REDU Revista de docencia Universitaria

Revista de Docencia Universitaria,
Vol.10 (Número especial, 2012), 229 - 241
ISSN:1887-4592

Fecha de entrada: 30-07-2012
*Fecha de aceptación:*17-09-2012

Variabilidad interobservador. Analizando algunas fuentes de error: heurísticas y categorizaciones

Interobserver variability. Analyzing some sources of error related to heuristics and categorization.

Alberto Alves de Lima
Instituto Cardiovascular de Buenos Aires. Argentina

Resumen

La educación médica ha intentado aprovechar la capacidad de observación humana para evaluar el desempeño de sus profesionales y estudiantes. Las herramientas de evaluación basadas en observadores han demostrado por un lado tener fortalezas en cuanto a su validez, fácil aplicabilidad, bajo costo y buena aceptación por parte de docentes y alumnos pero por otro una marcada debilidad en cuanto a su nivel de reproducibilidad. Son bien conocidas las diferencias de desempeño entre los estudiantes atribuible a la especificidad del contexto o del caso pero uno de los principales problemas es la variabilidad interobservador. Analizando la dificultad de este problema utilizando los marcos de referencia tradicionales, es necesario buscar otros puntos para analizar como las personas toman decisiones o como perciben a otras personas. Con respecto a la toma de decisiones es importante considerar que los evaluadores utilizan principios heurísticos de representatividad, de disponibilidad y anclajes que reducen las tareas mentales en el cálculo de probabilidades complejas en procesos simples, pero muchas veces promueven errores severos. Con respecto al acto de percibir a otra persona, podría describirse como una tarea de categorización analizable bajo 3 conceptos: a) formación de la impresión como una construcción de modelos de personas, b) formación de la impresión como la formación de un proceso de categorización nominal y c) formación de la impresión como un proceso de categorización multidimensional. Es posible que exista una discordancia entre forma en que los humanos perciben la información y la forma en que esta información es documentada. Hay mucho que aprender e investigar en relación al comportamiento de los observadores. Por el momento es necesario introducirnos en sus mentes, con el fin de comprender qué es específicamente lo que consideran importante al momento de tomar decisiones sobre el desempeño de un alumno y desde un punto de vista práctico, se impone la necesidad de realizar múltiples-mini observaciones con el fin de amortiguar estos sesgos.

***Palabras clave**:* observación directa, variabilidad interobservador, evaluación basada en el trabajo, categorización nominal, heurísticas, sesgos, error

Abstract

Medical education has tried to use human observation to evaluate students and professional performance. Assessment tools based on direct observation have shown, on one side, to have good results in terms of validity, feasibility, costs and acceptable satisfaction rates from teachers and students; but on the other side, they have as well shown a marked weakness in their reliability levels. The differences in performance among students related to case or context specificity are well known, but one of the main problems is the interrater variability. It is not easy to understand this problem using traditional frameworks, it is necessary to look to a different alternative to analyze how people make decisions or how they perceive others. Regarding to decision-making it is important to bear in mind that evaluators use heuristic principles of representativeness, availability and anchors to reduce complex mental process, but these principles often provoke errors. The act of perceiving another person could be described as an analyzable categorization task under 3 concepts: a) impression formation as a construction of person models, b) impression formation as a nominal categorization process and c) impression formation as a multidimensional categorization process. There is a mismatch between how humans perceive information and how this information is documented. There is too much to learn and research regarding the observers' behavior. At this moment it is necessary to gain access to their minds, to understand what specifically they consider important when they are making decisions about a student's performance, and from a practical point of view, we need to apply multiple-mini observations in order to reduce these biases.

Key words: Direct observation, interrater variability, nominal categorization, heuristics, biases, error.

Introducción.

La educación médica ha intentado aprovechar la capacidad de observación humana para evaluar el desempeño de sus profesionales y estudiantes.

Así, la evaluación basada en observadores es muy utilizada porque permite evaluar el desempeño profesional en la "acción" es decir, mientras éstos desarrollan actividades complejas que corresponden a altos niveles de la pirámide de Millar (Miller, 1990).

Ejemplos comunes incluyen el OSCE, y todas aquellas herramientas de evaluación basada en el trabajo como el Mini-CEX (Turner & Dankoski, 2008; Norcini & Fortuna, 2003; Alves de Lima et al., 2011).

Las herramientas de evaluación basadas en observadores han demostrado por un lado tener fortalezas en cuanto a su validez, fácil aplicabilidad, bajo costo y buena aceptación por parte de docente y alumnos pero por otro, una marcada debilidad en cuanto a su nivel de reproducibilidad.

Son bien conocidas las diferencias de desempeño entre los estudiantes atribuible a la especificidad del contexto o del caso. La especificidad de contexto o de caso es definida como, la observación en que el desempeño de un estudiante de una determinada habilidad o comportamiento frente a un problema particular y en un contexto particular, es pobremente predictivo del mismo desempeño o comportamiento un problema diferente o en un contexto diferente (Eva, 2003). Norman (1995) le pidió a 30 clínicos de distintos niveles de expertise que observen y evalúen una serie de 10 problemas clínicos en cardiología y reumatología. 2 casos de cada especialidad eran exactamente iguales pero lo único que cambiaba era el actor

que hacía de paciente. A pesar que el contenido de los caso eran exactamente iguales en los 2 problemas, el solo hecho de cambiar al actor, generó una correlación promedio entre los scores de desempeño de solo 0.28 (Norman, Tugwell, Feighter, Muzzin & Jacoby, 1995). En líneas generales la especificidad de contexto jerarquiza el valor de las circunstancias en relación a los atributos o rasgos emocionales de la persona. No significa necesariamente que los rasgos de las personas sean necesariamente estables sino más bien son dependientes del contexto y la situación. En un ejemplo clásico, Darley y Batson (1973) le pidieron a un grupo de estudiantes del seminario de Princeton (admitidos en el seminario entre otras cosa por sus valores de caridad) que atraviesen un campo hasta llegar a un lugar a dar una charla sobre la parábola del buen samaritano. A modo de recordatorio en esa parábola un sacerdote judío y un levita pasaban por delante de un judío apaleado sangrante. Ambos dejaron de darle asistencia hasta que llego un samaritano que se detuvo, lo acobijó y pagó los gastos de la atención. Cabe destacar que los samaritanos eran enemigos de los judíos. En este experimento los seminaristas se toparon con un hombre (actor que los participantes no conocían ni sabían de su presencia) en dificultades, que se desplomó frente a un portal. De los participantes a los cuales se les había dicho que estaban llegando tarde a la charla, se detuvo sólo el 10%, mientras a los que no se les dijo que llegaban tarde a la charla se detuvo el 63%. Este ejemplo ilustra claramente la influencia de "las circunstancias" en relación a los rasgos de personalidad de los individuos. Existe correlación significativa en la medición de la personalidad a través de múltiples situaciones, pero la correlación es baja (< 0.3) en relación al impacto de la situación.

Variabilidad interobservador.

Uno de los principales problemas es la variabilidad interobservador que ocurre cuando dos observadores observan el mismo desempeño (Downing, 2004, Downing, 2005).

En un ejemplo emblemático, 19 de 20 estaciones de OSCE tenían de 1 a 8 discrepancias en donde al menos un observador hizo un comentario evaluativo positivo en relación a presencia o ausencia en un comportamiento especifico observable, mientras otro observador hizo un comentario evaluativo negativo de ese mismo comportamiento observado (Mazor et al., 2007).

La razón de la variabilidad de las evaluaciones interobservador observando en mismo desempeño no es claro y abre un gran debate orientado a solucionarlo

Con este tema sobre la mesa los investigadores han diseñado escalas, formularios, formatos y sistemas con el fin de disminuir esta subjetividad pero las soluciones han tenido resultados limitados. Cook (2008) evaluó en forma randomizada el impacto de un programa de entrenamiento para evaluadores y no observó mejoría en la reproducibilidad interobservador ni en la precisión de los puntajes otorgados. Kogan (2010) observó que las habilidades clínicas personales de los observadores están asociadas con los puntajes otorgados a los estudiantes. A mejores habilidades clínicas personales del evaluador más exigente es al momento de poner un resultado. Govaerts (2011) observó diferencias entre observadores expertos y novatos en

relación al tiempo necesario para la representación del desempeño del estudiante y que a su vez eso depende de la complejidad de la habilidad observada.

La pobre mejoría de las mediciones a través del entrenamiento ha provocado a algunos investigadores a sospechar que los evaluadores médicos son impermeables al entrenamiento, sugiriendo que algunos son consistentemente buenos evaluadores y otros no. El primero no necesita entrenamiento y el segundo no mejora con entrenamiento (Newble & Hoare 1980).

Analizando la dificultad de este problema utilizando los marcos de referencia tradicionales, es necesario buscar otros puntos para analizar como las personas toman decisiones o como perciben a otras personas. Con respecto a la toma de decisiones es importante considerar que los evaluadores utilizan principios heurísticos y con respecto al acto de percibir a otra persona es posible que exista una discordancia entre forma en que los humanos perciben la información y la forma en que esta información es documentada.

Heurísticas.

En evaluación, los que deben tomar las decisiones tienen a su alcance reportes de desempeño, resultados estadísticos, análisis de contextos estandarizados y los resultados individuales de desempeño de diferentes estudiantes en los cuales deben tomar decisiones de aprobar o no aprobar. Deben estar confiados que están tomado la decisión correcta. La confianza de las personas en cualquier decisión va a depender de experiencias personales y de la forma en que la información es recogida y ponderada. La forma en que los evaluadores combinan y ponderan la información es un tema poco claro.

Ese proceso de toma de decisiones esta influencia por heurísticas y así propensas a sesgos (Tversky & Kahnman, 1974). Muchas decisiones están basadas en creencias en relación a la probabilidad de la ocurrencia de eventos inciertos; ej. ¿Cuál es la probabilidad que el estudiante A tenga un comportamiento X frente a la circunstancia J?. ¿Qué determina esta creencia?, ¿Cómo los evaluadores analizan las posibilidades de que esos eventos inciertos ocurran?. Los evaluadores podrían utilizan principios heurísticos. Los principios heurísticos que reducen las tareas mentales en el cálculo probabilidades complejas en procesos simples. Estos procesos son útiles pero muchas veces promueven errores severos y sistemáticos. El análisis de la heurística y los sesgos comenzaron con el trabajo del premio novel de economía Laureate Daniel Kahneman y del profesor Amos Tversky. Disconformes con las discrepancias de la economía clásica en explicar la toma de decisiones humanas, Kahneman y Tversky desarrollaron una disciplina hoy conocida como “comportamiento económico”. En contraposición a los modelos clásicos en que se describía al ser humano como un maximizador racional de decisiones costo-efectiva, estos autores desarrollaron un esquema simple del comportamiento humano basado en toma de decisiones bajo situaciones de incertidumbre, riesgo y ambigüedad. Ellos proponen que frente a estas situaciones en el cual deberíamos manejar una importante carga de información, los seres humanos reducen su complejidad a través de uso de heurísticas. En el momento de la simplificación del proceso mental, aplicamos sesgos cognitivos. Cabe destacar que el uso de heurísticas no produce errores siempre sino nos hace vulnerables a inducir

error. Las heurísticas incluyen: representatividad, disponibilidad y anclajes (Tversky & Kahnman, 1974).

La representatividad es una heurística que usan las personas para evaluar la probabilidad que un evento, persona u objeto corresponda a otra categoría de eventos personas u objetos mayor, si demuestra o se perciben característica de esa clase o categoría. Para ilustrar un proceso por representatividad consideremos el siguiente ejemplo: José es un estudiante retraído, calmo con algunos problemas de comunicación con sus compañeros de curso. En general se presenta desalineado y es impuntual. Cómo evalúan las personas la probabilidad que José sea un estudiante que corresponde al percentilos de desempeño inferior, medio o superior. En la heurística de representatividad, en este caso que José corresponda al percentilo de desempeño inferior, es evaluado por el grado en que José es representativo por o similar al estereotipo de ese percentilo. De acuerdo con estos autores los factores que influencia a la representatividad son a) la indiferencia a la probabilidad previa, b) al tamaño de la muestra, c) el concepto erróneo de chance, d) la validez ilusoria y e) el concepto erróneo de regresión a la media.

a) La indiferencia a la probabilidad previa podría explicarse usando el ejemplo anterior. El hecho que haya gran número de estudiantes del percentilo superior es el curso de José debe entrar en un análisis de estimación de probabilidad y esto influenciar lo que Juan nos representa. En un experimento a un grupo de personas se les mostraba una breve descripción de una serie de personas (100) con características que correspondían a ingenieros o abogados. Se les pedía a los participantes que evalúen para cada descripción la probabilidad que corresponda a un ingeniero o un abogado. En un grupo de se les dijo que del total 30 eran abogados y 70 ingenieros y en otro grupo al revés. La probabilidad de la primera secuencia debía ser mayor para los ingenieros y en la segunda para los abogados. En una clara violación a la reglas del análisis bayesiano los sujetos en ambos grupos realizaron juicios con la misma probabilidad. Aparentemente los participantes evaluaron la probabilidad que la descripción correspondía a un ingeniero o a un abogado en base al estereotipo haciendo escasa o ninguna referencia a la probabilidad previa (Tversky & Kahneman 1973).

b) La indiferencia al tamaño de la muestra es otro de los factores que puede influenciar la representatividad. En algunas oportunidades podemos tomar decisiones individuales en similitud a los parámetros promedio de la población que estos individuos representan. La probabilidad que la que un grupo de 10 egresados de la escuela de medicina de la universidad de Harvard sea muy buena está vinculada a la similitud de la muestra con el parámetro promedio de la población aplicada a esta muestra. Así, la similitud estadística de la muestra con la población considerada a través de la representatividad es independiente del tamaño de la muestra.

c) Con respecta al concepto erróneo de chance, supongamos que estamos observando una ruleta en el casino. Cuál de las siguientes frecuencias de rojos y negros sería la más probable?: RNRNRNR O RNRRNNR o RNNNNN. La respuesta es todas las chances tienen la misma probabilidad aunque la mayoría de las personas considerarían a la secuencia RNRNRNR como la más probable. Esta

sería la secuencia más popular porque las personas en general esperan que las chances se equilibren (50% R y 50% N) aunque matemáticamente cada secuencia tiene una chance de 1.56%. La implicancia es que nosotros inconscientemente evaluamos la probabilidad de los eventos futuros en basado en le representatividad de la secuencia no en probabilidades (Tversky & Kahneman 1973). Ahora consideremos en siguiente ejemplo: cuál de las siguientes situaciones es más probable: 1) José tendrá una crisis vocacional durante 2013 y 2) José tendrá problemas emocionales por conflictos con su pareja y tendrá una crisis vocacional en 2013. Si elige la opción 2, entonces Ud. está equivocado. Cuanto más específica es la descripción la probabilidad del evento es menor. Dos eventos ocurriendo en el mismo año son menos probables que un solo evento. La personas tienen la tendencia a considerar que un evento tendrá más probabilidad cuanto más información específica es analizada (Tversky & Kahneman 1973).

d) Con respecto a la validez ilusoria un ejemplo representa a este problema podría ser el valor que los evaluadores pueden darle a determinados comentarios o anécdotas sobre el estudiante sobre todo si el estudiante se correlaciona con el estereotipo de ese comentario, inclusive si esa descripción es escasa o desactualizada. La confianza no garantizada entre la información que uno recibe y la conclusión que alcanza se denomina validez ilusoria.

e) Con respecto al concepto erróneo de regresión a la media, vale la pena analizar el siguiente ejemplo. Supongamos un residente que debe realizar una punción pleural. En la primera oportunidad que tiene para realizarla su desempeño no es satisfactorio lo que le vale feedback negativo por parte del instructor. En una segunda oportunidad el desempeño mejora y recibe un feedback positivo por parte del instructor. En una tercera o cuarta oportunidad el desempeño vuelve a ser insatisfactorio. Este comportamiento podría llevar a la falsa conclusión que el feedback positivo es perjudicial para el aprendizaje. A través del concepto de regresión a la media descrito por Galton hace mas de 100 años, en procedimientos que se repiten, un buen desempeño va a ser seguido de otro no tan bueno y uno malo por otro mejor.

La disponibilidad es una heurística que aparece cuando los individuos evalúan la frecuencia de una clase o la probabilidad de un evento por la facilidad con que estas instancias u ocurrencias pueden ser traídas a la mente. Por ejemplo, uno puede dar mayor o menos magnitud al error de un residente durante un encuentro de Mini-CEX por recordar ese error en base a alguna experiencia personal o de algún conocido. Este análisis se denomina “disponibilidad”. Además los problemas prevalentes son recordadas mejor y más rápido que las menos prevalentes. Esta heurística tiene los siguientes sesgos potenciales: a) sesgos por recuperabilidad y b) desvios por imaginación.

a) Sesgos por recuperabilidad de los casos: un ejemplo de la influencia de este factor en la disponibilidad de los datos seria que la probabilidad subjetiva que un paciente sea portador de “porfiria” es mayor si la noche anterior leímos un artículo al respecto. En ese sentido la probabilidad subjetiva de un mal desempeño de un residente en una prueba, podría aumentar transitoriamente si uno viene observando malos desempeños durante las observaciones de en ese

día. Los fenómenos que ocurrieron recientemente están más disponibles que otros eventos.

b) Desvíos por imaginación: cuando nos enfrentamos a una situación nueva en la cual no tenemos ningún registro disponible en la memoria utilizamos nuestra imaginación como mecanismo subjetivo de premonición. Si frente a un determinada situación imaginamos situaciones negativas asumiremos una probabilidad baja de alcanzar el objetivo o eventualmente tomaremos recaudos exagerados.

Los anclajes y ajustes son heurísticas que ocurren cuando las personas hacen estimaciones desde un valor inicial que es utilizado para generar la respuesta final. El valor inicial puede ser sugerido en la formulación del problema. Estos ajustes producen gran variabilidad de los resultados, es decir diferentes puntos de inicio (gran expectativa) producen diferentes estimaciones que están sesgadas hacia el valor inicial. Esto se denomina fenómeno de anclaje (Tversky & Kahnman, 1974).

Impresiones de las personas desde la cognición social.

Dada las dificultades para comprender las razones de las diferencias en la reproducibilidad de los juicios de los evaluadores a través de los marcos de referencia tradicionales, Gingerich (2011) considera que podría ser de valor explorar otros marcos de referencia para comprender la manera en que las personas se representan y sacan conclusiones sobre otras. En este sentido un grupo de investigadores han alertado para considerar al proceso cognitivo de los evaluadores como un proceso cognitivo social y las implicancia que esto tiene al momento de medir un desempeño. El autor propone la necesidad de considerar a los evaluadores como procesadores de información activos que utilizan su criterio, razonamiento y estrategias de toma de decisiones para evaluar a los alumnos. Hace hincapié en la complejidad de las interacciones entre la formación de la impresión, la interpretación, la recuperación de la información a través de la memoria al momento de calificar. Además es posible que exista incongruencia entre los procesos de evaluación, los principios de medición psicométricos y la capacidad de evaluar del ser humano (Van der Vleuten & Schuwirth, 2005; Govearts, Van der Vleuten, Schuwirth & Muijtjens, 2007). Estos análisis explorados por estos autores están vinculados a la bibliografía de la formación de la impresión, un área en donde la cognición social que se focaliza en estudiar cómo los individuos toman impresiones de otros individuos en el contexto social (Gingerich, Regehrs & Eva, 2011).

Según Gingerich (2011) las impresiones se forman como parte del conocimiento de la otra persona. Son construidas con información fáctica, suposiciones y reacciones evaluativas en relación a la persona en cuestión (persona objetivo). Por otra parte los investigadores en cognición social están interesados en los procesos mentales usados por las personas para pensar en el mundo social. Analizan como la información social es codificada, guardada y recuperada desde la memoria y estructurada y representada como conocimiento. Estudian también los procesos usados para formar juicios y tomar

decisiones. Esta establecido que observadores diferentes van formarse una impresión diferente del mismo alumno inclusive cuando se les de la misma información. La máxima proporción de la varianza en evaluaciones de rasgos de personalidad no es atribuible por las diferencias percibidas entre los alumnos sino a la diferencia en la relación entre el observador, el alumno y el caso (Alves de Lima, Conde, Costabel, Corso &Van der Vleuten 2011). Margolis (2006) observo que la varianza atribuida a la exigencia/permisividad del observador contribuye más que la varianza real entre los estudiantes. Alves de Lima (2011) observó en una serie de 22 residentes que fueron evaluados frente a 3 pacientes simulados que la principal fuente de varianza fue el error general (34%) y seguida por el efecto de los evaluadores (18%).

Los hallazgos en la literatura en relación a la evaluación basada en el observador y la de la formación de la impresión, sugieren que exploraciones en este campo podrían ser de valor y colaborar a la comprensión del proceso cognitivo utilizado por los observadores en el contexto social de las herramientas de evaluación basadas en observadores

En psicología el acto de percibir a otra persona se describe como una tarea de categorización, aunque existen diferencias en la forma en que estos procesos cognitivos se cree que se representen.

Según Gingerich (2011) esta categorización puede ser conceptualizada bajo 3 conceptos: a) formación de la impresión como una construcción de modelos de personas, b) formación de la impresión como la formación de un proceso de categorización nominal y c) formación de la impresión como un proceso de categorización multidimensional

a) La formación de la impresión como construcción de modelos de persona parte del concepto que los juicios sociales son idiosincráticos y falibles bajo ciertas condiciones. Por ejemplo el humor de los evaluadores y las emociones pueden influenciar. Si el evaluado le hace recordar al evaluador a un ser querido o si el evaluador ha sido expuesto a alguna anécdota o antecedente del evaluado puede hacer que el evaluador tenga un comportamiento ambiguo consistente con esa descripción. A pesar de este comportamiento idiosincrático la formación de la impresión es habitualmente consistente a lo lago de los evaluadores. Park, De Kay y Kraus (1994) proponen que las impresiones podían agruparse en 3 patrones,(historia o personal models) más o menos definidos. En un estudio 69 participantes vieron y evaluaron el mismo video de 4 minutos. Se trataba de un alumno manteniendo una conversación con un amigo y luego con un miembro de su familia. Todos dieron una descripción escrita. Las descripciones pudieron agruparse en 3 categorías: el 69% la describió al alumno como activo, amigable y expresivo, el 15% la describió como inseguro y nervioso y el 16% como dominante brusco y ambicioso. En otras palabras las impresiones son idiosincráticas pero no infinitas y dan cierta explicación a la variabilidad interobservador.

b) En la impresión de las personas como un proceso de categorización nominal no se trata ahora de construir un texto en relación a la descripción de un comportamiento, sino la tendencia vincular a los evaluados en esquemas preexistentes. El valor de este mecanismo radica en que permiten a los

evaluadores aplicar los conocimientos preexistentes para ayudar a entender la información percibida de la persona evaluada. Si bien hay un claro peligro en la sobre generalización (como los estereotipos), con el uso de categorización los recursos cognitivos no necesitan ser usados para monitorear la categoría correspondiente a comportamiento del observado. En cambio el evaluador solo necesita señalar cualquier comportamiento incompatible con la categoría. El modelo de categorización actúa como un marco de referencia para tener posibles explicaciones frente a determinados comportamientos del evaluado en una situación dada. El contexto es un factor importante en determinar la categoría. Si aceptamos que los evaluadores categorizan a los evaluados esto puede tener gran implicancia en las evaluaciones basadas en el observador. La implicancia práctica es que esto es más parecido a una escala de evaluación nominal que a una ordinal. Las variables nominales tienen categorías pero no un orden lógico (cero verdadero o un intervalo). Las grillas de evaluación sugieren órdenes lógicos (ordinales). Si los evaluadores están juzgando el desempeño de los alumnos percibiéndolos como pertenecientes a una categoría particular, entonces como traducen esos juicios nominales en ordinales? .Esto podría ser otra fuente de error

c) En la impresión de las personas basadas en categorización por dimensiones a diferencia de la categorización nominal, existen otros autores que conceptualizan que la categorización implica un juicio en escalas dimensionales (Gingerich, Regehrs & Eva, 2011). Hay dos dimensiones más definidas en la literatura: una está basada en los rasgos deseados o no deseados en la sociedad que impactan directamente sobre los otros (incluye rasgos positivos como la honestidad y negativos tramposo) y una segunda dimensión, que tiene más variabilidad a lo largo de los estudios y se refiere a rasgos que tienden a influenciar más directamente sobre el éxito del individuo. Incluye positivos como la inteligencia y negativo como la indecisión o la ineficiencia. Se sugiere además que estas dimensiones pueden estar dicotomizados en valores alto versus valores bajos.

Conclusión.

Las herramientas de evaluación basadas en observadores han demostrado una marcada debilidad en cuanto a su nivel de reproducibilidad. Si bien la especificidad del caso ocupa un rol importante, la variabilidad inter-observador es una fuente muy significativa de error.

En este artículo hemos analizados 2 fuentes de error: los sesgos cognitivos vinculados asociados a heurísticas utilizadas para evaluar probabilidades y predecir valores y la impresión que tienen las personas sobre otras desde el punto de vista de la cognición social

Los sesgos cognitivos no son atribuibles a ilusiones ni a expresiones de deseo. Para juzgar probabilidades en forma adecuada la consistencia interna no es suficiente. Los juicios deben ser compatibles con la red de creencias que posee el individuo. Desafortunadamente no existe un procedimiento simple formal para evaluar la

compatibilidad del set de juicios de probabilidades con sistema de creencias del individuo (Tversky & Kahneman 1974)

La impresión de las personas a través de la categorización sobreviene espontáneamente y sin aviso. Esto podría explicar las dificultades en generar cambios a través del entrenamiento de los observadores. Los formularios que se utilizan para evaluar el desempeño profesional utilizan una serie predeterminada de dimensiones (ej: comunicación, examen físico, educación al paciente). Este formato teórico puede no ser compatible con el proceso cognitivo de categorización de la mente humana. Es posible que el proceso de traducción desde la categorización nominal a los juicios ordinales colabore también al error(Gingerich, Regehrs & Eva, 2011).

Hay mucho que aprender e investigar en relación al comportamiento de los observadores. Es necesario introducirnos en sus mentes con el fin de comprender que es específicamente lo que consideran importante al momento de tomar decisiones sobre el desempeño de un alumno. Por el momento desde un punto de vista práctico se impone la necesidad de realizar múltiples-mini observaciones con el fin de amortiguar estos sesgos. (Crossley, Johnson, Booth & Wade, 2011)

Referencias bibliográficas.

Alves de Lima, A., Conde, D., Costabel, J., Corso, J. & Van der Vleuten, C.P.M. (2011). A laboratory study on the reliability of the Mini-CEX.*Advances Health in Science Education Theory and Practice,*doi: 10.1007/s10459-011-9343-y

Alves de Lima, A., Barrero, C., Castillo, Y., Bortman, G., Conde, D., Carabajales, J., Galli, A. Degrange, G., & Van der Vleuten, C.P.M. (2007). Validity, reliability, feasibility and satisfaction of the mini-clinical evaluation exercise (mini-cex) for cardiology residency training. *Medical Teacher*, 29,785-90.doi: 10.1080/01421590701352261

Cook D.A., Dupras, D.M., Beckman, T.J., Thomas, K.G., &Pankratz, S. (2008). Effect of rater training on reliability and accuracy of the mini-cex scores: a randomized, controlled trial. *Journal of general internal medicine*, 24, 74-79. doi: 10.1007/s11606-008-0842-3

Crossley, J., Johnson, G., Booth, J., & Wade, W. (2011). Good questions, good answers: construct alignment improves the performance of workplace-based assessment scales. *Medical Education*, 45, 560-569.doi: 10.1111/j.1365-2923.2010.03913.x

Darley, J., & Batson, C.D. (1973). From Jerusalem to Jericho: a study of situational and dispositional variables in helping behaviors. *Journal of Personality and Social Psychology*, 27, 100-119.

Downing, S.M. (2004). Reliability: on the reproducibility on the assessment data.*Medical Education*, 38, 1006-1012.doi: 10.1097/SIH.0b013e318222fde9

Downing, S.M. (2005). Threats to the validity of clinical teaching assessments: what about rater error? *Medical Education*, 39, 353-355.doi: 10.1111/j.1365-2929.2005.02138.x

Eva, K. (2003). On the generality of specificity.*Medical Education*, 37, 587-588.doi: 10.1046/j.1365-2923.2003.01563.x

Gingerich, A., Regehr, G., & Eva, K. (2011). Rater-based assessments as a social judgements: rethinking the etiology of raters errors*. Academic medicine: journal of the Association of American Medical Colleges*, 86, S1-S7. doi: 10.1097/ACM.0b013e31822a6cf8

Govearts, M.J., Schuwirth, L., Van der Vleuten, C.P.M., &Muijtjens J., (2011).Workplace-based assessment: effects of rater expertise. *Advances in Health Science Education Theory no practice*, 16, 151-165. doi: 10.1007/s10459-010-9250-7

Govearts, M.J., Van der Vleuten, C.P.M., Schuwirth L., &Muijtjens A.M., (2007). Broadering perspectives on clinical performance assessment: rethinking tha nature of thintraining assessment. *Advances in Heath Science Education Theory and Practice*, 12, 239-269.doi: 10.1007/s10459-006-9043-1

Kogan, J.R., Hess, B.J., Conforti, L., &Holmboe, E. (2010). What drives faculty ratings of residents´clinical skills?.The impact of faculty´s own clinical skills.*Academic Medicine,* 85; S25-S28. doi: 10.1097/ACM.0b013e3181ed1aa3

Margolis, M., Clauser, B., Cuddy, M., Cicone, A., Mee, J., Harik, P., & Hawkins, R. (2006). Use of the Mini-Clinical evaluation exercise to rate examinee performance on a multiple-station clinical skills examination: A validity study. *Academic Medicine,* 81(10): S56-S60. doi: 10.1097/01.ACM.0000236514.53194.f4

Mazor, K.M., Zanetti, M.L., Alper, E.J., Hatem, D., Barrett, S.V., Meterko, V., Gammon, W., &Pugnaire M.P. (2007). Assessment professionalism in the context odf a structuresd clinical examination: An in-depth study of the rating process. *Medical Education,* 41:331-340. doi: 10.1111/j.1365-2929.2006.02692.

Miller, G.E. (1990). The assessment of clinical skills/competence/performance.*Academic Medicine*, 65, S63-67.

Newble, D.I., & Hoare, J. (1980). The selection and training of examiners for clinical examinations..*Medical Education,* 14(5) 345–349. doi: 10.1111/j.1365-2923.1980.tb02379.x

Norcini, J., &Fortna, G. (2003).The Mini-CEX: A method for assessing clinical skills. *Annals of Internal Medicine*, 138:476-81.

Norman, G.R., Tugwell, P., Feightner, J.W., Muzzin, L.J., & Jacoby, L,L. (1995). Knowledge and clinical solving problem solving.*Medical Education,* 19: 344-56. doi: 10.1111/j.1365-2923.1985.tb01336.x

Park, B., DeKay, M.L., & Kraus, S. (1994) Aggregating social behavior into person models: perceiver-induced consistency. *Journal of Personality and Social Psychology*, 1; 66:437-459. doi: 10.1037/0022-3514.66.3.437

Turner, J.L., &Dankoski, M. (2008). Objective structure clinical exams: a critical review. *Family Medicine*, 40:574-578. doi:10.1186/1471-244X-11-85

Tversky, A., &Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science,* 185, 1124-1131.doi: 10.1126/science.185.4157.1124

Van der Vleuten, C., &Schuwirth, L.W. (2005). Assessing professional competence: from methods to programmes. *Medical Education*, 39:309-317. doi: 10.1111/j.1365-2929.2005.02094.x

Cita del artículo:

Alves de Lima, A. (2012). Variabilidad interobservador. Analizando algunas fuentes de error: heurísticas y categorizaciones.Revista de Docencia Universitaria. **REDU**. Vol.10. Número especialdedicado a la*Docencia en Ciencias de la Salud*. Pp. 229-241 Recuperado el (fecha de consulta) en http://redaberta.usc.es/redu

Acerca del autor

Alberto Alves de Lima

Instituto Cardiovascular de Buenos Aires
Departamento de docencia e Investigacion

Mail:aealvesdelima@icba.com.ar

Director del Departamento de docencia e Investigacion. Sub-jefe de Cardiologia clinica del Instituto Cardiovascular de Buenos Aires.

Médico especializado en Cardiología Clínica en el Instituto Cardiovascular de Buenos Aires (ICBA), entidad afiliada a la Facultad de Medicina de la Universidad de Buenos Aires. Master en educación en las profesiones de la salud de la Universidad de Maastricht y Doctor en medicina de la Universidad de Buenos Aires. Director de la residencia de cardiología clínica del ICBA y Profesor Titular de Cardiología de la Universidad del Salvador. Es profesor del Master Health Profession Education de la Universidad de Maastricht con sede en Brasil, profesor invitado del Instituto Universitario del Hospital Italiano y de la Universidad Nacional del Sur.

Su área de interés académico es la evaluación del desempeño profesional a través de la observación directa. Los temas de sus publicaciones giran alrededor de la formula de utilidad del Mini-CEX.