

REVISTA INTERAMERICANA DE PSICOLOGÍA OCUPACIONAL

VOLÚMEN 23, NÚMERO 1, 2004, ISSN 120-3800

DIRECCIÓN

Fernando Toro Alvarez

ADMINISTRACIÓN

Teresa Chávez de Toro

COMITÉ EDITORIAL ASESOR

Ramiro Alberto Alvarez Cuadros, Ph. D.

COLOMBIA

Carlos Dávila Ladrón De Guevara, Ph. D.

COLOMBIA

Eduardo Ponce De León, M. A.

COLOMBIA

Néstor Vargas Castro, M. A.

COLOMBIA

Fernando Arias Galicia, Ph. D.

MÉXICO

Carlos Andújar Rojas, Ph. D.

PUERTO RICO

Julio César Casales F., Ph. D.

CUBA

DIRECCIÓN POSTAL

Calle 34 N° 85C-20

Apartado Aéreo 65021 - Medellín

e-mail: cincel@epm.net.co

INFORMACIÓN LEGAL

La Revista Interamericana de Psicología Ocupacional ha sido registrada ante el Ministerio de Gobierno, oficina de Registro de Propiedad Intelectual y Comunicaciones, y se le ha eximido de caución mediante Resolución 1004 de mayo 2 de 1985.

Todos los derechos de reproducción están reservados. Ninguna parte de esta publicación puede ser reproducida, archivada o transmitida en ninguna forma o por ningún medio electrónico, mecánico de fotocopias, grabaciones o cualquier otro medio, sin la autorización previa y por escrito de la Revista Interamericana de Psicología Ocupacional.

Cada autor es responsable de los conceptos expresados en su escrito.

La Revista Interamericana de Psicología Ocupacional está inscrita en el servicio de informática del Colegio Oficial de Psicólogos de España.

CONFIABILIDAD EN PROCESOS DE EVALUACIÓN DE 360 GRADOS¹

Fernando Toro Álvarez
*Centro de Investigación e Interventoría
en Comportamiento Organizacional*

Resumen

Esta investigación tuvo por objeto examinar la existencia de consenso y la variabilidad en la confiabilidad de instrumentos de evaluación en los que participan diversos calificadores que califican a un mismo evaluado. Este tema tiene relevancia por cuanto se ha vuelto una práctica empresarial frecuente la evaluación de 360 grados. Se realizaron dos estudios distintos. Uno examinó estos asuntos a partir de un método e instrumento para evaluación de competencias organizacionales en un grupo de empleados (102) que se autoevaluaron, fueron evaluados por algunos de sus pares (196) y por sus jefes (60) en una empresa estatal. El otro estudio examinó los mismos asuntos a partir de una evaluación del liderazgo del personal de mando de una empresa del sector industrial. En este caso los calificadores fueron 217 empleados que evaluaron a sus propios jefes (63). Estos jefes se autoevaluaron y al tiempo fueron evaluados por su respectivo jefe y por 102 pares o colegas. Tal como se reportó en otros estudios, se registraron diferencias importantes en la confiabilidad de las medidas dependiendo del tipo de calificador. También se encontraron diferencias estadísticamente significativas entre las calificaciones dadas por cada tipo de evaluador. Se encontró que las calificaciones más confiables y consistentes fueron las del jefe y los colaboradores. Fueron menos confiables y consistentes la auto evaluación y la calificación de los pares.

Abstract

The main purpose of this study was to explore the existence of agreement and reliability of measurements provided by methods and instruments in which several different raters participate. This process is called 360 degrees assessment. Two different studies were done. In the first case an instrument and method designed to assess organizational competences was used. 102 employees evaluated themselves, were evaluated by 196 colleagues and 60 supervisors. In the second case 217 employees evaluated 63 supervisors. Those bosses also evaluated themselves and were evaluated by 102 peers. As reported in previous research, there were found important differences in reliability related to type of rater. It also were found statistically significant differences between raters. The more reliable and consistent were supervisors and subordinates and the less were self evaluations and peers'.

Palabras Claves: Evaluación, confiabilidad, correlación, auto evaluación.

Key Words: Assessment, reliability, correlation, self evaluation.

¹ Esta investigación se realizó con el auspicio y financiación del Centro de Investigación e Interventoría en Comportamiento Organizacional, Abril de 2003.

La mayoría de los programas de retroalimentación de 360 grados están diseñados para el mejoramiento de capacidades gerenciales y de liderazgo (London y Beatty, 1.993; Van Velsor y Leslie, 1.991). La finalidad central de estos programas o sistemas es, en concepto de London y Smither (1.995), proveer a los evaluados información proveniente de diferentes fuentes para ayudarlos a comprender cómo son vistos por otras personas y, en última instancia para ayudarles a mejorar su desempeño gerencial (Hazucha et al., 1.993). Aunque la evidencia informal y anecdótica sugiere que esos sistemas pueden funcionar, poca investigación ha demostrado mejoras en el desempeño a partir de estos sistemas multifuente (Greguras y Robie, 1.998).

Existen algunas razones y hechos que sugieren la existencia de una baja consistencia *inter – evaluador* en sistemas de medida basados en el concepto de evaluación de 360 grados. Este tipo de sistema evaluativo busca generar una medida de uno o varios atributos a partir de los conceptos de diferentes jueces o evaluadores, quienes suelen utilizar el mismo instrumento o los mismos criterios de evaluación. Suelen ser evaluadores típicos, en el contexto organizacional, el jefe, los colaboradores, los colegas o pares, los clientes y el propio evaluado. No siempre se incluyen todos estos evaluadores en un proceso. En ocasiones solo dos o tres de ellos participan.

Este tipo de evaluación suele ser problemático por varias razones. Entre dos o tres evaluadores suelen existir divergencias importantes al calificar un atributo. En estos casos siempre cabe preguntarse acerca de quién de ellos se aproxima de modo más cercano a la verdad.

Cada uno de los evaluadores suele contar con hechos, datos y juicios bien distintos acerca de un evaluado común a todos ellos. Es decir, que el jefe percibe a su evaluado en relación con metas, procesos, desempeños o expectativas que son propios de su rol de jefe

(Pulakos, Schmitt y Chan, 1.996). El cliente percibe desde su perspectiva de usuario de un producto o servicio, pero no desde el punto de vista de los procesos o deberes organizacionales de su evaluado. Los compañeros o pares, a su turno, perciben a su evaluado de acuerdo con el trato y relación que mantienen con él (Conway, 1.999), pero difícilmente pueden construir buenos juicios acerca de lo que percibe el cliente externo o de lo que percibe el jefe inmediato. Esto puede significar que, a pesar de que contemos con los mismos criterios de evaluación, las mismas variables o atributos o un mismo instrumento de medida, cada juez cuenta con datos obtenidos en un escenario de actuación diferente. Por tanto, los hechos y datos que alimentan sus juicios pueden ser bien distintos (Scullen, Mount y Judge, 2.003).

La naturaleza selectiva de los procesos perceptivos garantiza, en cierta medida, que dos o más personas que perciben una misma realidad pueden construir imágenes y juicios relativamente distintos de tal realidad (Álvarez, 1.992; James, James, Ashe, 1.990; Krech, Crutchfield y Ballachey, 1.965; Toro, 2.001). Si a este hecho agregamos que cada evaluador cuenta con juicios alimentados por diferentes realidades y percepciones debemos asumir que las calificaciones de diferentes evaluadores, con respecto a un mismo evaluado, pueden tener muy poco en común, aunque converjan en un mismo resultado cuantitativo.

Existen varios tipos de errores perceptivos que suelen distorsionar los juicios de los evaluadores. Cirino (1.993) menciona y analiza varios de ellos:

- *El Efecto de Halo*, consistente en una tendencia a evaluar o juzgar a una persona a partir de una impresión generalizada, usualmente creada a partir de una característica sobresaliente de la persona, positiva o negativa.
- *El Error de Tendencia Central*, por el cual los evaluadores tienden a calificar a una

persona utilizando de preferencia los valores centrales de la escala de medida y eludiendo los extremos.

- *El Error de Severidad o Indulgencia*, que consiste en que algunos evaluadores tienden a calificar con excesiva severidad o indulgencia a sus evaluados. De este modo algunos evaluados son calificados menos favorablemente y otros con más generosidad, dependiendo más del estilo del calificador y menos de la actividad evaluada.
- *El error de Contraste y Similitud*, que se refiere a la tendencia a evaluar a una persona según se diferencie o se parezca al evaluador en su forma de ser o de actuar.
- *El Prejuicio Personal*, suele ser un juicio de carácter emotivo o irracional que determina la valoración que un evaluador hace de su evaluado.

De otro lado, la literatura referente a la confiabilidad de sistemas de evaluación de 360 grados también ha evidenciado algunos hechos y datos que soportan las consideraciones precedentes. Algunos investigadores han reportado que los supervisores tienden a exhibir una más alta confiabilidad intercalificador que los pares (Conway y Huffcutt, 1.997; Scullen, Mount y Sytsma, 1.996; Viswesvaran, Ones y Schmidt, 1.996). Por otra parte, Conway y Huffcutt, (1997), Scullen et al., (1.996) reportaron que los supervisores registraron mejores confiabilidades que los subordinados. No obstante, poco se conoce sobre las puntuaciones de los pares y de los subordinados (Greguras y Robie, 1.998). A este respecto Fox y Bizman (1.988) encontraron que los subordinados tendían a enfatizar, en la evaluación de sus jefes, las habilidades gerenciales, las relaciones interpersonales y el conocimiento profesional, por encima de otros criterios de evaluación.

En resumen, hay evidencias que muestran que los jefes, los pares y los subordinados centran su atención en diferentes dimensio-

nes del desempeño en sus evaluaciones y esto puede explicar en parte las diferencias halladas en la confiabilidad al comparar a diferentes evaluadores.

Dadas las consideraciones y datos precedentes propusimos una investigación orientada a examinar la similaridad o divergencia de las calificaciones de varios evaluadores y el comportamiento de la confiabilidad del sistema para cada tipo de evaluador. A diferencia de otros estudios, incluimos una nueva fuente de información: la autoevaluación. Propusimos dos hipótesis para esta investigación:

- 1) Antes que consenso existen diferencias significativas entre las calificaciones dadas a un mismo evaluado por diferentes evaluadores, aún en el caso de utilizar los mismos criterios y la misma escala de medida.
- 2) La confiabilidad del sistema de evaluación difiere según cada tipo de evaluador.

MÉTODO

Muestra

Este estudio se llevó a cabo en relación con dos instrumentos de evaluación distintos utilizados en dos organizaciones diferentes.

En el primer caso se trató de una empresa privada del sector industrial. Allí se empleó la escala ALG que evalúa 15 tipos de errores o deficiencias de liderazgo (Toro, 2.001) y cuenta con una versión para cada tipo de evaluador. Fue validada en el medio colombiano y se registró una confiabilidad de .89 calculada mediante Alfa de Crombach. Para el propósito se contó con las siguientes cantidades de evaluadores: 63 personas de mando fueron evaluadas por sus jefes inmediatos, también fueron evaluadas por 217 colaboradores inmediatos (entre 3 y 4 por persona), igualmente fueron evaluadas por 102 pares o colegas de su mismo nivel (entre 1 y 2 por persona) y finalmente, 58 de las 63 personas de referencia se autoevaluaron.

En el segundo caso se trató de de una empresa proveedora de servicios de agua, energía, gas y telecomunicaciones, del sector público. Allí se estaba evaluando la calidad psicométrica de un sistema de calificación de competencias organizacionales que busca monitorear cinco competencias que se desagregan en 12 dimensiones o facetas específicas. Un conjunto de 106 descriptores conductuales operacionalizan tales competencias y sus dimensiones. Los evaluadores fueron 60 jefes que calificaron a 112 colaboradores inmediatos. Estos colaboradores se evaluaron ellos mismos y fueron también evaluados por 196 compañeros de trabajo (cerca de 3 compañeros por cada colaborador).

Procedimiento

En los dos casos se recogieron los datos de acuerdo con protocolos de aplicación, en un período corto, y se llevaron a cabo diversos análisis estadísticos cuyos resultados se registran a continuación.

En el primer caso se invitó a participar, como parte de un programa institucional, a todo el personal de dirección. Por razones de fuerza mayor algunas personas del equipo directivo no pudieron participar en el proceso. En el caso de la empresa de servicios pú-

blicos se eligieron a azar las personas que participaron en el proceso.

RESULTADOS

1. Correlación de las puntuaciones asignadas por los diferentes evaluadores.

Para este propósito se calcularon correlaciones de Pearson, por parejas, entre las calificaciones dadas por los distintos evaluadores a cada una de las variables del respectivo sistema de evaluación, es decir de la escala ALG o del sistema de Competencias. La tabla No. 1 muestra los valores de tales correlaciones en el caso de la empresa del sector privado. Aquí los evaluadores fueron la propia persona (un mando), su jefe inmediato, algunos de sus colaboradores inmediatos y algunos de sus pares o compañeros, tal como se expresó previamente.

Es evidente que, con muy pocas excepciones, la gran mayoría de las correlaciones no alcanzan a ser estadísticamente diferentes de cero, es decir, hay ausencia de correlación entre las calificaciones dadas a un mismo evaluado por unos y otros evaluadores. De hecho, solo el 16% de las correlaciones obtenidas son significativas. Adicionalmente los pocos resultados significativos tienden a ser relativamente bajos, según puede apreciarse.

Tabla No. 1 *Correlaciones de Pearson entre evaluadores que emplearon el ALG. (n = 70)*

VARIABLES DEL ALG	Auto vs. Colabor.	Auto vs. Jefe	Auto vs. Par	Jefe vs. Colabor.	Jefe vs. Par	Colabor. vs. Par
1. Inadecuada Supervisión	.08	-.02	.27	.40**	.54**	.23
2. Crítica al Colaborador	.24	-.00	.32*	.03	.33*	.22
3. Negación de Apoyo	.18	.21	.30*	.29*	.05	-.09
4. Toma Unilateral de Decisiones	-.13	-.08	.18	.06	.20	.04
5. Resistencia a Nuevas Ideas	.22	.27	-.02	.21	.15	-.07
6. Desinterés por el desempeño	.20	-.19	.15	.15	.31*	-.01
7. Pérdida de Ascendencia	.44**	.29*	.26	.00	.23	.13
8. Desconfianza del Colaborador	.06	.14	-.03	.32*	.23	.19
9. Falta de Reconocimiento	.13	.29*	.03	-.04	.28*	.01
10. Retroinformación Insuficiente	.12	.05	-.05	.14	.33*	.09
11. Asignación de Trabajo Tedioso	.12	.10	.10	.10	-.11	-.04
12. Robo o Negación de Méritos	-.01	.02	.13	-.04	.11	-.12
13. Falta de Acción Correctiva	.16	.02	.26	-.09	.36**	.11
14. Pedir Acciones no Éticas	-.02	.09	.07	.09	.20	-.03
15. No Cumplir lo Prometido	.21	-.10	-.13	.10	.37**	.01

** Nivel de significación $p < .01$ * Nivel de significación $p < .05$

La tabla No. 2 presenta el mismo tipo de resultados para la organización de servicios públicos. En este caso los evaluadores fueron la propia persona, su jefe inmediato y uno o varios compañeros, en proporciones que ya se describieron. El sistema evaluativo emplea-

do buscaba examinar el nivel de desarrollo de cinco categorías de competencias organizacionales y varias dimensiones dentro de ellas. Para tener una visión más detallada se presentan en la tabla los datos relativos a las dimensiones.

Tabla No. 2 *Correlaciones de Pearson entre evaluadores que calificaron las competencias organizacionales. (n = 113)*

VARIABLES	Auto vs. Jefe	Auto vs. Compa.	Jefe vs. Compa.
1. Escuchar	.14	.05	.24
2. Preguntar	.06	.08	.38*
3. Hablar Efectivamente	.19	.17	.13
4. Entender al Cliente	.35*	.01	.20
5. Satisfacer al Cliente	.07	.02	.34*
6. Orientación a Resultados	.16	.19	.39*
7. Optimización de Recursos	.19	.04	.41*
8. Mejoramiento Continuo	.33*	.20	.42*
9. Innovación	.25	.08	.44**
10. Construcción de Relaciones	.10	.20	.30
11. Trabajo en Equipo	.13	.20	.47**
12. Pensamiento Analítico	-.09	.24*	.34*
13. Búsqueda de Información	.11	.14	.35*
14. Sentido de Oportunidad	.17	.13	.18

** Nivel de significación $p < .01$ * Nivel de significación $p < .05$

Al igual que en el caso precedente, se evidencia que la mayoría de las correlaciones no alcanzan a ser significativamente diferentes de cero. Solo el 29% alcanza niveles de significación estadística y, aún en este caso, los valores obtenidos tienden a ser bajos.

Llama la atención que, tal como ocurrió en el primer caso, también en esta organización el mayor número de correlaciones significativas se encuentra entre el jefe y los compañeros del evaluado.

Este conjunto de datos pone en evidencia varios hechos de interés que proveen un buen soporte a nuestra primera hipótesis de trabajo.

- La mayor divergencia entre evaluadores se registra consistentemente entre la autoevaluación y la evaluación por otro calificador: jefe, par o compañero.

- Por contraste, se aprecia una mayor convergencia entre dos tipos de evaluadores: el jefe y el par o compañero. Es preciso advertir que par es una persona de similar nivel jerárquico pero que se desempeña en otra área o dependencia. El compañero de trabajo es alguien que pertenece a la misma área de trabajo o dependencia y comparte el espacio, los compañeros y los jefes.
- En cada caso algunas variables tienden a exhibir correlaciones más altas entre calificadores. Esto sugiere que su contenido o las conductas que representan facilitan mayores niveles de consenso entre evaluadores.
- Es evidente la escasa covariación entre las calificaciones de los diferentes evaluadores.

2. Significación de las diferencias en las calificaciones dadas por los diferentes evaluadores.

Otra manera de examinar el grado de convergencia entre las calificaciones dadas a un mismo evaluado por diversos evaluadores consistió en realizar un análisis de varianza y un test de significación de las diferencias por parejas (Tukey).

En el caso de la organización industrial, que realizó la evaluación con la escala de liderazgo ALG, es preciso tener en consideración que este instrumento demanda que el evaluador califique de acuerdo con un criterio de frecuencia de la conducta observada, en una escala de tipo Likert con rango entre 1 y 5, cuyos extremos equivalen a los conceptos de "rara vez o nunca" y "con mucha frecuencia o siempre" respectivamente (Toro, 2.001).

Se encontró que en 7 de los 15 criterios de evaluación se registraron diferencias significativas entre evaluadores, según puede apreciarse en la tabla No.3. Las autocalificaciones tendieron a alcanzar los valores más bajos. Por tratarse de una apreciación de la frecuencia con que se incurre en un desacierto de liderazgo la calificación será más positiva entre más cercana esté de 1, que significa "rara vez o nunca". De acuerdo con este criterio resultaron ser jueces más exigentes los jefes, en primera instancia, luego los pares y por último los colaboradores. La más benigna de las calificaciones fue la autoevaluación.

En cerca de la mitad de los criterios de calificación (53%) se encontró un relativo consenso entre evaluadores, justificado por la ausencia de diferencias significativas en el ANOVA. A pesar de este consenso de tipo estadístico, fue evidente la tendencia de los autoevaluadores a calificarse más favorablemente y la de los jefes y colaboradores a ser más exigentes.

Es notable que exista este tipo de divergencia entre calificadores que utilizan un instrumento cuyas condiciones psicométricas son adecuadas. Es evidente, entonces, la dificultad para decidir cuál de los resultados contiene un reflejo más fiel y creíble de la "verdad".

La tabla No. 4 muestra el mismo tipo de análisis para la empresa de servicios públicos que evaluó competencias organizacionales. Allí puede apreciarse que en casi todas las variables se encontraron diferencias significativas entre la autoevaluación y la evaluación por los compañeros. Solamente en uno de los criterios de calificación hubo consenso entre los tres evaluadores. También se encontraron algunas diferencias entre la autoevaluación y la calificación del jefe o entre éste y los compañeros, con una sola excepción, según puede apreciarse en la tabla.

También es evidente en la tabla No. 4 que las calificaciones más positivas corresponden a la autoevaluación y las menos favorables a la calificación dada por los compañeros. Los jefes dieron calificaciones con valores ubicados en algún punto intermedio entre estas dos evaluaciones extremas.

De acuerdo con estos resultados no resulta claro cómo construir un concepto válido sobre el evaluado, que inspire decisiones acertadas. Tal como lo asumimos en la primera hipótesis, se aprecian más diferencias que consenso entre evaluadores en este tipo de proceso.

Las evaluaciones del jefe tienden a mantener una distancia relativamente equivalente en relación con los otros dos calificadores (con excepción de dos criterios), razón por la cual no se diferencian significativamente de ninguno de ellos, es decir, conforman un cierto consenso con dos calificadores que difieren entre sí. Esto hace que no haya cómo decidir cuál de los evaluadores está más cerca de la verdad. Además cada uno podría estar refiriendo a una realidad en alguna medida diferente.

Tabla No. 3 ANOVA de las calificaciones dadas por diferentes evaluadores que emplearon el ALG.

VARIABLES	Diferencias entre Evaluadores	F	Significación Tukey
1. Inadecuada Supervisión	-	1.61	N.S
2. Crítica al Colaborador	-	1.93	N.S
3. Negación de Apoyo	-	1.52	N.S
4. Toma Unilateral de Decisiones	Jefe 2.2 - Auto 1.7 Par 2.3 - Auto 1.7	6.90	.003 .000
5. Resistencia a Nuevas Ideas	-	2.45	N.S
6. Desinterés por el desempeño	Auto 2.5 - Col. 2.2	4.83	.005
7. Pérdida de Ascendencia	Col. 2.0 - Auto 1.5 Par 2.1 - Auto 1.5	5.92	.002 .001
8. Desconfianza del Colaborador	Par 2.2 - Auto 1.7	5.16	.001
9. Falta de Reconocimiento	Par 1.8 - Col. 1.5	6.43	.001
10. Retroinformación Insuficiente	Jefe 2.1 - Col. 1.7 Jefe 2.1 - Auto 1.4	8.15	.007 .000
11. Asignación Trabajo Tedioso	-	3.79	N.S
12. Robo o Negación de Méritos	Col.1.4 - Auto1.1 Par 1.6 - Auto1.1	7.81	.001 .000
13. Falta de Acción Correctiva	-	3.00	N.S
14. Pedir Acciones no Éticas	-	1.37	N.S
15. No Cumplir lo Prometido	-	0.49	N.S

Tabla No. 4 ANOVA de las calificaciones dadas por diferentes evaluadores que calificaron las competencias organizacionales.

VARIABLES	Diferencias entre Evaluadores	F	Significación Tukey
1. Escuchar	Auto 4.2 - Comp. 3.9	6.57	.001
2. Preguntar	Auto 4.3 - Comp. 4.0	5.49	.004
3. Hablar Efectivamente	Auto 4.3 - Comp. 4.0 Jefe 4.4 - Comp. 4.0	8.63	.016 .001
4. Entender al Cliente	Auto 4.2 - Comp. 3.9 Jefe 4.2 - Comp. 3.9	8.60	.000 .049
5. Satisfacer al Cliente	Auto 4.2 - Comp. 3.9	5.02	.006
6. Orientación a Resultados	-	7.15	N.S
7. Optimización de Recursos	Auto 4.5 - Jefe 4.3 Auto 4.5 - Comp. 4.3	7.13	.031 .001
8. Mejoramiento Continuo	Auto 4.4 - Jefe 4.0 Auto 4.4 - Comp. 4.0	12.79	.002 .000
9. Innovación	Auto 4.0 - Comp. 3.7	8.39	.000
10. Construcción de Relaciones	Auto 4.3 - Jefe 4.0 Auto 4.3 - Comp. 3.9	10.51	.029 .000
11. Trabajo en Equipo	Auto 4.3 - Comp. 3.9	10.86	.000
12. Pensamiento Analítico	Auto 4.2 - Comp. 4.0	5.95	.002
13. Búsqueda de Información	Auto 4.3 - Comp. 4.1	4.24	.011
14. Sentido de Oportunidad	Auto 4.3 - Comp. 4.0	8.17	.000

3. Confiabilidad del sistema o instrumento de medida.

Una última opción analítica consistió en calcular la confiabilidad del instrumento o sistema de evaluación por cada tipo de evaluador. De este modo obtuvimos varias confiabilidades que se presentan en la tabla No. 5. Para el propósito se calculó el Alfa de Crombach con cada grupo de calificadores en cada empresa.

Tal como se propuso en la segunda hipótesis, encontramos que la confiabilidad del instrumento o del sistema de medida varía en función del evaluador que lo aplica. Es evidente que el nivel más bajo de confiabilidad se encontró, en ambas empresas, en la autoevaluación. Esta es la calificación que se dio a sí misma cada persona, independientemente de que fuera jefe o no. Aunque el resultado es satisfactorio en ambos casos, la consistencia interna de sus respuestas es inferior a la hallada para los demás evaluadores.

Tabla No. 5 *Coefficientes de Consistencia Interna Alfa por tipo de calificador para cada instrumento de evaluación.*

CALIFICADOR	Escala de Liderazgo ALG	Competencias Organizacionales
Autoevaluación	.76	.76
Jefe	.86	.82
Colaborador	.86	-
Par	.89	-
Compañero	-	.90

Por otra parte, los demás evaluadores tendieron a mostrar una mayor consistencia interna en sus respuestas y, al menos en la organización industrial, alcanzaron niveles de confiabilidad muy equivalentes. Algo diferente ocurrió en el caso de la organización estatal.

DISCUSIÓN

El conjunto de resultados precedentes pone en evidencia, una vez más, el hecho de la baja confiabilidad *inter-evaluador* que se ha reportado en la literatura especializada. Por tres métodos diferentes examinamos el asunto y hallamos soporte para nuestras hipótesis y confirmación de lo reportado en otras investigaciones.

Por otra parte, logramos establecer que la autoevaluación es un tipo de calificación relativamente menos confiable que las cali-

caciones de otros evaluadores. Pero adicionalmente tiende a ser la calificación más benigna y favorable que encontramos al compararla con las calificaciones de los demás evaluadores. Este tipo de resultado se puede entender desde varias perspectivas.

De otro lado, preservar un nivel conveniente de autoestima puede conllevar a minimizar la importancia de errores o desaciertos personales de los que tiene conciencia el individuo. Sobrevalorar los propios méritos o aciertos puede ayudar también al incremento o al sostenimiento de niveles de autoestima convenientes para la persona. (Bednar y Peterson, 1996).

La Teoría de la Atribución también arroja luz para la comprensión de este tipo de resultado. Este campo de la investigación ha puesto en evidencia que las personas tienden a subestimar sus desaciertos asumiendo que son causados por agentes externos o por algo aje-

no a su voluntad, como la mala suerte. Por otra parte, los éxitos o aciertos suelen ser explicados subjetivamente como dependiendo del esfuerzo o de la habilidad personal. Por contraste, los aciertos de otros suelen ser percibidos como determinados por agentes externos a la persona y sus desaciertos son imputados a su incapacidad o mala voluntad (Heider, 1.958; De la Coleta, 1.982; Toro, 1.986). Estas tendencias explicativas del acierto o del desacierto, tanto propio como ajeno, muestran que cada calificador cuenta con disposiciones personales evaluativas no solo diferentes sino aún opuestas a las de los otros.

El hecho de que este tipo de evaluaciones ocurra en un contexto organizacional puede mover a las personas que se autoevalúan a experimentar temor por las críticas, por las sanciones o por proyectar una imagen inconveniente. Estas preocupaciones pueden determinar las calificaciones más positivas encontradas en este estudio entre autoevaluadores. Los demás evaluadores, en calidad de observadores y jueces externos, pueden no verse afectados por estas ideas o sentimientos, dado que no necesitan defenderse de sus temores o poner en tela de juicio su propia autoestima.

Greguras y Robie (1.998) concluyeron en su estudio que el jefe o supervisor puede ser la fuente de información más confiable y más creíble por varias razones. Puede compartir cierto número de valores y conceptos con el evaluado. Puede estar mejor habilitado como evaluador por ser parte de su rol prestar atención y dar retroalimentación a su evaluado. También puede ser un observador más experimentado que otros evaluadores y contar con más oportunidades para observar la conducta de su evaluado. Está más familiarizado con las posibilidades de acierto o desacierto de su colaborador.

El caso de los colaboradores o subalternos resulta también de interés. Estas personas tendieron a ser jueces benignos. Es probable

que su calificación esté influenciada por el temor a ser identificado, a suscitar sanciones o represalias por parte de su jefe o por una actitud de aceptación pasiva de la autoridad. En la cultura colombiana las personas de autoridad suelen ser vistas y tratadas de modo ambiguo por sus colaboradores. Alvarez (1.982; 1.983) en estudios colombianos referidos a estos asuntos puso en evidencia que los subalternos tienden, por una parte, a experimentar sentimientos de respeto, temor y reverencia hacia las personas investidas de autoridad formal. Al mismo tiempo pueden experimentar odio o rencor, en particular si el trato recibido les resulta demandante o agresivo. Es probable que en muchos casos la actitud aceptante y reverencial haga sentir al colaborador más seguro cuando se ve en la necesidad de calificar a su jefe.

Por otra parte, otros analistas explican la baja confiabilidad *inter - evaluador* por el tipo de programa de entrenamiento al que se sometió a los evaluadores (Church y Bracken, 1.997), por las oportunidades de observar, que suelen ser diferentes para cada tipo de evaluador (Rothstein, 1.990) o por la ocasión en que se genera la calificación (Greguras y Robie, 1.998), tal como decidir sobre una mejora salarial, la posibilidad de un ascenso o aún la posibilidad de un despido.

Greguras y Robie (1.998), en un estudio basado en el Análisis de Generalizabilidad, encontraron que las siguientes son las cantidades mínimas de evaluadores y reactivos que se requerirían para lograr niveles de confiabilidad aceptables en un sistema evaluativo de 360 grados: una escala o instrumento de medida con un mínimo de 5 items por cada variable, la calificación de 4 supervisores o jefes, la de 8 pares y la de 9 subordinados. Evidentemente esto puede resultar costoso, poco práctico y complejo como para justificar este tipo de evaluación. No obstante, en el concepto de estos investigadores, elevar el número de evaluadores de

cada tipo contribuiría de modo sustancial a mejorar la confiabilidad de estos sistemas.

CONCLUSIONES

Este tipo de evidencia que hemos analizado cuestiona la utilidad de los sistemas evaluativos y de retroalimentación de 360 grados. No se puede saber cuál evaluador y cuál resultado contiene más información cierta. Son tantas y tan diversas las circunstancias que inciden sobre el juicio y el concepto de cada evaluador que cada uno de ellos podría estar suficientemente equivocado como para no merecer crédito alguno. Igualmente, cada uno de ellos podría ser testigo de una verdad merecedora de todo crédito. No obstante, el sistema no permite distinguir estas condiciones.

Desde el punto de vista práctico, no resulta conveniente aumentar el número de evaluadores de cada tipo por varias razones. Por una parte puede resultar costoso y consumir mucho tiempo. No suele ser posible aumentar a más de 5 el número de jefes, dado que normalmente cada persona tiene un número reducido de jefes o supervisores.

Si ha de emplearse este modo de evaluación y retroalimentación conviene tener en consideración que cada tipo de evaluador puede ser mejor o peor que otros, desde el punto de vista de la confiabilidad de sus respuestas. En este estudio encontramos que el jefe y el compañero o par tendieron a calificar de modo más convergente y más confiable. El colaborador y la autoevaluación resultaron ser menos confiables.

Encontramos que existe una manera conveniente de sacar partido de un sistema evaluativo de 360 grados, a pesar de las limitaciones que hemos constatado. A este propósito contribuye notablemente el concepto de Autoconciencia Gerencial. Se refiere a la habilidad de un gerente para juzgar con exactitud los propios comportamientos y habili-

dades, tal como se manifiestan en las interacciones que ocurren en el lugar de trabajo (Van Velsor, Taylor y Leslie, 1.993). Para Ashford y Tsui (1.991) se trata de una habilidad que se asocia a los niveles de excelencia gerencial. En reporte reciente de investigación se ha encontrado que gerentes de alto rendimiento fueron consistentemente más capaces de evaluar con exactitud sus propios comportamientos, mostrando mayor congruencia entre su autoevaluación y la evaluación que le realizaron otras personas (Church, 1.997).

La manera usual de evaluar la Autoconciencia Gerencial consiste en establecer la diferencia entre la autoevaluación de la persona y la que le realizan otros evaluadores (Church, 1.997). Será mejor la autoconciencia en la medida en que tal diferencia se aproxime a cero. Una diferencia de cero significa una alta coherencia entre la visión que una persona tiene de sus aciertos y desaciertos, de sus fortalezas y limitaciones y la que uno o varios observadores externos tienen de los mismos hechos o realidades. En un estudio realizado a este respecto en empresas colombianas obtuvimos varias conclusiones que son relevantes en la presente investigación (Toro, 2.001). Vistos los resultados de diversas evaluaciones practicadas a una misma persona por distintos evaluadores, siendo uno de ellos el propio evaluado, se puede tener la seguridad de que los resultados no reflejan la percepción independiente de los jefes ni los juicios poco ilustrados de los colaboradores, muestran mas bien el nivel de consenso que existe entre unos y otros alrededor de modos de comportamiento específicos. Por otra parte, se concluyó que la retroalimentación dada sobre la base de la poca o mucha convergencia entre evaluadores tiene una alta capacidad de impacto personal, de credibilidad y un buen potencial de confrontación de los esquemas y construcciones personales con la realidad, tal

como se ve en conjunto con otros miembros del mismo equipo (Toro, 2.001).

Este tipo de análisis no resuelve los inconvenientes de confiabilidad inter – evaluadores que se han reportado pero permite obtener beneficios prácticos a la hora de utilizar esta modalidad evaluativa.

En esta investigación no se tuvieron en cuenta las diferencias demográficas de los evaluadores. Es probable que el género, la experiencia laboral y la experiencia en el rol de jefe, la antigüedad en la organización y en el cargo, el nivel de educación formal, el tipo de entrenamiento recibido como evaluador y otras condiciones similares puedan incidir sobre la confiabilidad de estos sistemas de evaluación o sobre la confiabilidad en la calificación de un evaluador particular.

Tampoco se hizo un control del instrumento de medida. Resulta probable que la confiabilidad entre evaluadores pueda ser mejor o peor dependiendo del tipo de variables que se evalúen y del tipo de instrumento de medida. Quizás puede ser más objetivo y basado en hechos un evaluador que debe juzgar modos de actuación y desempeños laborales específicos que uno comprometido con la evaluación de atributos, cualidades o com-

petencias. Por otra parte, puede ser más objetivo y confiable diligenciar un instrumento de medida breve, con descripciones conductuales y criterios de calificación objetivos y simples que uno con muchas definiciones, variables y distinciones conceptuales complejas.

Se omitió también, por razones de delimitación, controlar la ocasión de la evaluación. Esta es una variable compleja. Probablemente resulte diferente la calidad de una evaluación cuando se trata de decidir sobre si conceder o no una mejora salarial o de conservar o despedir a una persona, de iniciar una acción educativa o sencillamente de proveerle retroalimentación sobre su desempeño.

Sugerimos a otros investigadores tener en consideración restricciones como las aquí mencionadas para abordar el estudio de estos campos y variables que no fue posible controlar en la presente investigación.

Sugerimos a los profesionales que tienen a cargo este tipo de sistemas que verifiquen siempre sus condiciones psicométricas. Que constaten permanentemente su validez y confiabilidad, que entrenen exhaustivamente a los calificadores y usuarios del sistema y que les ayuden a entender y aceptar las importantes limitaciones del sistema.

Referencias

- Alvarez, G. (1.992). La percepción de la organización: Clave para la comprensión del comportamiento del individuo en la organización. *Revista Interamericana de Psicología Ocupacional*. Vol. 11, No. 1 y 2.
- Alvarez, R. (1.982). Dimensiones semánticas de las relaciones de poder. *Revista Interamericana de Psicología Ocupacional*, Vol. 1, No. 3
- Alvarez, R. (1.983). Dimensiones del poder en las organizaciones públicas colombianas. *Revista Interamericana de Psicología Ocupacional*, Vol. 2, No. 2
- Ashford, S.J. y Tsui, A.S. (1.991). Self-regulation of managerial effectiveness: The role of active feedback seeking. *Academy of Management Journal*, 34, 251 - 280
- Bednar, R.L., y Peterson, S.R. (1.996). *Self-esteem. Paradoxes and innovations in clinical theory and practice*. Segunda Edición. Washington: APA

- Cirino, G. (1993). *La evaluación dinámica del desempeño*. República Dominicana: Ed. Corripio.
- Conway, J.M. (1999). Distinguishing contextual performance from task performance for managerial jobs. *Journal of Applied Psychology*, 84, 3 - 13
- Conway, J.M. y Huffcutt, A.I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer and self-ratings. *Human Performance*, 10, 331 - 360.
- Church, A.H. (1997). Managerial self-awareness in high-performance individuals in organizations. *Journal of Applied Psychology*, Vol. 82, No. 2, 281 - 292.
- Church, A.H. y Bracken, D.W. (1997). Advancing the state of the art of 360-degree feedback. *Group and Organization Management*, 22, 149 - 161.
- Dela Coleta, J.A. (1982). *Atribuição de causalidade. Teoria e pesquisa*. Rio de Janeiro: Fundação Getulio Vargas.
- Greguras, G.J., y Robie, C. (1998). A new look at within-source interrater reliability of 360-degree feedback ratings. *Journal of Applied Psychology*, 83, 960 - 968.
- Hazucha, J.F., Hezlet, S.A., y Schneider, R.J. (1993). The impact of 360-degree feedback on management skills development. *Human Resource Management*, 32, 325 - 351
- Heider, F. (1958). *The Psychology of Interpersonal Relations*. New York: Wiley.
- James, L.R., James, L.A. y Ashe, D.K. (1990). The meaning of organizations: The role of cognitions and values. En Schneider, B. *Organizational climate and culture*. San Francisco: Jossey - Bass.
- Krech, D., Crutchfield, R.S. y Ballachey, E.L. (1965). *Psicología Social*. Madrid: Biblioteca Nueva.
- London, M. y Smither, J.W. (1995). Can multi-source feedback change perceptions of goal accomplishment, self evaluations, and performance-related outcomes? Theory-based applications and directions for research. *Personnel Psychology*, 48, 803 - 839.
- London, M. y Beatty, R.W. (1993). 360 degree feedback as competitive advantage. *Human Resource Management*, 32, 353 - 373.
- Pulakos, E.D., Schmitt, N. y Chan, D. (1996). Models of job performance rating: an examination of rater race, rate gender and rater level effects. *Human Performance*, 9, 103 - 119
- Rothstein, H.R. (1990). Interrater reliability of job performance ratings: growth of asymptote level with increasing opportunity to observe. *Journal of Applied Psychology*, 75, 322 - 327
- Scullen, S.E., Mount, M.K. y Judge T.A. (2003). Evidence of construct validity of developmental ratings of managerial performance. *Journal of Applied Psychology*, 88, 50 - 66
- Scullen, S.E., Mount, M.K. y Sytsma, M.R. (1996). *Comparisons of self, peer, direct report, and boss ratings of managers' performance*. Ponencia presentada en la 11 reunión anual de la Sociedad de Psicología Industrial y Organizacional, San Francisco. CA.
- Toro, F. (1986). Teoría de la Atribución: Un aporte a la comprensión del desempeño ocupacional. *Revista Interamericana de Psicología Ocupacional*. Vol.5, No.2 y 3
- Toro, F. (2001). *El clima organizacional. Perfil de empresas colombianas*. Medellín: Cincel.
- Toro, F. (2001). *Autoconciencia Gerencial en un grupo de gerentes del sector industrial*. Ponencia presentada en el XXVIII Congreso Interamericano de Psicología. Santiago de Chile.

- Van Velsor, E., y Leslie J. B. (1991). *Feedback to managers: Vol. 1. A guide to evaluating multi - rater feedback instruments*. Greensboro, NC. Center for creative leadership.
- Van Velsor, E., Taylor, S. y Leslie, J.B. (1993). An examination of the relationship among self-perception accuracy, self-awareness, gender and leader effectiveness. *Human Resource Management*, 32, (2-3), 249 – 263.
- Viswesvaran, C., Ones, D.S., y Schmidt, F.L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, 81, 557 – 574.

Fernando Toro Álvarez
Centro de Investigación e Interventoría en Comportamiento Organizacional
Calle 34 No. 85 C – 20
cincel@epm.net.co
Medellín, Colombia.