo que podemos afirmar que las competencias evaluadas miden aspecto desempeño y que el instrumento usado para evaluar posee estabilidad para medir, por lo que arrojará resultados similares en diversas mediciones de un mismo individuo. Concluimos, por tanto, que el instrumento utilizado para evaluar el desempeño es fiable para situar a un trabajador en un baremo.

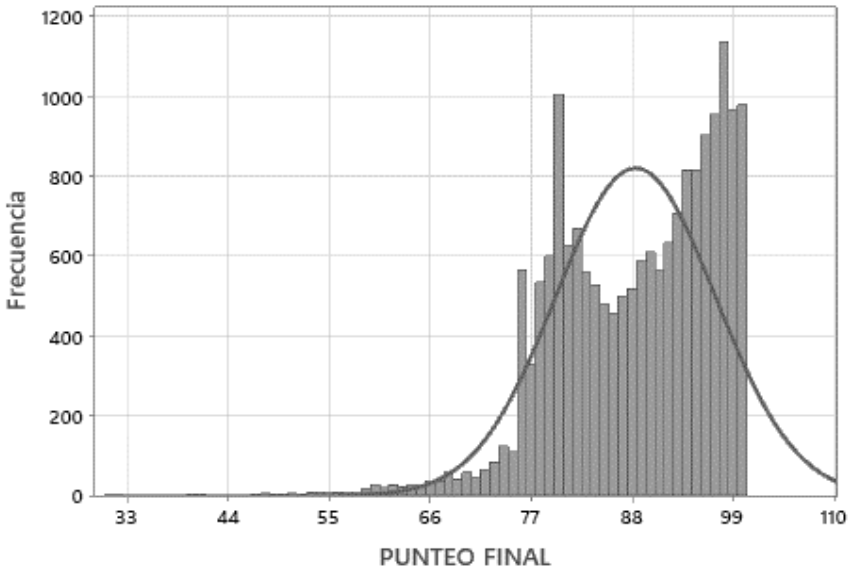
3.3. Objetivo 2. Detectar si existe parcialidad en la asignación de calificaciones de evaluación de parte de los evaluadores, si existen tendencias que afecten a los evaluados, a través análisis de varianza. Con el fin de recomendar formas de mejora de este punto

3.3.1. Análisis gráfico de tendencias

A fin de detectar si existe parcialidad en la asignación de calificación de evaluación de parte de los evaluadores y si existen tendencias extremas por medio de pruebas paramétricas y no paramétricas, se hace necesario el análisis gráfico de los resultados de la evaluación de la variable punteo o calificación obtenida por los evaluados, que presenta las siguientes características:

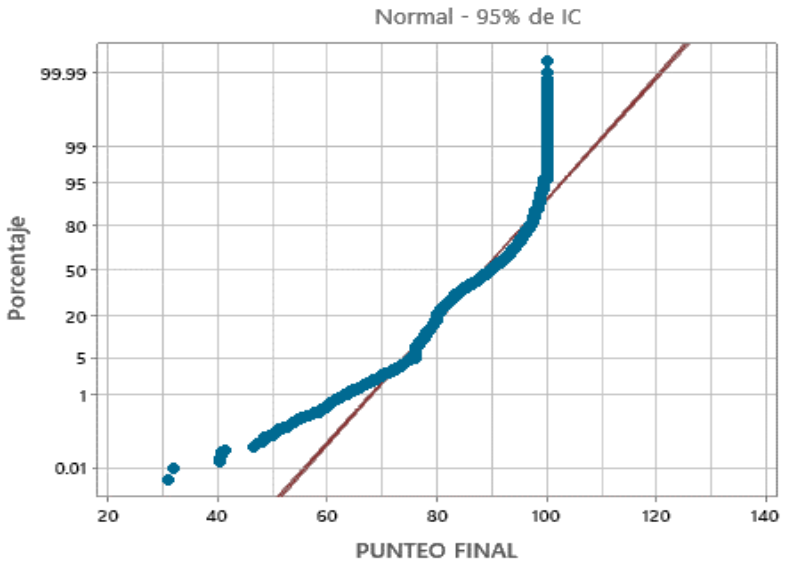
- N = número de evaluados = 17,918
- Media = 88.387
- Desviación estándar = 8.720
- Mínimo = 31
- Mediana = 89.600
- Máximo = 100
- Curtosis = 0.63

Figura 4. **Histograma con curva normal de las calificaciones**



Fuente: elaboración propia, realizado con IBM SPSS Statistics.

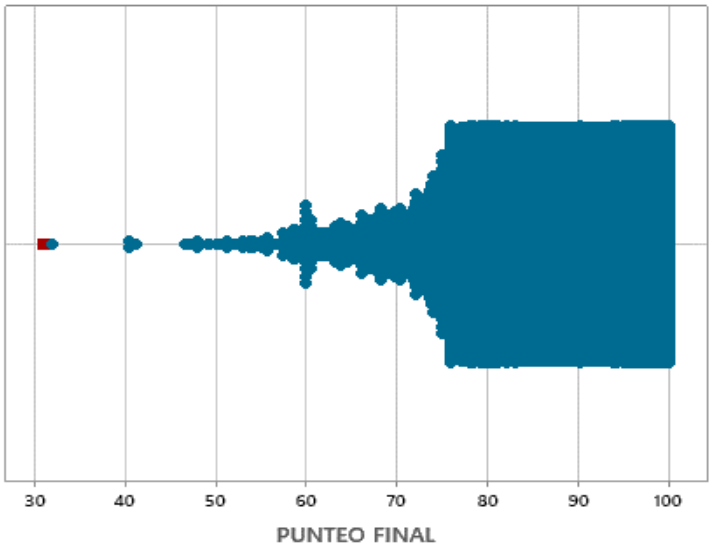
Figura 5. **QQ plot de calificaciones de desempeño**



Fuente: elaboración propia, realizado con IBM SPSS Statistics.

Como se puede observar en la figura 5, el supuesto de normalidad de la curva no se cumple y sus medidas centrales tienden hacia la derecha como podemos observar por simple inspección de la figura 5.

Figura 6. **Gráfica de valores atípicos de la calificación de la evaluación**



Fuente: elaboración propia, realizado con Minitab Statistics.

Tabla IX. **Prueba de normalidad**

Hipótesis nula	Todos los valores de los datos provienen de la misma población normal
• Hipótesis alterna	El valor más pequeño o grande de los datos es un valor atípico
• Nivel de significancia	α = 0.05

Fuente: elaboración propia.

Tabla X. **Prueba de Grubbs**

Variable	N	Media	Desv.Est.	Mín.	Máx.	p-valor
Calificación	17917	88.390	8.713	31.000	100.000	0.000

Fuente: elaboración propia.

Valor atípico de la variable punteo final, fila 10044 valor 31.

3.3.2. Cálculo de muestra para una población finita

Por el número de evaluadores y tamaño de la población, se decidió extraer una muestra, el cálculo de la muestra se hizo sobre el número de evaluadores, no sobre el número total de evaluados, debido al objetivo.

La población de estudio estuvo constituida por el total de trabajadores evaluados (17,919) y los evaluadores (1,123) del ejercicio de evaluación de desempeño realizado en diciembre del año 2019. La muestra obtenida fue probabilística de acuerdo con la fórmula de tamaño muestral para poblaciones finitas y estuvo conformada por 287 evaluadores y sus evaluados (4,735), como se estudia la influencia del evaluador sobre las calificaciones asignadas. La selección de la muestra se realizó por muestreo probabilístico aleatorio simple. A continuación, se presentan los cálculos:

$$n = \frac{N * Z_{\alpha p}^2 pq}{e^2(N-1) + Z_{\alpha}^2 pq} \qquad \text{(Ec. 32)}$$

En este caso los datos son los siguientes:

Tabla XI. **Variables para cálculo de tamaño de muestra**

Var.	Descripción	Dato
n =	Tamaño de muestra buscado	¿?
N =	Número de evaluadores	1123
Z =	Parámetro estadístico, para nivel de confianza 5 %	1.96
e =	Error de estimación máximo aceptado	0.05
p =	Probabilidad de que ocurra el evento estudiado	0.5
q =	(1-p)	0.5

Fuente: elaboración propia.

La probabilidad de que haya o no preferencia en la asignación de calificaciones de parte del evaluador es asumida como 0.5, dado que no contamos con precedentes estudiados. El error de estimación que se definió es del 5 % y el parámetro estadístico para el 5 % es de 1.96.

$$n = \frac{1123 * 1.96^2_{\alpha=.05}(0.5 * .05)}{0.05^2(1123-1)+1.96^2_{\alpha=.05}(.5 * .5)} = 286.43 \approx 287$$

Por el número de evaluadores (287) y de resultados de evaluación de evaluados (4735), el total de datos es de 4735.

Tabla XII. **Estadísticos de las calificaciones de la muestra**

N	**Válido**	**4735**
	Perdidos	**0**
Media		88,2882
Mediana		89.4000

Fuente: elaboración propia.

Por el tipo de distribución de las calificaciones asignadas por evaluador, que no es normal, se analizó la igualdad de medianas de Mood y por ser muy

cercana la media y la mediana se hizo también un ANOVA, la conclusión es la misma en ambos casos.

3.3.3. Prueba de igualdad de medianas y medias prueba de la mediana de Mood

- Prueba de Mood

    - Hipótesis nula: la mediana de las calificaciones asignadas por cada evaluador es la misma, por lo que no hay tendencias significativas hacía ser más estrictos o bondadosos al realizar la evaluación del desempeño.

    - Hipótesis alternativa: la mediana de las calificaciones asignadas por cada evaluador es diferente, por lo que existen tendencias significativas hacía ser más estrictos o bondadosos al realizar la evaluación del desempeño.

    - Resultado: grados de libertad 271, Chi cuadrada 2272.43, Valor-p 0.000, el Valor-p es cercano a cero, por lo que se rechaza la hipótesis nula, se concluye que existen evaluadores con tendencia a ser más estrictos o bondadosos que otros al asignar calificaciones.

- ANOVA

    - Hipótesis nula: la media de las calificaciones asignadas por cada evaluador es la misma, por lo que no hay tendencias significativas

hacía ser más estrictos o bondadosos al realizar la evaluación del desempeño.

$$H_0: \mu(E_1)=\mu(E_2)=\cdots\mu(E_{287})$$

- Hipótesis alternativa: la media de las calificaciones asignadas por cada evaluador es diferente, por lo que hay tendencias significativas hacia ser más estrictos o bondadosos al realizar la evaluación del desempeño.

$$H_1: \mu_i \neq \mu_j \text{ para algún } i \neq j$$

Tabla XIII. **ANOVA tendencias de evaluación**

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Entre grupos	182879.353	286	639438	18.108	.000
Dentro de grupos	157070.062	4448	35313		
Total	339949.415	4734			

Fuente: elaboración propia.

El nivel de significancia de la prueba de Mood y el ANOVA es de .0000 menor a 0.05 por lo que se rechaza la hipótesis nula y se concluyó que si existen evaluadores con tendencia a ser más estrictos y otros son más bondadosos al asignar calificaciones de evaluación de desempeño a los colaboradores. No se realiza prueba post hoc debido a que son 1,123 evaluadores y no se pueden

realizar acciones para cambiar a los evaluadores, pero sí hacer recomendaciones acerca de la forma de evaluar.

3.4. Objetivo 3. Proponer una forma de categorización del desempeño de los colaboradores de la organización, para que se tome una decisión adecuada respecto de su talento, por medio de la creación de baremos tomando como base las normativas basadas en la metodología de nueve cajas

En esta parte del estudio se expone la forma en que se categorizaron las calificaciones del desempeño de los colaboradores para tomar una decisión adecuada respecto de su talento. Lo anterior por medio de la creación de baremos, pues al contar con las normas para categorizar las calificaciones por grupos, se pudo hacer la equivalencia para asignarlas bajo una norma a las categorías definidas por la metodología de nueve cajas.

La metodología de nueve cajas clasifica a los trabajadores en tres grupos:

- Bajo
- Moderado
- Alto

Para estas categorías se crearon los baremos; la creación de los baremos se hizo de acuerdo con la estructura de la base de datos de la evaluación de desempeño para definir cuántos baremos son necesarios debido a las variables que influyen en los mismos. En este momento es importante aclarar que únicamente interesa la calificación de los evaluados.

Tabla XIV. **Variables de la base de datos**

Variable	Descripción	Clase
Código del evaluador	Identificador del evaluador	Identificador
Nivel del evaluador	Nivel organizacional del evaluador(*)	Sociodemográfica (variable independente)
Código del evaluado	Identificación del evaluado	Identificador
Nivel del evaluado	Nivel organizacional del evaluado(*)	Sociodemográfica (variable independente)
Dependencia	Área organizacional a la que pertenecen el evaluador y el evaluado (**)	Sociodemográfica (variable independente)
Conocimiento sobre el rol de la institución	Factor evaluado 1	Escala de likert
Resolución de problemas de trabajo	Factor evaluado 2	Escala de likert
Ejecutar su trabajo y rendición de cuentas	Factor evaluado 3	Escala de likert
Liderazgo	Factor evaluado 4	Escala de likert
Trabajo en equipo	Factor evaluado 5	Escala de likert
Comunicación e influencia	Factor evaluado 6	Escala de likert
Servicio	Factor evaluado 7	Escala de likert
Administración de talento	Factor evaluado 8	Escala de likert
Modelar conductas esperadas	Factor evaluado 9	Escala de likert
Calificación	Punteo final de acuerdo con los 9 factores evaluados	De medición de 1 a 100

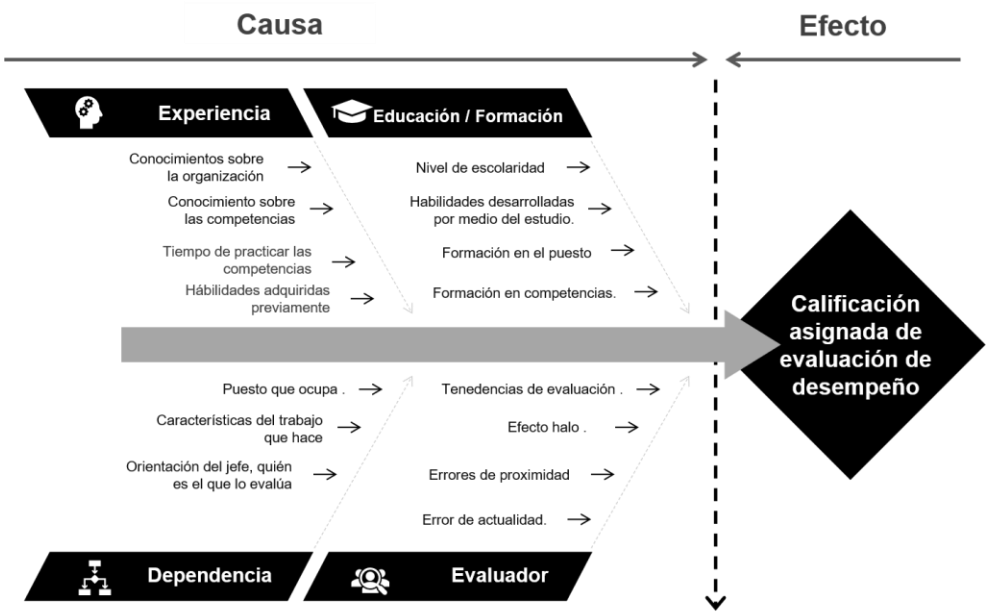
Fuente: Elaboración propia.

3.4.1. Determinación de variables sociodemográficas que inciden en la calificación

Para hacer la categorización el primer paso fue determinar las variables sociodemográficas que podían afectar los resultados de la evaluación. El contenido de estas variables dependía del evaluado y el evaluador, por lo que se

hizo un análisis y se utilizó la técnica de espina de pescado para identificar las razones de esta influencia y luego una validación de las que estaban más representadas en las variables sociodemográficas, así como la practicidad de evaluarlas para la confección del baremo.

Figura 7. **Análisis de variables que influyen sobre las calificaciones**

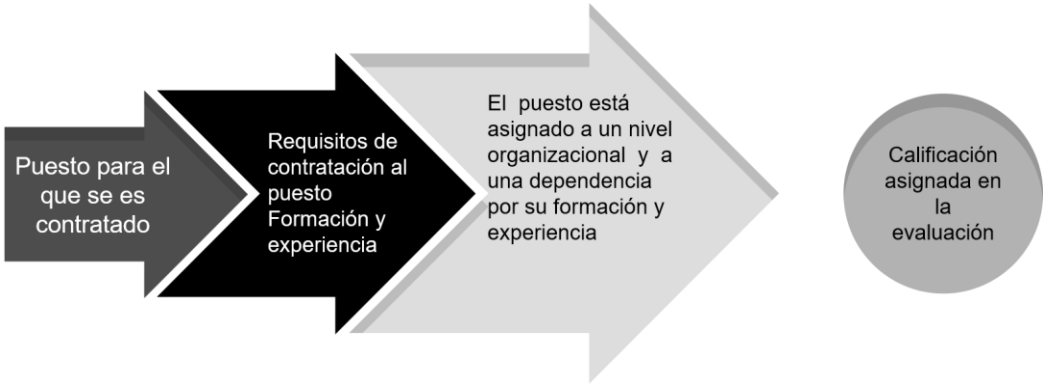


Fuente: elaboración propia, realizada con PowerPoint.

Resultado del análisis: las condiciones del evaluado, educación y experiencia son las que tienen mayor incidencia en su calificación. Estas dependen de su entendimiento de las competencias evaluadas, el conocimiento sobre la organización, el tiempo en el puesto y los entrenamientos tomados respecto a la evaluación los que probablemente tengan mayor impacto sobre la calificación que le asigne su evaluador. El nivel organizacional al que está asignado su puesto es debido a su nivel educacional y experiencia, por lo que se

puede utilizar este elemento como clasificador de incidencia sobre el puesto asignado. A continuación, se presenta un diagrama de análisis que resume e ilustra la deducción hecha sobre el análisis de espina de pescado:

Figura 8. **Resumen del análisis de variables**



Fuente: elaboración propia, realizado con PowerPoint.

Tabla XV. **Niveles organizacionales**

Nivel organizacional	Código
Operativo	1
Especializado Operativo	2
Especializado	3
Técnico	4
Técnico Profesional	5
Profesional	6
Jefatura	7
Asistencia Profesional	8
Ejecutivo	9
Directivo	10

Fuente: elaboración propia.

Las variables sociodemográficas pueden afectar los resultados de la evaluación, puede ser tanto el nivel del evaluado, como la dependencia a que

pertenece y el nivel del evaluador. Por ser la dependencia a que pertenece una variable que puede identificar a la institución, esta fue codificada con numeración de 1 a 280. La tabla de niveles organizacionales se comparte debido a que está asociada a la mayoría de las instituciones gubernamentales.

Se evaluó con base en el análisis de espina de pescado la incidencia de la dependencia sobre la calificación final, al revisar vemos que la dependencia está relacionada con el puesto y el puesto está relacionado con el nivel organizacional, razón por la cual no se evalúa, ya que se duplicaría la incidencia de los factores de origen ya razonados en el párrafo anterior. Nos queda al final el evaluador, esté como se ve en el análisis anterior incide sobre la calificación; sin embargo, no es práctico hacer 1,123 baremos y la incidencia de esta variable se piensa trabajar de otra forma, por lo que quedó únicamente evaluar si incide el nivel organizacional sobre el puntaje asignado a los evaluados. Para ello se calculó una muestra representativa de la población y se procedió a hacer la evaluación por medio de un ANOVA.

3.4.2. Cálculo de muestra para una población finita

Para iniciar el análisis se decidió nuevamente extraer una muestra de la población estudiada que estuvo constituida por el total de trabajadores evaluados (17,918). No se utilizó un registro por no estar registrada su categoría. La selección de la muestra se realizó por muestreo probabilístico aleatorio simple. A continuación, se presentan los cálculos:

$$n = \frac{N * Z_{\alpha p}^2 pq}{e^2(N-1) + Z_{\alpha}^2 pq} \qquad \text{(Ec. 33)}$$

En este caso los datos son los siguientes:

n =	Tamaño de muestra buscado	=	¿?
N =	Número de evaluados	=	17,918
Z =	Parámetro estadístico, para nivel de confianza 5%	=	1.96
e =	Error de estimación máximo aceptado	=	0.05
p =	Probabilidad de que ocurra el evento estudiado	=	0.5
q =	(1-p)	=	0.5

La probabilidad de que incida o no el nivel organizacional sobre la asignación de calificaciones a los evaluadores, se asume como 0.5, dado que no contamos con precedentes estudiados. El error de estimación que se definió es del 5 % y el parámetro estadístico para el 5 % es de 1.96.

$$n = \frac{17{,}918 * 1.96^2_{\alpha=.05}(0.5 * .05)}{0.05^2(17918-1)+1.96^2_{\alpha=.05}(.5 * .5)} = 376.11 \approx 377$$

Al verificar el número se decidió seguir analizando con la muestra por evaluadores que tiene un total de 4735 registros y que fue extraída aleatoriamente.

3.4.3. Factores que influyen en las calificaciones

Como primer paso para la elaboración del baremo se determinó si existe dependencia entre las calificaciones asignadas a los trabajadores y los niveles a que sus puestos pertenecen. Es la única variable evaluada debido a que el nivel del colaborador es asignado acorde al puesto y para su contratación en la institución.

- ¿Existen diferencias significativas de calificaciones debidas al nivel organizacional de los evaluados? Por la distribución de la muestra se hace

la prueba de Mood y por la cercanía de la media y la mediana es válido hacer un ANOVA.

Tabla XVI. **Prueba de Mood nivel organizacionales vs. Punteo final**

Estadísticas descriptivas

Nivel evaluado	Mediana	Mediana general de N <=	Mediana general de N >	Q3 – Q1	IC de la mediana de 95 %
1	82.0	213	79	12.80	(80.6; 83.8449)
2	81.6	104	42	13.35	(80.8; 83.2)
3	86.6	446	317	14.80	(85; 87.6130)
4	89.0	1244	1116	13.20	(88.4; 89.4)
5	93.0	83	136	14.00	(91.3062; 94.2938)
6	93.0	278	409	13.60	(91.8660; 93.7340)
7	97.6	21	198	4.60	(96.7062; 98.0938)
8	96.4	10	34	5.60	(95.2268; 97.9911)
General	89.4				

Fuente: elaboración propia.

Prueba

H_0: las medianas de población son todas iguales

H_1: las medianas de población no son todas iguales

Chi-cuadrada	Valor-p
309.61	0.000

La hipótesis planteada utilizando el ANOVA para responder a esta pregunta es:

- Hipótesis nula: la media de las calificaciones asignadas en la evaluación de desempeño en la muestra es la misma por cada nivel.

$$H_0\colon\ \mu_{N_1} = \mu_{N_2} = \cdots \mu_{N_9}$$

- Hipótesis alternativa: existe alguna o varias diferencias entre las medias de las calificaciones asignadas por nivel organizacional en la evaluación de desempeño.

$$H_1\colon\ \mu_i \neq \mu_j\ para\ algún\ i \neq j$$

Tabla XVII. **ANOVA influencia nivel organizacional sobre calificaciones**

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Entre grupos	28086.956	7	4012.422	60.8150	<.001
Dentro de grupos	311544.525	4722	65.977		
Total	339631.482	4729			

Fuente: elaboración propia.

Con este análisis de varianza se determinó la necesidad de hacer tablas independientes para los niveles de los evaluados, dado el resultado del p- valor de la prueba, este nos indica que si hay diferencias significativas entre las calificaciones de los grupos de nivel organizacional. Se utilizó la prueba de Bonferroni, cuya tabla se presenta a continuación y se determinó para cuáles niveles es necesario hacer normativas separadas de calificación. Depende del nivel de significancia en su comparación y la diferencia de medias.

3.4.4. Bonferroni por nivel organizacional

Se puede observar en la tabla de comparaciones múltiples de Bonferroni que la diferencia entre medias de los grupos de nivel organizacional 1 es mínimo 2 y su significancia es mayor a .05 por lo que para el nivel organizacional 1 y 2 se utilizará una misma tabla. Lo mismo sucede con los grupos 3 y 4, 5 y 6, 7 y 8.

Tabla XVIII. **ANOVA influencia nivel organizacional sobre calificaciones**

(I) Código nivel evaluado		Diferencia de medias (I-J)	Error estándar	Sig.	Intervalo de confianza al 95%	
					Límite inferior	Límite superior
1	2	-0.4904109589041070	0.823	1.000	-3.064	2.0829
	3	-3.596716278568621*	0.559	0.000	-5.344	-1.8497
	4	-4.491408172742041*	0.504	0.000	-6.066	-2.9165
	5	-7.027853881278517*	0.726	0.000	-9.297	-4.7584
	6	-6.794474686446875*	0.567	0.000	-8.568	-5.0209
	7	-12.243378995433787*	0.726	0.000	-14.513	-9.9740
	8	-10.371046077210480*	1.314	0.000	-14.477	-6.2655
2	1	0.4904109589041070	0.823	1.000	-2.083	3.0637
	3	-3.106305319664514*	0.734	0.001	-5.400	-0.8130
	4	-4.000997213837934*	0.693	0.000	-6.166	-1.8359
	5	-6.537442922374410*	0.868	0.000	-9.250	-3.8250
	6	-6.304063727542768*	0.740	0.000	-8.618	-3.9905
	7	-11.752968036529680*	0.868	0.000	-14.465	-9.0405
	8	-9.880635118306373*	1.397	0.000	-14.247	-5.5145
3	1	3.596716278568622*	0.559	0.000	1.850	5.3437
	2	3.106305319664515*	0.734	0.001	0.813	5.3996
	4	-0.8946918941734200	0.338	0.230	-1.952	0.1626
	5	-3.431137602709896*	0.623	0.000	-5.377	-1.4849
	6	-3.197758407878254*	0.427	0.000	-4.533	-1.8625
	7	-8.646662716865166*	0.623	0.000	-10.593	-6.7005
	8	-6.774329798641858*	1.259	0.000	-10.710	-2.8382
4	1	4.491408172742041*	0.504	0.000	2.916	6.0663
	2	4.000997213837934*	0.693	0.000	1.836	6.1661
	3	0.8946918941734200	0.338	0.230	-0.163	1.9520
	5	-2.536445708536476*	0.574	0.000	-4.330	-0.7431
	6	-2.303066513704834*	0.352	0.000	-3.404	-1.2025
	7	-7.751970822691746*	0.574	0.000	-9.545	-5.9586
	8	-5.879637904468439*	1.236	0.000	-9.742	-2.0168

Continuación de la tabla XVIII.

(I) Código nivel evaluado		Diferencia de medias (I-J)	Error estándar	Sig.	Intervalo de confianza al 95%	
					Límite inferior	Límite superior
5	1	7.027853881278517*	0.726	0.000	4.758	9.2973
	2	6.537442922374410*	0.868	0.000	3.825	9.2499
	3	3.431137602709896*	0.623	0.000	1.485	5.3773
	4	2.536445708536476*	0.574	0.000	0.743	4.3298
	6	0.2333791948316420	0.630	1.000	-1.737	2.2035
	7	-5.215525114155270*	0.776	0.000	-7.642	-2.7894
	8	-3.3431921959319600	1.342	0.357	-7.537	0.8510
6	1	6.794474686446875*	0.567	0.000	5.021	8.5680
	2	6.304063727542768*	0.740	0.000	3.990	8.6177
	3	3.197758407878254*	0.427	0.000	1.863	4.5330
	4	2.303066513704835*	0.352	0.000	1.202	3.4036
	5	-0.2333791948316420	0.630	1.000	-2.203	1.7367
	7	-5.448904308986911*	0.630	0.000	-7.419	-3.4788
	8	-3.5765713907636000	1.263	0.130	-7.525	0.3714
7	1	12.243378995433787*	0.726	0.000	9.974	14.5128
	2	11.752968036529680*	0.868	0.000	9.040	14.4654
	3	8.646662716865166*	0.623	0.000	6.700	10.5929
	4	7.751970822691746*	0.574	0.000	5.959	9.5453
	5	5.215525114155270*	0.776	0.000	2.789	7.6416
	6	5.448904308986911*	0.630	0.000	3.479	7.4190
	8	1.8723329182233100	1.342	1.000	-2.322	6.0665
8	1	10.371046077210480*	1.314	0.000	6.266	14.4766
	2	9.880635118306373*	1.397	0.000	5.515	14.2467
	3	6.774329798641858*	1.259	0.000	2.838	10.7104
	4	5.879637904468439*	1.236	0.000	2.017	9.7424
	5	3.3431921959319600	1.342	0.357	-0.851	7.5374
	6	3.5765713907636000	1.263	0.130	-0.371	7.5245
	7	-1.8723329182233100	1.342	1.000	-6.067	2.3219

Fuente: elaboración propia.

3.4.5. Baremo de evaluación de desempeño

Acorde a lo definido en el numeral anterior, se calcularon los percentiles de las calificaciones de la evaluación de los cuatro niveles organizacionales y se diseñó el baremo de acuerdo a los percentiles 1, 10, 20, 30, 40, 50, 60, 70, 80, 90 y 99, bajo las reglas usuales de baremación.

Tabla XIX. **Deciles de calificaciones por niveles organizacionales**

		Nivel 1 y 2	**Nivel 3 y 4**	**Nivel 4 y 5**	**Nivel 6 y 7**
	N =	438	3123	906	263
Percentiles	1	61.478	65.048	66.800	76.560
	10	74.760	77.800	79.800	86.280
	20	76.600	80.000	82.200	92.800
	30	78.600	82.600	85.400	95.400
	40	80.000	85.400	89.600	96.400
	50	81.900	88.400	93.000	97.400
	60	85.000	91.200	94.840	98.200
	70	88.600	93.600	96.400	98.760
	80	92.640	95.840	97.800	99.400
	90	96.22	98.000	99.000	100.00
	99	100.00	100.000	100.000	100.000

Fuente: elaboración propia.

Tabla XX. **Baremo de calificaciones con equivalencias al sistema de nueve cajas de la evaluación de desempeño**

Percentil	Interpretación Sistema 9 cajas	Nivel Organizacional			
		1 y 2	3 y 4	5 y 6	7 y 8
1	Inferior	61	65	67	77
10	Bajo, bajo	62-74	66-78	68-80	78-86
20	Bajo	75-77	79-80	81-82	87-93
30	Bajo alto	78-79	81-83	83-85	94-95
40	Promedio bajo	80-81	84-85	86-90	96
50	Promedio	82	86-88	91-93	97
60	Promedio alto	83-85	91	94-95	98
70	Alto bajo	86-88	92-94	96	-
80	Alto	89-93	95-96	97-98	99
90	Alto, alto	94-96	97-98	99	-
99	Superior	97-100	99-100	100	100

Fuente: elaboración propia.

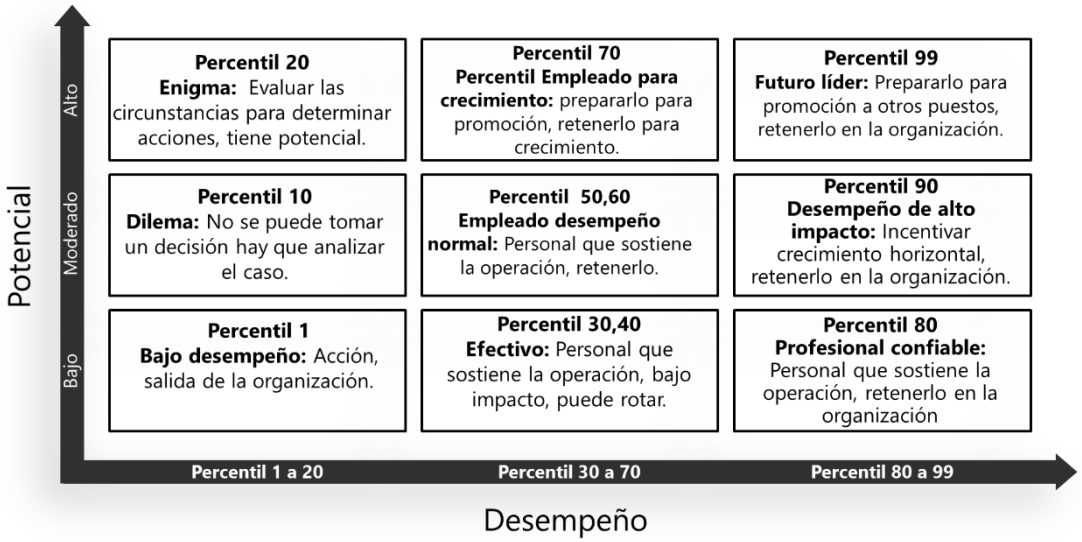
Para hacer la asignación de percentiles a las tres categorías de la metodología de 9 cajas, se propone hacerlo de acuerdo con la interpretación siguiendo la siguiente regla:

Tabla XXI. **Asignación de percentiles de calificaciones a metodología de 9 cajas**

Percentil	Interpretación	Desempeño
1-20	Bajo a inferior	bajo
30-70	Bajo alto, promedio y Alto bajo	moderado
80-99	Alto a superior	alto

Fuente: elaboración propia.

Figura 9. **Sistema de 9 cajas con percentil de desempeño**



Fuente: elaboración propia, realizado con PowerPoint.

4. DISCUSIÓN DE RESULTADOS

4.1. Análisis interno

Luego de presentar los resultados del análisis del ejercicio de evaluación de desempeño se confirmó que el instrumento de evaluación es fiable y válido, el Alfa de Cronbach (95.5 %) y *Spearman Brown* (93.7 %), no se estudió la validez de contenido debido a que el instrumento de evaluación fue diseñado por un experto reconocido a nivel nacional. Se pretendía en este estudio determinar las tendencias de evaluación por cada evaluador; sin embargo, al discutir con expertos en el tema, esto no aporta al estudio, la importancia del análisis realizado es que se puede, mediante la identificación de diferencia de medias o medianas de las calificaciones asignadas por evaluador, definir la necesidad de alinear conceptos de evaluación de los evaluadores y evidencia su importancia.

En este análisis, el hecho de que la población y la muestra de calificaciones no tenga una distribución normal, hizo que el planteamiento inicial de utilizar un ANOVA no fuera suficiente, en el manual de MINITAB se expone que es posible utilizar pruebas paramétricas para poblaciones con distribuciones no normales, por lo que se utilizó la prueba no paramétrica de Mood y el ANOVA, además varios autores utilizan ANOVA cuando la media y la mediana tienen poca diferencia como es el caso de la población estudiada. Finalmente el encontrar que efectivamente la variable de nivel organizacional afecta las calificaciones asignadas a los evaluados al utilizar Bonferroni, donde es notorio que el nivel organizacional es directamente proporcional a la calificación asignada en la evaluación; a mayor nivel más altas son las calificaciones, a menor nivel más bajas son las calificaciones, hace mucho sentido dentro de la organización

estudiada, donde estos niveles son representativos de la formación y experiencia de sus empleados y hace suponer que este resultado es aplicable a la mayoría de organizaciones.

4.2. Análisis externo

Este análisis tiene como objetivo el conferir mayor imparcialidad y validez al ejercicio de evaluación de desempeño y recomendar acciones de mejora que las promuevan. El resultado del coeficiente Alfa de Cronbach (0.95) y Spearman Brown (0.93) con que se evaluó la fiabilidad y validez del instrumento, lo definen válido y confiable de acuerdo con Frías Navarro, (2019) quienes exponen como aceptables los valores mayores que 0.7. Abarzúa (2017) plantea que el diseño de un baremo permite conferir imparcialidad a una prueba y que una diferencia de resultados entre distintos grupos puede ser reflejo de las inequidades que viven los propios grupos. Esto coincide con el resultado obtenido al distinguir los grupos de nivel organizacional. No todos en la organización tienen las mismas características debido a la naturaleza, requerimientos y compensaciones de un puesto. Por ello sistematizar estadísticamente este análisis ha sido acertado y propone mejoras para ejecutar un ejercicio de evaluación que brinde confiabilidad al asignar apropiadamente la categorización del desempeño de los empleados al eje horizontal del sistema de nueve cajas.

CONCLUSIONES

1. La interpretación de los resultados de los cálculos de las mediciones de fiabilidad (95.5 %) y validez del instrumento (93.7 %) indican que es válido y fiable.

2. El análisis gráfico permite observar que hay tendencias de evaluación distintas. El ANOVA confirma la hipótesis sobre que estas diferencias existen.

3. El nivel organizacional influye significativamente sobre las calificaciones de evaluación del desempeño. La prueba de Bonferroni confirmó que hay que separar las normas por niveles organizacionales. Se concluyó que a menor nivel organizacional las calificaciones son más bajas y a mayor nivel organizacional las calificaciones son más altas.

4. Se puede afirmar que el nivel de confiabilidad y credibilidad de la evaluación del desempeño se incrementan al aplicar la metodología desarrollada. Por la certeza que proporcionan los resultados de las pruebas estadísticas aplicadas, lo que puede mejorar la percepción de mayor justicia en el ejercicio.

RECOMENDACIONES

1. Diseñar un instrumento para evaluar el desempeño es altamente recomendable verificar estadísticamente su fiabilidad y validez.

2. Formar a los evaluadores sobre la forma de evaluar, es altamente recomendable por sus tendencias extremas al evaluar.

3. Detectar la existencia de variables socio demográficas que afectan las calificaciones, es necesario, para poder normar la interpretación y que esta sea acertada y justa.

4. Sustentar la objetividad de la evaluación y tomar decisiones acertadas sobre las carreras de los trabajadores es recomendable aplicar la metodología que se presenta en este estudio.

REFERENCIAS

1. Aiken, L. R. (2003). *Test psicológicos y evaluación.* Pearson Educación.

2. Alfaro, K. y Montero, E. (2013). Aplicación del modelo de *Rasch*, en el análisis psicométrico de una prueba de diagnóstico en matemática. Revista Digital: *Matemática, Educación e Internet*. 13. 10.18845/rdmei. v13i1.1628.

3. Alveiro Montoya, C. (2009). Evaluación del desempeño como herramienta para el análisis del capital humano. *Revista Científica Visión de Futuro*, 11(1). Recuperado de https://www.redalyc.org/articulo.oa?id=3579/357935472005ernández.

4. Abarzúa Morasso, A. (2017). Confiabilidad, validez e imparcialidad en evaluación educativa. Recuperado de https://www.inee.edu.mx/wp-content/uploads/2019/08/P2A352.pdf.

5. Brown, F. (1999). Principios de la medición en psicología y educación. Manual Moderno.

6. Cerdas, D. y Montero, E. (2017). Uso del modelo *Rasch* para la construcción de tablas de especificaciones: Propuesta metodológica aplicada a una prueba de selección universitaria. *Actualidades Investigativas en educación*, 1-16 DOI. Recuperado de https://dx.doi.org/10.15517/aie.v17i1.27299.

7. Fox, D., y López, E. (1981). *El proceso de investigación en la educación*. Universidad de Navarra.

8. Frías Navarro, D. (2019). *Apuntes de consistencia interna de las puntuaciones de un instrumento de medida*. Universidad de Valencia. España.

9. George, D., y Mallery, P. (2003). SPSS for Windows step by step: A Simple Guide and Reference. 11.0 Update (4ª ed.). Boston, EE.UU.: Allyn & Bacon.

10. Gamboa, J. y Heredia, C. (2017). *Propuesta de implementación de una plataforma informática para mejorar los procesos del plan de sucesión de talentos de la empresa NETAFIM* (Tesis de maestría). Universidad Privada del Norte, Perú. Recuperado de https://repositorio.upn.edu.pe/handle/11537/11861.

11. García Leal, J. y Lara Porras, A. M. (1998). Diseño Estadístico de Experimentos. Análisis de la Varianza. Grupo Editorial Universitario.

12. Gorriti Bontigui, M. (2007). La evaluación del desempeño en las Administraciones Públicas de España. La Evaluación del Desempeño: Análisis, retos y propuestas. Una aplicación a la Comunidad Autónoma de Aragón. *Revista Aragonesa de Administración Pública*, ISSN 1133-4797, 13, 297-387. Recuperado de https://dialnet.unirioja.es/servlet/articulo?codigo=5547442.

13. Gutiérrez Castillo, J. J., Cabero Almenara, J., y Estrada Vidal, J. (2017). Diseño y validación de un instrumento de evaluación de la competencia digital del estudiante universitario. *Espacios*, 38(10). Recuperado de https://idus.us.es/handle/11441/54725.

14. Hidrugo, J., Pucce, D. (2016). *El rendimiento y su relación con el desempeño laboral del talento humano en la Clínica San Juan de Dios – Pimentel* (Tesis de Licenciatura). Universidad de Sipan, Chiclayo, Perú. Recuperado de http://repositorio.uss.edu.pe/bitstream/handle/uss/2285/Tesis%20de%20Hidrugo%20V%E1squez%20y%20Pucce%20Castillo.pdf;jsessionid=C9BC0A44EAF02B0AFA917952F8D4031B?sequence=.

15. HSU, J. (1996). *Multiple Comparisons Theory and Methods.* Chapman and Hall/CRC.

16. Kaufman, R. (2009). Meta Thinking and Planning: An introduction to Defining and Delivering Individual and Organizational Success*. Performance Improvement Quarterly*, 22(2), 5-15.

17. Leiva, D. (2016). *Sesgo de Escalada del compromiso en la evaluación del desempeño* (Tesis de maestría). Universidad de Chile, Santiago. Recuperado de http://repositorio.uchile.cl/bitstream/handle/2250/138625/Sesgo%20de%20escalada%20del%20compromiso%20en%20la%20evaluaci%C3%B3n%20del%20desempe%C3%B1o.pdf?sequence=1yisAllowed=y#:~:text=su%20error%20p%C3%BAblicamente.-,El%20sesgo%20de%20Escalada%20del%20Compromiso%2C%20puede%20estar%20presente%20en,que%20debe%20evaluar%2

0su%20desempe%C3%B1o.ytext=La%20variable%20independiente%20del%20experimento,con%20un%20empleado%20promovido%20previamente.

18. Lloret-Segura, S., Ferreres-Traver, A., Hernández-Baeza, A. y Tomás-Marco, I. (2014). El Análisis factorial exploratorio de los ítems: Una guía práctica, revisada y actualizada. *Anales de Psicología*, 30(3), Recuperado de https://dx.doi.org/10.6018/analesps.30.3.199361.

19. Macías Calvillo, E. (2011). *Validación y confiabilidad de pruebas de opción múltiple para la evaluación de habilidades* (Tesis de Maestría). Centro de Investigación en Matemáticas (CIMAT), Guanajato, México. Recuperado de https://cimat.repositorioinstitucional.mx/jspui/bitstream/1008/245/2/TE%20373.pdf.

20. Martínez López, E. J. (2004). Elaboración de baremos de calificación en Educación Física con la hoja de cálculo Excel 2000. *Revista efedeportes.com*, 10(69). Recuperado de https://www.efdeportes.com/efd69/baremos.htm.

21. Pérez, R., García, J., Gil, J. y Galán, A. (2009). *Estadística aplicada a la educación*. Pearson educación, S. A.

22. Pulido, H. y Salazar, R. (2008). *Análisis y diseño de experimentos*. Segunda edición. McGraw Hill / Interamericana Editores, S. A. de C. V.

23. Rojas García, G. V. (2007). *Muestreo para Correlaciones por Contingencias y de Pearson* (Tesis de Maestría).Universidad Central “Marta

Abreu" de las Villas, Santa Clara, Cuba. Recuperado de https://dspace.uclv.edu.cu/handle/123456789/10910.

24. Rossie-Casé, L., Lopetegui, S., Doná, S., Biganzoli, B. y Garzaniti, R. (2014). Matrices Progresivas de Raven; Efecto *Flinn* y actualización de baremos. *Revista de Psicología,* 2(23). Recuperado de https://www.redalyc.org/pdf/264/26435341002.pdf.

25. Sullivan, J. (2011). Performance Appraisal, the Most Dreaded HR Process – A List of the Top 50 Problems. Recuperado de https://drjohnsullivan.com/articles/performance-appraisal-the-most-dreaded-hr-process-a-list-of-the-top-50-problems/.

26. Walpole, R., Nyers, R., Myers, S. y Ye, K. (2012). *Probabilidad y estadística para ingeniería y ciencias*. Pearson Edición.