WPOM

Working Papers on Operations Management.
Vol. 6, Nº2 (49-63)
ISSN: 1989-9068

http://dx.doi.org/10.4995/wpom.v6i2.4032

Protocol: Comparing advantages and disadvantages of Rating Scales, Behavior Observation Scales and Paired Comparison Scales for behavior assessment of competencies in workers. A systematic literature review.

Protocolo: comparación de las ventajas e inconvenientes de las "Rating Scales", "Behavior Observation Scales", y " Paired Comparison Scales" para la evaluación basada en comportamientos, de las competencias de los trabajadores de empresas. Una revisión sistemática de la literatura.

Juan A. Marin-Garcia[a], Lucia Ramirez Bayarri[b], Lorena Atares Huertas[c]

[a] ROGLE. Dpto. de Organización de Empresas. Universitat Politècnica de València. Camino de Vera S/N 46021 Valencia. jamarin@omp.upv.es, [b]Department d'Ensenyament de la Generalitat de Catalunyac luraba@gmail.com, [c]Departamento Tecnología de Alimentos, Universitat Politècnica de València, loathue@tal.upv.es

Recibido: 2013-04-25 Aceptado: 2014-07-07

Abstract

This is the protocol for a review and there is no abstract. The objectives are as follows: Identify the characteristics of each of the types of scale and how they differ from each other. Estimate the extent to which they are used by organizations in the assessment of skills of their employees. Summarize the advantages and disadvantages of each. Propose which of them would be more appropriate for assessing the competence of innovation in people (performance evaluation, promotion of workers, recruitment, etc.) and the mode of administration (self, peer, expert assessment).

Keywords: *Protocol; Systematic literature review; Rating Scales; Behavior Observation Scales; Paired Comparison Scales; behavior assessment.*

Resumen

Por tratarse de un protocolo para una revisión sistemática no existe un resumen propiamente dicho. En esta revisión nos planteamos los siguientes objetivos: identificar qué características tiene cada uno de los tipos de escala y cómo se diferencian unas de otras. Estimar el grado en que son usadas, unas u otras, por las organizaciones en la evaluación de competencias de sus empleados. Resumir las ventajas e inconvenientes que presentan cada una de ellas en general y a la hora de evaluar la competencia transversal de innovación en las personas. Proponer cuál de ellas sería más adecuada, atendiendo a los objetivos específicos de evaluación de la competencia de innovación en las personas (evaluación del desempeño, promoción de trabajadores, selección de personal, etc.) y el modo de administración (self, peer, expert assessment).

Palabras clave: *Protocolo; Revisión Sistemática de literatura; Rating Scales; Behavior Observation Scales; Paired Comparison Scales; evaluación comportamiento.*

Antecedentes para la revisión sistemática

La evaluación de competencias, consiste en una valoración formal del desempeño por medio de ciertos indicadores. Estos indicadores pueden ser de naturaleza más o menos objetiva, en función de las dimensiones de la competencia que se desee evaluar.

Actualmente se observa un creciente interés en las organizaciones por la evaluación de las competencias de sus trabajadores. Quizás este movimiento sea debido a la necesidad de acreditar unas capacidades y/o competencias de modo que tengan una validez o una representatividad más general. O, quizás, sea debido a que se considera un complemento necesario a la evaluación de resultados como elemento para predecir el desempeño futuro de la persona o del grupo/departamento en el que trabaje. Sea como fuere, la evaluación de competencias parece tener implicaciones importantes en los procesos de reclutamiento, evaluación del desempeño y el desarrollo de los profesionales que trabajan en una organización (Boyatzis, 2008; Moore, Cheng, & Dainty, 2002; Rowe, 1995)

Esta evolución se aprecia también en las universidades. Si bien es cierto que se ha empezado a usar mucho más tarde que en las empresas, ahora no es raro oír hablar de "evaluar las competencias" de los alumnos, e incluso algunas universidades se plantean como objetivo aumentar la incidencia de la evaluación basada en competencias, contraponiéndola a la evaluación basada en conocimientos, con el objetivo de fomentar un aprendizaje más orientado a las demandas que las organizaciones o la sociedad exigirán a los profesionales que se forman en sus aulas.

En ambos casos, empresas y universidades, parece que la evaluación de competencias no es una tarea fácil. Por un lado requiere, de los evaluadores, una inversión de tiempo que probablemente sea mayor que otros tipos de evaluación. Un tiempo, que muchos mandos o profesores no son capaces de reservar, urgidos por sus múltiples obligaciones diarias. Esto origina que, algunos evaluadores, se sientan incómodos con los instrumentos propuestos para realizar la tarea de evaluar competencias y la realicen de manera poco comprometida, rellenando formularios porque "hay que rellenarlos", pero sin poder dedicar el tiempo que consideran necesario para hacer estas evaluaciones con garantías. Por otro lado, la evaluación de competencias se suele asociar a un modo subjetivo de evaluar. Esto puede restar credibilidad a la evaluación, tanto del lado del evaluador como del lado del evaluado, achacándose una posible falta de validez o de fiabilidad a las puntuaciones (Baartman, Bastiaens, Kirschner, & van der Vleuten, 2006; Hartig, 2008; Lenburg, 1999; Marin-Garcia, Aragonés Belgran, & Melón, 2014).

En definitiva, podemos considerar la evaluación de competencias como una tarea importante y complicada, tanto cuando se aplica a trabajadores de empresas, como en evaluación de estudiantes universitarios, que luego van a ser profesionales en las empresas. Probablemente, el primer paso en la evaluación de competencias, es seleccionar qué dimensiones de desempeño o tareas clave se desean evaluar (Lohmann & Prumper, 2006; Marin-Garcia, Perez-Peñalver, & Watts, 2013) y en qué situaciones o contextos evaluarlo. Una vez completado, la siguiente etapa es establecer cómo tomar los datos para decidir el grado en que las personas superan cada una de las dimensiones o tareas. Es, precisamente, en el contexto de esta etapa en la que enmarcamos este proyecto, para identificar qué ventajas e inconvenientes tiene cada tipo de escala de medida.

Descripción del contexto: evaluación de competencias por observación de comportamientos (*behavior assessment*)

Una de las formas de obtener datos para la evaluación de competencias, es a través de la observación de comportamientos. La evaluación del comportamiento se centra en "identificar aspectos claramente observables y el modo en que la persona interactúa con su entorno" (Groth-Marnat, 2003). Los principales instrumentos para la evaluación del comportamiento son los cuestionarios y estrategias de observación (Dowdy, Twyford, & Sharkey, 2013). Los cuestionarios presentan la ventaja de que pueden usarse para auto-evaluación y para evaluación por parte de pares o mandos/profesores. Mientras que las estrategias de observación, normalmente, requieren de observadores externos dedicados a esta función (estos observadores suelen ser expertos instruidos para tal fin).

Descripción de la intervención: tipos de escalas

Dentro de los cuestionarios, podemos considerar varias escalas para recogida de los datos. Tres de ellas recogen los datos en términos absolutos: las escalas de puntuación (*rating scales*), las escalas de evaluación ancladas en comportamientos (BARS) y las escalas de observación de frecuencia de comportamientos (*BOS*). La cuarta utiliza la comparación, bien sea pareada (*paired comparisons*) o con listas ordenadas (*rankings*) de personas (Dolan, Valle Cabrera, Jackson, & Schuler, 2007; Dowdy et al., 2013; Hatzinger & Dittrich, 2012; Marin-Garcia, Garcia-Sabater, Maheut, Valero-Herrero, & Andres-Romano, 2012).

Las escalas de puntuación se caracterizan porque las categorías de respuesta, compartidas por diferentes preguntas o ítems en el cuestionario, se identifican con números, o distancia respecto a límites o anclas verbales (Marin-Garcia, Ramirez Bayarri, & Andreu Andres, 2015) (ver también http://www.oxfordreference.com/view/10.1093/oi/authority.20110803095904227?rskey=VE2Fjz&result=3) (Figura 1).

Dimensión de desempeño:	Novato 1	2	3	4	Experto 5
Comunicación interpersonal - Escucha					
Innovación en métodos de trabajo					
Otra dimensión de desempeño o tarea clave …					

Figura 1: Ejemplo escala de puntuación

Por otra parte, las escalas de evaluación ancladas en comportamientos (BARS), identifican los aspectos claves de un puesto de trabajo y las posibles conductas que pueden mostrar las personas que ocupan esos puesto al desarrollar sus tareas, ordenándolas desde las más ineficientes o indeseables, hasta las más eficientes o deseables. A diferencia de las escalas de puntuación, en las BARS, cada uno de los aspectos claves, tiene unas descripciones distintas para cada uno de los niveles de desempeño (Marin-Garcia et al., 2015)(ver también http://www.oxfordreference.com/view/10.1093/oi/authority.20110803095456792?rskey=cWbwX8&result=3). En la Figura 2 mostramos un ejemplo de BARS asociado las mismas dimensiones de desempeño

que hemos mostrado en el ejemplo de la Figura 1. Al construir las BARS, al igual que en las escalas tipo Guttman, se supone que los niveles de desempeño superiores sólo pueden lograrse si se han superado los inferiores. Por lo tanto, en cada una de las filas se seleccionaría sólo una casilla: la correspondiente al nivel de desempeño más elevado que se haya observado en la persona evaluada.

Dimensión de desempeño:	1	2	3	4
Comunicación interpersonal - Escucha	No mira al resto de personas cuando hablan y parece no prestar atención a lo que se trata en la reunión (más pendiente de otros estímulos que de la reunión de su grupo)	Mira y asiente cuando los demás hablan	Reformula oralmente las intervenciones de los demás	Toma nota y registra las intervenciones de los demás, reformulándolas por escrito, para que puedan ser usadas como "memoria" efectiva de la reunión.
Innovación en métodos de trabajo	Usa los procedimientos tradicionales sin plantearse nada más (por ejemplo, los explicados en clase o en los recursos básicos de la asignatura)	Es consciente de que hay limitaciones y problemas con los procedimientos o métodos tradicionales. aunque no sabe identificar cuáles son esos problemas	Identifica y define las limitaciones y problemas de los métodos tradicionales	Propone/aplica nuevos procedimientos y acciones en el propio proceso de trabajo
Otra dimensión de desempeño o tarea clave …	…	…	…	..

Figura 2.- Ejemplo de BARS

Las escalas de observación de frecuencia de comportamiento (BOS), al igual que las BARS, identifican una serie de conductas observables en las personas a evaluar. Sin embargo, se recogen datos de la frecuencia en que los evaluados muestran esas conductas a lo largo del tiempo establecido para la observación. Las puntuaciones de las observaciones de cada conducta se pueden ponderar con diferentes pesos para calcular una puntuación agregada en cada dimensión de desempeño. Son especialmente adecuadas cuando se quiere evaluar el proceso, no sólo el resultado o la productividad de las personas (Marin-Garcia et al., 2015) (ver también: http://www.oxfordreference.com/view/10.1093/oi/authority.20110803095456852). En la Figura 3 mostramos un ejemplo de BOS para la primera de las dimensiones de la BARS del ejemplo de la Figura 2. Nótese que cada una de las opciones de la BARS, que es un comportamiento observable, da origen a una medida en BOS, que pueden ponderarse para calcular la puntuación de la dimensión. En ocasiones se puede usar una escala de 0% a 100% del tiempo y la suma de todos los comportamientos de una dimensión de desempeño deben sumar 100%.

Dimensión de desempeño: Comunicación interpersonal - Escucha	No mira al resto de personas cuando hablan y parece no prestar atención a lo que se trata en la reunión (más pendiente de otros estímulos que de la reunión de su grupo)	Mira y asiente cuando los demás hablan	Reformula oralmente las intervenciones de los demás	Toma nota y registra las intervenciones de los demás, reformulándolas por escrito, para que puedan ser usadas como "memoria" efectiva de la reunión.	No se observa ninguna
Con qué frecuencia te comportas de las siguientes maneras cuando formas parte de grupos que buscan soluciones u oportunidades (toma como referente grupos de trabajo. Si no estás trabajando, puedes usar los grupos para tareas como estudiante u otros grupos en los que participes –por ejemplo asociaciones, etc.-)	0 Nunca 1 Muy pocas veces 2 Pocas veces 3 Bastantes veces 4 Muchas veces 5 Casi siempre Sin respuesta	0 Nunca 1 Muy pocas veces 2 Pocas veces 3 Bastantes veces 4 Muchas veces 5 Casi siempre Sin respuesta	0 Nunca 1 Muy pocas veces 2 Pocas veces 3 Bastantes veces 4 Muchas veces 5 Casi siempre Sin respuesta	0 Nunca 1 Muy pocas veces 2 Pocas veces 3 Bastantes veces 4 Muchas veces 5 Casi siempre Sin respuesta	0 Nunca 1 Muy pocas veces 2 Pocas veces 3 Bastantes veces 4 Muchas veces 5 Casi siempre Sin respuesta

Figura 3.- Ejemplo de BOS

Por último las comparación pareada (*Paired comparison*) es un tipo de escala donde parejas de personas y el evaluador debe decidir cuál de los dos presenta más cantidad, frecuencia, habilidad o competencia en la dimensión de desempeño que se está evaluando (Marin-Garcia et al., 2014). El número de comparaciones a realizar es n*(n-1)/2 por cada dimensión de desempeño a evaluar (siendo n el número de sujetos a evaluar). Una vez rellenadas las matrices de comparación, se pueden establecer diferentes procedimientos para crear una lista ordenada de mayor a menor desempleo (Dolan et al., 2007; Marin-Garcia et al., 2014; Saaty, 1980, 1996).

Cómo funcionan los diferentes tipos de escalas

De los cuatro modelos, el más sencillo de crear y de administrar es el de *Ratings*. Sólo es necesario tener identificadas las dimensiones de desempeño (ítems) y la cantidad de niveles de respuesta (por ejemplo; Sí/No; poco/moderado/mucho; excelente/aceptable/necesita mejorar…). Estos cuestionarios, suelen requerir menos tiempo por parte del evaluador, porque hay que leer menos. Sin embargo, las *Ratings* pueden dar lugar a diferentes interpretaciones de lo que se quiere medir, o incluso de lo que representa cada uno de los niveles (tanto en un propio evaluador cuando la aplica en diferentes momentos, como cuando intervienen diferentes evaluadores). Por lo tanto, es posible que este tipo de escalas, a pesar de ser más sencillas y rápidas de desarrollar, tengan problemas de validez o de fiabilidad.

Las BARS, aparentemente, presentan la ventaja de describir con más claridad el comportamiento asociado a cada nivel de desempeño, de modo que es más probable interpretarlo de manera similar por varios evaluadores, incluso por el propio evaluado. Por este motivo, las BARS se consideran un método más fiable y objetivo que las escalas de puntuación. También ofrecen un patrón al evaluado para mejorar su desempeño sabiendo a qué comportamiento debería aspirar, partiendo del nivel que tiene. Debido a ello, este tipo de escalas presentan un potencial de cara a la formación y desarrollo de la competencia. Sin embargo, las BARS son muy costosas de desarrollar, se pueden tardar meses y se necesita la intervención de personas expertas en la creación de estos instrumentos. No suele ser fácil crear una anidación completa de los comportamientos, de modo que es posible que una persona, esté mostrando simultáneamente comportamientos asociados a diferentes niveles de desempeño, pero el evaluador tiene que optar por una única puntuación. Por el contrario, se consideran más fiables y las decisiones a tomar suelen ser más sencillas porque está más claro qué se debe observar.

Las BOS, superan alguna de las limitaciones de las BARS porque cada nivel de desempeño se puede evaluar de manera independiente de los otros. Sin embargo requieren más tiempo de creación porque a la complejidad de las BARS hay que añadir el protocolo para capturar y anotar las observaciones. Además exigen de una dedicación muy intensa y sistemática del evaluador para cumplir el protocolo, lo que puede hacerlas difíciles de aplicar, sobre todo en contextos con mucho evaluados (por ejemplo en procesos de selección con muchos candidatos). También requieren de más tiempo de evaluador para recoger una muestra representativa de comportamientos, por lo que su aplicación acaba resultando más costosa que las dos escalas anteriores.

Por último, las comparaciones pareadas suelen generar resistencias tanto en evaluadores como evaluados porque consideran que no es correcto comparar a dos personas. Parece que lo asocian a algún tipo de problema ético al tener que elegir entre la valía de dos seres humanos. Sin embargo, llama la atención que las personas que tienen objeciones para usar la comparación pareada, consideran muy adecuado usar las escalas de puntuación o las BARS o BOS para evaluar el desempeño de las personas. No obstante, puesto que las decisiones que tiene que tomar el evaluador son muy sencillas, cada par es evaluado muy rápido y los datos recogidos de esta forma suelen gozar de más fiabilidad y validez que los recogidos con otro tipo de escalas. En el fondo se pasa de tomar unas pocas decisiones muy complejas, con los otros tipos de escalas, a tomar muchas decisiones, pero mucho más sencillas con las comparación pareada. Relacionado con esto, a medida que se incorporan personas a evaluar, el número de comparaciones crece muy rápidamente, por lo que resulta poco práctico emplearlo cuando hay más de 5-7 personas en el grupo a comparar. Existen algunas investigaciones que intentan reconstruir la matriz de comparaciones a partir de un conjunto más reducido de datos, sin necesidad de comparar a todos los pares posibles, pero aún no se ha acordado un método definitivo para lograrlo.

Revisiones o publicaciones anteriores sobre el tema

Existen algunos trabajos que resumen las ventajas y limitaciones de las observaciones de comportamiento frente a las escalas de calificación, la frecuencia con las que se usan o las recomendaciones para la puesta en práctica en general de esta estrategia de evaluación. En el área de psicología clínica parte de esa información se pueden encontrar en Dowdy et al (2013). Marin-Garcia et al. (2015) hacen un análisis del uso en educación, donde las Rúbricas pueden considerarse el equivalente de las BARS/BOS.

En el campo de la gestión de los recursos humanos, parece que las revisiones se han centrado en el uso de las escalas para la evaluación del desempeño. En este contexto, a pesar de que las *rating scales* parecen estar más extendidas que otros modelos de escalas, propician ciertos sesgos que podrían ser resueltos con la aplicación de las BOS (Tziner & Kopelman, 2002). Sin embargo, el uso de BARS, BOS o Paired comparison no parecía demasiado extendido en el momento de algunas revisiones y los resultados acerca de las ventajas de cada tipo de escala no eran muy concluyentes a principios de 1990 (Bretz, Milkovich, & Read, 1992). Parece que una fuente importante de varianza del sesgo en la evaluación se debe a las diferentes interpretaciones que los evaluadores hacen de la escala (Hoyt & Kerns, 1999), pero queda por demostrar qué tipo de escala favorece una interpretación más homogénea de los niveles de puntuación.

En la revisión más reciente localizada (DuVernet, Dierdorff, & Wilson, 2015), se propone que el tipo de escala podría tener consecuencias en la fiabilidad y/o validez de los datos obtenidos. El enfoque de estos autores se centra en analizar la longitud de la escala, el tiempo necesario para contestarla y otras características que no encajan directamente con los objetivos de nuestra investigación, sin embargo, podrían ser una fuente interesante de referencias a extraer por el procedimiento de bola de nieve. De manera análoga, otras revisiones parecen confirmar que el tipo de escala o los contenidos de la escala afectan a la fiabilidad y validez, pero no queda claro cuál es mejor, o en qué condiciones es mejor una escala que otra (Heidemeier & Moser, 2009). Para los objetivos de esta revisión nos parece especialmente relevante la revisión de Voskuijl y Sliedregt (2002) que concluyen que las evaluaciones que usan escalas tipo BARS o BOS tienen una fiabilidad más alta que las escalas tipo *rating*. Sin embargo, el conjunto de artículos revisados (38) es limitado y bastante antiguo (publicados antes de 1998). Además, se obtuvieron con una estrategia de búsqueda no especificada y con unos criterios de inclusión muy restrictivos en cuanto a la base de datos utilizada para búsqueda automática como a las condiciones para ser seleccionados.

Por qué es importante hacer esta revisión

Desde el punto de vista práctico, sería de gran ayuda para los responsables de recursos humanos contar con una evidencia científica acerca del tipo de escala que sería más conveniente para evaluar determinadas competencias. De este modo, disfrutarían de una mejor estimación de las evaluaciones y podrían tomar decisiones con datos de mejor calidad, al tiempo que evitarían actuar por "ensayo y error" en el desarrollo de unos instrumentos que son muy caros de desarrollar y afecta a la motivación y clima en la empresa.

Para los investigadores, esta revisión aportaría evidencias para la teoría de medición de comportamiento y, basándose en ella, podrían desarrollar modelos de decisión para la aplicación de la evaluación de competencias en las organizaciones.

Ninguna de las revisiones localizadas hasta el momento es capaz de dar una respuesta concluyente a las preguntas que nos planteamos, por lo que consideramos necesaria nuestra revisión para saber si algún estudio individual ha resuelto alguna de ellas o si, a partir de la colección de artículos publicados es posible dar respuesta a nuestras preguntas. Y, en caso negativo, plantear una propuesta de investigación de campo en los próximos años que permita proporcionar las evidencias necesarias para abordar una revisión exitosa dentro de unos años.

Preguntas de investigación

En esta revisión nos planteamos los siguientes objetivos:

1. Identificar qué características tiene cada uno de los tipos de escala y cómo se diferencian unas de otras
2. Estimar el grado en que son usadas, unas u otras, por las organizaciones en la evaluación de competencias de sus empleados. Y si el grado de uso depende del sector, antigüedad, número de empleados u otras variables.
3. Resumir las ventajas e inconvenientes que presentan cada una de ellas en general y a la hora de evaluar una competencia transversal como la innovación en las personas.
4. Proponer cuál de ellas sería más adecuada, atendiendo a los objetivos específicos de evaluación de la competencia de innovación en las personas (evaluación del desempeño, promoción de trabajadores, selección de personal, etc.) y el modo de administración (self, peer, expert assessment)

Metodología

Criterios de inclusión y exclusión de trabajos en la revisión

Criterios de inclusión:

1. Artículos científicos publicados en revistas con proceso de revisión por pares, en inglés, castellano, alemán, italiano o francés.
2. Indexados en Scopus, Web Of Science o google Scholar
3. Con fecha de publicación entre 1965 y julio de 2015
4. Estudios teóricos, revisiones de literatura o estudios de campo/experimentos/cuasi-experimentos que definan las características de los diferentes tipos de escalas o que proporcionen evidencias del grado de uso o las ventajas o inconvenientes de cada una de las escalas cuando se aplican a evaluar la competencias de las personas que trabajan en una organización
5. Publicados en el revistas del área de Bussiness Management o psychology, especialmente en el campo de gestión de recursos humanos (HRM) en empresas (reclutamiento, evaluación del desempeño o formación)

Criterios de exclusión:

1. Observación de comportamientos patológicos o enfermedades psicológicas o educación especial (autismo, trastornos bipolares, esquizofrenia, etc…)
2. Áreas diferentes de Social Science (por ejemplo: psiquiatría, educación especial…)
3. Competencias agregadas a nivel de grupo de personas o de una organización

Estrategias de búsqueda automáticas para la identificación de estudios relevantes (búsquedas realizadas en junio 2015)

Scopus	Resultados
(TITLE-ABS-KEY ("performance assessment" OR "performance appraisal" OR "human source*" OR recruitment OR training) AND SUBJAREA (mult OR arts OR busi OR deci OR econ O EA (mult OR arts OR busi OR deci OR econ OR psyc OR soci)) AND ((TITLE-ABS-KEY ((behavi* OR rating) AND scale*) OR TITLE-ABS-KEY ("pair* Comparison*")) AND SUBJAREA (mult OR arts OR busi OR deci OR econ OR psyc OR soci)) AND (LIMIT-TO (SUBJAREA , "PSYC") OR LIMIT-TO (SUBJAREA , "BUSI")) AND (EXCLUDE (SUBJAREA , "MEDI") OR EXCLUDE (SUBJAREA , "NEUR") OR EXCLUDE (SUBJAREA , "ARTS") OR EXCLUDE (SUBJAREA , "BIOC") OR EXCLUDE (SUBJAREA , "NURS") OR EXCLUDE (SUBJAREA , "HEAL") OR EXCLUDE (SUBJAREA , "PHAR") OR EXCLUDE (SUBJAREA , "ECON") OR EXCLUDE (SUBJAREA , "COMP") OR EXCLUDE (SUBJAREA , "COMP") OR EXCLUDE (SUBJAREA , "ENGI") OR EXCLUDE (SUBJAREA , "DECI") OR EXCLUDE (SUBJAREA , "AGRI") OR EXCLUDE (SUBJAREA , "MATH") OR EXCLUDE (SUBJAREA , "ENVI") OR EXCLUDE (SUBJAREA , "ENER") OR EXCLUDE (SUBJAREA , "EART") OR EXCLUDE (SUBJAREA , "MATE") OR EXTE") OR EXCLUDE (SUBJAREA , "CENG") OR EXCLUDE (SUBJAREA , "IMMU") OR EXCLUDE (SUBJAREA , "CHEM"))	1036
WOS	Resultados
TOPIC: (("performance assessment" OR "performance appraisal" OR "human re-source*" OR recruitment OR training)) AND TOPIC: (((behavi* OR rating) AND scale*) OR "pair* Comparison*") Refined by: [excluding] WEB OF SCIENCE CATEGORIES: (PSYCHIATRY OR LINGUISTICS OR PSYCHOLOGY CLINICAL OR PUBLIC ENVIRONMENTAL OCCUPATIONAL HEALTH OR SOCIAL SCIENCES BIOMEDICAL OR REHABILITATION OR TRANSPORTATION OR HEALTH CARE SCIENCES SERVICES OR CLINICAL NEUROLOGY OR NURSING OR MEDICINE GENERAL INTERNAL OR PSYCHOLOGY DEVELOPMENTAL OR GERIATRICS GERONTOLOGY OR EDUCATION EDUCATIONAL RESEARCH OR NEUROSCIENCES OR PEDIATRICS OR GERONTOLOGY OR HEALTH POLICY SERVICES OR EDUCATION SCIENTIFIC DISCIPLINES OR SPORT SCIENCES OR PSYCHOLOGY EDUCATIONAL OR SUBSTANCE ABUSE OR PSYCHOLOGY EXPERIMENTAL OR MEDICINE RESEARCH EXPERIMENTAL OR SOCIAL SCIENCES INTERDISCIPLINARY OR EDUCATION SPECIAL OR PHARMACOLOGY PHARMACY OR FAMILY STUDIES OR ONCOLOGY OR LANGUAGE LINGUISTICS OR ERGONOMICS OR ENVIRONMENTAL STUDIES OR SOCIAL WORK) AND RESEARCH AREAS: (PSYCHOLOGY OR BUSINESS ECONOMICS) Indexes=SSCI Timespan=1965-2015	455
Google Scholar en full text	Resultados
https://scholar.google.es/scholar?q=behavior+anchor+scale+%22performance+appraisal%22+BARS&btnG=&hl=es&as_sdt=0%2C5&as_ylo=1965&as_yhi=2015 (Es una buena estrategia, salen pocos falsos positivos)	2060 (Stop 320)
https://scholar.google.es/scholar?hl=es&as_sdt=0,5&q=BARS+%22behaviorally+anchored+scale%22	131
https://scholar.google.es/scholar?as_q=&as_epq=behavior+observation+scale&as_oq=&as_eq=&as_occt=any&as_sauthors=&as_publication=&as_ylo=1965&as_yhi=2015&btnG=&hl=es&as_sdt=0%2C5 (la mayoría de los resultados de psiquiatría o medicina. Ningún resultado a retener en los 80 primeros)	703 (Stop 80)
https://scholar.google.es/scholar?as_q=human+resource+management&as_epq=behavior+observation+scale&as_oq=&as_eq=&as_occt=any&as_sauthors=&as_publication=&as_ylo=1965&as_yhi=2015&hl=es&as_sdt=0%2C5 (muchos falsos postivos)	290 (Stop 120)
https://scholar.google.es/scholar?q=+%22behavioral+observation+scale%22+human+resource+management&btnG=&hl=es&as_sdt=0%2C5&as_ylo=1965&as_yhi=2015	254 (Stop 140)
https://scholar.google.es/scholar?q=%22performance+appraisal%22+%22paired+comparison%22&btnG=&hl=es&as_sdt=0%2C5&as_ylo=1965&as_yhi=2015	583 (stop 280)
https://scholar.google.es/scholar?as_q=%22rating+scale%22&as_epq=performance+appraisal&as_oq=employee+worker&as_eq=&as_occt=any&as_sauthors=&as_publication=&as_ylo=1965&as_yhi=2015&hl=es&as_sdt=0%2C5	4540 (stop 500)

Método usado para filtrar los estudios

Debido al elevado número de referencias de google scholar y las limitaciones para realizar filtrados automáticos más restrictivos, uno de los autores (JAM) ha hecho un filtro previo en las referencias obtenidas en Google Scholar, para ello:

1. Se ordenan los resultados de cada una de las búsquedas por relevancia mostrando 20 referencias por página.
2. Se descargan todas las referencias de la página que, por título y resumen, cumplan los criterios de la revisión. Cuando en una página no haya ninguna referencia útil se ignoran todas las demás de esa búsqueda y se registra el número de referencia en el que se ha detenido el proceso.

La lista de referencias únicas se obtendrá tras eliminar los duplicados en los resultados de las bases de datos consultadas.

Los resultados de las búsquedas serán revisados, de manera independiente por tres autores de la revisión (LRB, JAM, LAH). En primer lugar, se filtrarán por título y resumen. Si con el título y resumen queda claro que el trabajo no encaja en los objetivos (criterios de inclusión y exclusión), será excluido. En los casos que parezca claro o dudoso, se conseguirán los textos completos y se procederá a su filtrado en el proceso de codificación y extracción de la información. Para ello, la primera fase será en común entre dos de los autores (LRB, JAM) (bien en reunión presencial o por video conferencia):

1. Se ordenarán las referencias alfabéticamente por apellido autor (eso evita que haya un sesgo por año de publicación o por "relevancia")
2. Éste será el orden o identificador (ID) de las referencias para todo lo que sigue
3. Se empezará por la primera referencia y, en común, se decide en base a título y resumen, si se retiene o excluye la referencia, verbalizando el motivo por el que se retiene o se rechaza
    a. Libro de códigos para esta fase:
        i. Dudoso
            1. Definición: referencias que no está claro si cumplen los criterios de inclusión o si se ven afectados por los criterios de exclusión
            2. Cuándo usar: cuando se carece de resumen del artículo y en el título no queda claro si encajan en el objetivo de la revisión. Siempre que dudemos mantendremos la referencia y la analizaremos con el texto completo
            3. Cuándo no usar: si está claro que se cumple alguno de los criterios de exclusión
            4. Ejemplo: Ansari, M. A., & Baumgartel, H. (1981). The Critical Incident Technique: Description and Current Uses. Journal of Social & Economic. Studies, 9(2),
        ii. Seleccionado, estudio individual
            1. Definición: referencias que en el título y resumen queda claro que interesan para la revisión y se trata de una investigación de campo
            2. Cuándo usar: cuando se cumplan los criterios de inclusión y no afecten los de exclusión
            3. Cuándo no usar: si se trata de una revisión o meta-análisis

4. Ejemplo: Benson, P. G., Buckley, M. R., & Hall, S. (1988). The Impact Of Rating Scale Format On Rater Accuracy. Journal of Management, 14(3), 415

iii. Seleccionado, revisión o meta-análisis
1. Definición: referencias que en el título y resumen queda claro que interesan para la revisión y se trata de una revisión o meta-análisis
2. Cuándo usar: cuando se indica que el objetivo de la investigación es resumir los resultados de varios estudios y para analizar alguna de las variables clave de nuestra revisión
3. Cuándo no usar: cuando no está claro si es una revisión de literatura (en este caso se etiquetarían como estudio individual)
4. Ejemplo: Aggarwal, A., & Thakur, G. S. M. (2013). Techniques of Performance Appraisal-A Review. International Journal of Engineering and Advanced Technology

iv. No seleccionado, no encaja en los objetivos
1. Cuándo usar: cuando no se cumplen los criterios de inclusión
2. Ejemplo: Albion, M. J., Fernie, K. M., & Burton, L. J. (2005). Individual differences in age and self-efficacy in the unemployed. AUSTRALIAN JOURNAL OF PSYCHOLOGY, 57(1), 1119. http://doi.org/10.1080/00049530412331283417

v. No seleccionado, competencias de grupos u organizaciones
1. Definición: cumple algunos criterios de inclusión y trata sobre cosas que pueden interesar, pero mide competencias de grupos u organizaciones y no de personas.
2. Cuándo usar: se cumplen los criterios de inclusión pero les afecta el criterio de exclusión 3
3. Ejemplo: Laplante, N., & Harrisson, D. (2008). Conditions for the development of trust between managers and union representatives in a context of innovation [Les conditions de la confiance entre gestionnaires et représentants syndicaux dans un contexte d'innovations]. Relations Industrielles, 63(1), 85–107+161

vi. No seleccionado, comportamiento patológico o educación especial
1. Definición: cumple algunos criterios de inclusión y trata sobre cosas que pueden interesar, como fiabilidad o tipos de escala... pero para entornos médicos o de educación y no de gestión de recursos humanos.
2. Cuándo usar: se cumplen los criterios de inclusión pero les afecta los criterios de exclusión 1 ó 2
3. Ejemplo: Allen, R. A., Robins, D. L., & Decker, S. L. (2008). Autism spectrum disorders: Neurobiology and current assessment practices. Psychology in the Schools, 45(10)

4. Se repetirá el proceso hasta encontrar 15 artículos a retener o llegar a la referencia nº100
5. Se actualizarán los criterios de inclusión o exclusión si fuese necesario

A continuación se trabajará de manera independiente con el resto de referencias:

1. Uno de los autores (JAM) revisará todas las referencias. Otro (LAH) revisará dos tercios de las referencias El tercer autor (LRB) revisará el tercio restante.
2. Después de las 100 primeras referencias revisadas por separado se comparará el grado de acuerdo entre los evaluadores (LRB y JAM) y se resolverán la discrepancias fijando o modificando los criterios de inclusión/exclusión en una reunión de los tres autores (LRB, JAM, LAH)
3. Se continuará hasta el final de las referencias
4. Se comprobará el grado de acuerdo entre los tres evaluadores.
5. Se dibujará el diagrama de flujo del proceso de selección (Moher, Liberati, Tetzlaff, Altman, & The, 2009)

Una vez seleccionado el conjunto de artículos incluidos, se revisarán las referencias citadas para extraer nuevas referencias seleccionadas con el procedimiento "bola de nieve" (Bryman & Bell, 2011) a partir del título de los artículos en los que aparezca Scale, BARS, BOS, paired o alguno de sus sinónimos. Para hacer el proceso más sencillo nos apoyaremos en la capacidad de búsqueda de los lectores PDF. Uno de los autores (JAM) buscará de manera automática en todos los documentos, los otros autores (LRB y LAH) repasarán manualmente y de manera independiente, la lista de referencias de 10 artículos (revisiones) y se compararán los resultados con la búsqueda automática.

Codificación y extracción de la información de los textos completos

La extracción de la información se hará de manera coordinada por parte de los tres autores, usando ATLAS-Ti como herramienta de apoyo:

1. Libro de códigos para etiquetar la extracción de la información:
    a. País donde se toman los datos de campo (si no hay datos de campo de trabajadores evaluados con alguna escalas, no se recoge este código)
    b. Tipo de organización donde trabajan los evaluados (actividad, sector, tamaño, …)
    c. Descripción demográfica de la muestra de los evaluados (sexo, edad, nivel de estudios,..)
    d. Competencia evaluada con la escala
    e. Tipo de escala utilizada (BARS, BOS, Rating Scale…)
    f. Característica de la escala (definición o propiedades de ese tipo de escala que la diferencia de otros tipos de escalas)
    g. Resultados psicométricos de la escalas (grado de acuerdo entre evaluadores…)
    h. Ventajas de la escala
    i. Inconvenientes de la escala
    j. Limitaciones de la investigación o investigación futura
    k. Utilidad de investigar el tipo de escala para académicos o mandos de empresa
2. Utilizando el listado de códigos común para los tres autores, se codificarán de manera conjunta los 10 primeros artículos de la lista, en una sesión de trabajo común (LRB, JAM, LAH):
3. Si alguna de las referencias es excluida en esta fase se procederá a etiquetar el motivo de exclusión
4. Se procederá a codificar de manera independiente (LRB, JAM, LAH) las 10 referencias siguientes

5. Se comprobará el grado de acuerdo entre los autores y se procederá a depurar el libro de códigos aclarando los significados su fuese necesario
6. Si el grado de acuerdo es elevado se procederá a repartir el resto de referencias entre los tres autores, de modo que cada uno de ellos extraiga la información del 40% de las referencias restantes (produciéndose un 20% de solape)
7. Se comprobará el grado de acuerdo en la extracción de información del 20% de referencias solapadas. Si es elevado se cerrará esta etapa, en caso contrario, se resolverán dudas y los autores realizarán la extracción de otro conjunto de 40% de referencias diferentes del inicial
8. Se comprobará el grado de acuerdo final entre autores por medio del coding Analysis Toolkit (CAT) (http://www.qdap.pitt.edu/cat.htm) (Hayes & Krippendorff, 2007) y se resolverán las diferencias por consenso en una reunión de grupo.

Procedimiento de análisis cualitativo de los resultados

Se seguirá el procedimiento de codificación por etapas *(Initial, Focused, Theoretical*) (Charmaz, 2006) con la ayuda del programa Atlas.ti (Friese, 2012).

Plan de trabajo

Literature search	June 2015
Filtering and retrieval of studies	July-September 2015
Pilot testing of study codes	October 2015
Extraction of data from studies and double coding	November 2015
Analysis	December 2015-January 2016
Preparation of report	February – May 2016
Submission of final report	June 2016
Response to review comments	Upon receipt

Acknowledgments

Este trabajo ha sido realizado con la financiación de la Unión Europea ["FINCODA" proyecto 554493-EPP-1-2014-1-FI-EPPKA2-KA] (The European Commission support for the production of this publication does not constitute an endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein) y PIME/2014/A/013/A- Barómetro INCODE: Evaluación de Competencias de lnnovación en la Empresa y en la Universidad.

References

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & van der Vleuten, C. P. M. (2006). The wheel of competency assessment: Presenting quality criteria for competency assessment programs. *Studies in Educational Evaluation, 32*(2), 153-170. doi:http://dx.doi.org/10.1016/j.stueduc.2006.04.006

Boyatzis, R. E. (2008). Competencies in the 21st century. *Journal of management development, 27*(1), 5-12. doi:doi:10.1108/02621710810840730

Bretz, R. D., Milkovich, G. T., & Read, W. (1992). THE CURRENT STATE OF PERFORMANCE-APPRAISAL RESEARCH AND PRACTICE - CONCERNS, DIRECTIONS, AND IMPLICATIONS. *Journal of Management, 18*(2), 321-352. doi:10.1177/014920639201800206

Bryman, A., & Bell, E. (2011). *Business Research Methods*. USA: Oxford University Press.

Charmaz, K. (2006). *Constructing grounded theory. A practical guide through qualitative analysis*. London: SAGE.

Dolan, S. L., Valle Cabrera, R., Jackson, S. E., & Schuler, R. S. (2007). *La gestión de los recursos humanos. Cómo atraer, retener y desarrollar con éxito el capital humano en tiempos de transformación*. Madrid: Mcgraw-Hill.

Dowdy, E., Twyford, J., & Sharkey, J. D. (2013). Methods of Assessing Behavior: Observations and Rating Scales. In D. H. Saklofske, V. L. Schwean, & C. R. Reynolds (Eds.), *The Oxford Handbook of Child Psychological Assessment* (pp. 623-650): Oxford University Press.

DuVernet, A. M., Dierdorff, E. C., & Wilson, M. A. (2015). Exploring Factors That Influence Work Analysis Data: A Meta-Analysis of Design Choices, Purposes, and Organizational Context. *Journal of Applied Psychology*. doi:10.1037/a0039084

Friese, S. (2012). *Qualitative Data Analysis with ATLAS.ti*. London: SAGE Publications Ltd.

Groth-Marnat, G. (2003). *Handbook of psichological assessment*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Hartig, J. (2008). Psychometric models for the assessment of competencies. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 69-90). Guttingen: Hogrefe & Huber Publishers.

Hatzinger, R., & Dittrich, R. (2012). Prefmod: An R package for modeling preferences based on paired comparisons, rankings, or ratings. *Journal of Statistical Software, 48*(10). Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84863314424&partnerID=40&md5=fe41f891ec6e3f17584078fe2d73b5ef

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*, 77-89.

Heidemeier, H., & Moser, K. (2009). Self-Other Agreement in Job Performance Ratings: A Meta-Analytic Test of a Process Model. *Journal of Applied Psychology, 94*(2), 353-370. doi:10.1037/0021-9010.94.2.353

Hoyt, W. T., & Kerns, M. D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4*(4), 403-424. doi:10.1037//1082-989x.4.4.403

Lenburg, C. (1999). The Framework, Concepts and Methods of the Competency Outcomes and Performance Assessment (COPA) Model

Online Journal of Issues in Nursing, 4(2).

Lohmann, A., & Prumper, J. (2006). Questionnaire for direct participation in the office (FdP-B) - results concerning its reliability and validity. *Zeitschrift fur Arbeits-und Organisationspsychologie, 50*(3), 119-134. Retrieved from <Go to ISI>://000239034600001

Marin-Garcia, J. A., Aragonés Belgran, P., & Melón, G. (2014). Intra-rater and inter-rater consistency of pair wise comparison in evaluating the innovation competency for university students. *Working Papers on Operations Management, 5*(2), 24-46. doi:http://dx.doi.org/10.4995/wpom.v5i2.3220

Marin-Garcia, J. A., Garcia-Sabater, J. J., Maheut, J., Valero-Herrero, M., & Andres-Romano, C. (2012). *Gestión de recursos humanos para ingenieros de la rama industrial*. Harlow: Pearson Education.

Marin-Garcia, J. A., Perez-Peñalver, M. J., & Watts, F. (2013). How to assess innovation competence in services: The case of university students. *Direccion y Organizacion*(50), 48-62. Retrieved from http://www.revistadyo.com/index.php/dyo/article/viewFile/431/451

Marin-Garcia, J. A., Ramirez Bayarri, L., & Andreu Andres, M. A. (2015). *Comparación de los métodos de escalas y frecuencia de comportamiento para valorar la competencia de innovación. El punto de vista de alumnos y profesor en el caso de una asignatura de máster.* Paper presented at the Congreso In-Red 2015-Universitat Politècnica de València.

Marin-Garcia, J. A., & Santandreu-Mascarell, C. (2015). What do we know about rubrics used in higher education? *Intangible Capital, 11*(1), 118-145. doi:http://dx.doi.org/10.3926/ic.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The, P. G. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med, 6*(7), e1000097. doi:10.1371/journal.pmed.1000097

Moore, D. R., Cheng, M. I., & Dainty, A. R. J. (2002). Competence, competency and competencies: performance assessment in organisations. *Work Study, 51*(6), 314-319. doi:doi:10.1108/00438020210441876

Rowe, C. (1995). Clarifying the use of competence and competency models in recruitment, assessment and staff development. *Industrial and Commercial Training, 27*(11), 12-17. doi:doi:10.1108/00197859510100257

Saaty, T. L. (1980). *The Analytic Hierarchy Process*. New York: McGraw-Hill.

Saaty, T. L. (1996). *Decision Making with Dependence and Feedback: The Analytic Network Process*. Pittsburgh, PA.: RWS Publication.

Tziner, A., & Kopelman, R. E. (2002). Is there a preferred performance rating format? A non-psychometric perspective. *Applied Psychology, 51*(3), 479-503. doi:10.1111/1464-0597.00104

Voskuijl, O. F., & Van Sliedregt, T. (2002). Determinants of interrater reliability of job analysis: A meta-analysis. *European Journal of Psychological Assessment, 18*(1), 52-62. doi:10.1027//1015-5759.18.1.52