DISEÑOS LONGITUDINALES:

UN CASO PRÁCTICO

Pablo Delgado Herrera
Nicolás Pérez Gómez

Introducción:

El trabajo que a continuación exponemos va a ocuparse de aplicar los conocimientos teóricos que hemos adquirido con la asignatura de Diseños Longitudinales. Para ello, mediante un caso práctico, analizaremos la evolución de la variable "satisfacción" mostrada por diferentes grupos de asistentes a programas de formación continua en ofimática; contando con los datos resultantes de encuestas de satisfacción aplicadas al final de cada curso formativo, antes y después de haber implantado una intervención.

Partimos de la siguiente situación: en el servicio de formación del P.A.S. se realizan anualmente una serie de cursos de formación continua en ofimática. En los años 2000 y 2001 se recogieron datos de satisfacción de los usuarios tras la finalización de cada curso que indicaron la existencia de una gran falta de la misma. A tal efecto, se decidió intervenir sobre el estilo educativo del profesor y sobre los contenidos de la materia, continuando con la recogida de datos.

Nuestro **objetivo** es, por tanto, comprobar si la intervención realizada por un formador sobre el estilo educativo y los contenidos de la materia, produce un cambio significativo en el grado de satisfacción que los asistentes muestran con el curso recibido y analizar la validez de la intervención.

Metodología*:*

Brevemente, expondremos los procedimientos de intervención y recogida de datos:

-La intervención fue llevada a cabo por un formador que trató de transmitir al profesor, *población diana*, nuevos contenidos y formas pedagógicas, siendo la *población objetivo* el alumnado participante de los cursos de ofimática.

-Los datos parten de los años 2000, 2001, 2002 y 2003, en los que, en cada semestre, se impartieron cinco cursos de tres horas de duración cada uno.

-A todos estos cursos, impartidos siempre por un mismo educador y siempre en el mismo edificio, asistieron cada año los mismos sujetos.

-Las medidas proceden de encuestas de satisfacción realizadas a tal efecto y recogidas al final de cada curso, que constan de doce ítems que miden (o tratan de medir) diferentes aspectos acerca del constructo "satisfacción", mediante una escala tipo likert de 1 a 5. Ejemplos de estos doce ítems son: "Se han alcanzado los objetivos previstos"; "El material entregado es de calidad"; o "Valoración global del curso".

Debemos señalar que tales encuestas no fueron ni realizadas ni recogidas por nosotros, por lo que en el apartado en el que analizamos la validez del estudio, haremos una crítica al método utilizado.

Identificación del estudio:

Como vimos al principio de la asignatura, existen tres grandes grupos de diseños longitudinales, que, a modo de recordatorio, comentaremos:

-**Diseños de medidas repetidas**, que se caracterizan porque en la recogida de datos se realiza tan sólo una medida anterior a la intervención y varias medidas tras ésta, tanto en uno como en varios grupos de sujetos.

-**Diseños de panel**, en los que se realizan pocas tomas de medidas pero de varias variables dependientes siempre en un único grupo de sujetos, antes y después del tratamiento.

-Por último, nos encontramos con los **diseños de series temporales (DST)**, en los que entraremos más a fondo por ser el tipo de estudio que nos ocupa. Este tipo de diseño se basa en la toma de un alto número de medidas tanto anteriores como posteriores al tratamiento. Pero una característica esencial es el análisis de un solo sujeto o unidad, que generará un gran conjunto de datos. Por tanto, como nosotros vamos a trabajar con todo un grupo de sujetos, reduciremos éste a una unidad (utilizando datos agregados) mediante el cálculo de la media de todos los datos.

Dentro de los DST, se pueden realizar dos tipos de estudios: los **diseños de series temporales bivariables**, en los que se registra de forma

paralela la evolución de dos VDs para después estudiar la relación entre ellas, y los **Diseños de Series Temporales Interrumpidas**, que será el elegido para nuestro trabajo. Aquí, donde sólo consideramos una VD, que en nuestro caso es la variable "satisfacción", encontramos otros dos subtipos de estudios:

-Diseños *inter-series temporales y de series combinadas*, modalidad que no se ajusta a nuestro trabajo debido a que se utiliza para comparar los efectos que diferentes VI tienen sobre una VD (*tratamiento alterno*), o el efecto que un tratamiento aplicado a una VD tiene sobre otras VD's de las cuales hemos establecido también una línea base (*línea de base múltiple*).

-Diseños *intra-series temporales*, que será el utilizado en nuestro estudio, más concretamente el modelo clásico A-B, que, aunque presenta importantes amenazas a la validez interna, lo consideramos más adecuado que un diseño de retirada A-B-A (con el que mejoraríamos el control experimental). Aún así, podríamos mejorar el diseño A-B utilizando un grupo control. Comentaremos estos aspectos en el siguiente apartado, que se ocupará de la validez del estudio.

Análisis de datos

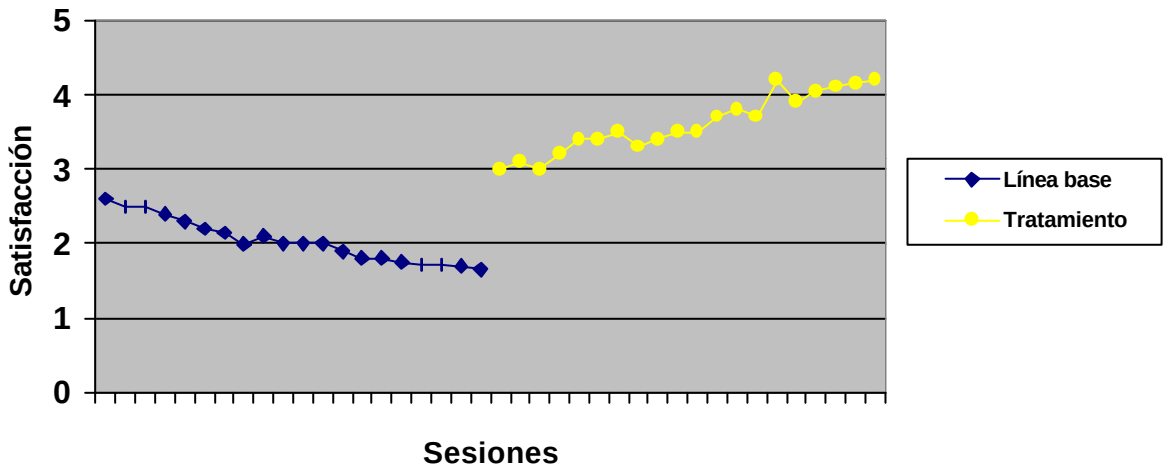
Nos ha parecido muy importante a lo largo de las exposiciones temáticas de la asignatura referidas al análisis de los datos, el hecho de que la evaluación del impacto en los diseños de series temporales presenta un importantísimo problema a la hora de inferir las hipótesis experimentales: la dependencia serial de los datos. Ésta se define como el hecho de que las respuestas emitidas por un sujeto o grupo de sujetos en un determinado momento, están estrechamente relacionadas con las emitidas por el mismo sujeto o grupo en un tiempo pasado de la serie (Vallejo, 1986ª cit. en Bono, 2001), y está causada principalmente por las tendencias o ciclos, que son la dirección natural en el nivel de conducta observada que es independiente de la intervención externa. Nos parece muy importante este efecto porque hace que se rompa el acuerdo

entre el análisis visual y el análisis estadístico de las series temporales, y, por tanto, los dos tipos de análisis deben ser complementarios (Arnau, 1995)

En nuestro caso, estaríamos hablando de autocorrelación en los datos cuando si la satisfacción mostrada por los asistentes hubiera seguido una tendencia natural independiente de la intervención del formador para elevar dicha satisfacción.

Análisis visual

Como sabemos, pese a ser un método comúnmente utilizado para evaluar el efecto de la intervención, el análisis visual presenta numerosos problemas de entre los que podríamos recordar y destacar el hecho de detectar, en muchas ocasiones, intervenciones significativas cuando de hecho no las hay (Arnau, 1995). Por ello, nuestro análisis se ha basado fundamentalmente en un análisis estadístico, si bien, como vemos a continuación, presentamos gráficamente la evolución de los datos tanto en su línea base como en la fase de tratamiento:



Como podemos observar, partimos de una línea base con una tendencia estable y descendente; y observamos cómo, en la línea de tratamiento hay tanto un cambio de nivel como un cambio en la tendencia, siendo ahora ascendente. Así, bajo un análisis puramente visual, podríamos decir que hay efecto de tratamiento.

Análisis estadístico

Nos encontramos aquí de nuevo con el problema de autocorrelación de los datos, que va a hacer que las pruebas estadísticas típicas no puedan ser utilizadas. La autocorrelación sesga las estimaciones de la varianza del error y, por consiguiente, viola el supuesto de independencia de los residuales, implícito en las técnicas estadísticas convencionales, la prueba *t de Student* y la prueba *F Snedecor* (Bono, 2001). Para ello, se han propuesto procedimientos no paramétricos como análisis alternativos, y se ha recomendado la utilización del AST mediante *modelos ARIMA*. Sin embargo, el uso de esta técnica requiere un número de observaciones por fase superior al usual en investigaciones conductuales, concretamente 50 observaciones antes y después de la intervención y, además, en la práctica, aún cuando el analista tenga la suficiente práctica (que no es nuestro caso), con frecuencia se suele identificar incorrectamente el modelo (Arnau, 1995).

Como procedimiento alternativo, se plantea el uso del *estadístico C*, método apropiado para evaluar los efectos de intervención en diseños de series temporales cortas (Bono, 2001). Este estadístico, nos proporciona únicamente la información de si existe o no algún tipo de tendencia estadísticamente significativa en nuestros datos, es decir, si existen variaciones sistemáticas que se apartan de la variación aleatoria. Pero, lógicamente, para utilizar estas pruebas, primero ha de comprobarse el supuesto de autocorrelación, mediante un estimador convencional que viene dado por la siguiente expresión:

$$r_k = \frac{\sum (Y_t - \overline{Y})(Y_{t+k} - \overline{Y})}{\sum (Y_t - \overline{Y})^2}$$

En nuestro caso, el coeficiente de autocorrelación ha sido de 0.798, con lo que nos hemos visto obligados a abandonar las pruebas convencionales para utilizar el estadístico C. Para ello, hemos seguido a Blumberg (1984 cit. en

Bono 2001), aplicando sistemáticamente el diagrama de flujo de tres estrategias propuesto por Bono y Arnau (1997ª; p.52).

Así, en la primera estrategia, aplicamos el estadístico C para probar si hay tendencia en la línea base. Dicho estadístico viene definido por la fórmula

$$C = 1 - \frac{\sum (Y_t - Y_{t+1})^2}{2\sum (Y_t - \overline{Y})^2}$$

Con nuestros datos, obtuvimos un estadístico igual a 4.573, que, para un nivel de confianza α=0.05 resultó ser estadísticamente significativo. En consecuencia, rechazamos la hipótesis nula (la línea de base no presenta tendencia alguna) y concluimos que existe algún tipo de tendencia en nuestros datos de la línea base.

De esta manera, pasamos a la estrategia tercera, donde tratamos de comprobar si en la fase de tratamiento se observa la misma tendencia (hipótesis nula) que en la línea base o bien la tendencia cambia (hipótesis alternativa). Por lo tanto, volvimos a aplicar el estadístico C, obteniendo ahora un resultado de 8.983, que, para un nivel de confianza α=0.05, volvió a ser estadísticamente significativo, con lo que llegamos a la conclusión de que, estadísticamente, el tratamiento es significativo.

Análisis de la validez

A continuación haremos un repaso a los aspectos que pueden considerarse una amenaza para la validez del estudio, concepto éste que suponemos ya conocido, aunque definiremos algunos aspectos que nos parece necesario recordar:

Interna

Como ya sabemos, podremos considerar nuestro estudio internamente válido cuándo estemos en disposición de asegurar que los cambios producidos en la satisfacción son debidos exclusivamente al tratamiento. Esto es algo difícil de controlar en los estudios de encuesta como en el nuestro.

Partiendo de la base de que nuestras expectativas consideran que la relación causal entre tratamiento y cambio en la satisfacción será directa, a continuación comentaremos una serie de argumentos que pueden influir en dicha relación y atentar así contra la validez interna de nuestro estudio.

Con respecto a la selección de la muestra a la cual se aplicó el tratamiento, tenemos que decir que se realizó mediante el procedimiento de accesibilidad. Creemos que el procedimiento idóneo hubiera sido la aleatorización pero las condiciones de estudio se ajustaron a los participantes del curso de ofimática. Al ser un procedimiento no aleatorio, puede ocurrir que la muestra no sea representativa con respecto a la población (personal del P.A.S.) en lo referente a características previas de los sujetos que pueden influir en la satisfacción con el curso. Éstas pueden ser sus conocimientos previos de ofimática (lo que puede influir, p.e., en las puntuaciones de la línea base -pretest-), su nivel de exigencia (que pueden influir, p.e., en el cambio de la satisfacción entre las puntuaciones de la línea base y las del tratamiento), nivel cultural, entusiasmo con la asignatura, etc. , aspectos todos que pueden atentar contra la validez interna de nuestro estudio, introduciendo modificaciones en la relación causal entre las variables.

Ya comentamos en el apartado anterior que utilizar un diseño experimental A-B-A, en el que el tratamiento es retirado para volver a una fase de toma medidas sin tratamiento y comprobar si los resultados cambian, mejora mucho la validez de la investigación, pero en nuestro caso creemos que no es oportuno, ya que no consideramos adecuado retirar unas supuestas mejoras en el estilo educativo para volver a un nivel pedagógico inferior en calidad. Aún así un diseño A-B-A hubiera mejorado la interpretación de los resultados.

Además, no se utilizó grupo control, algo que necesita de la aleatorización para hacerlo equivalente al grupo experimental, que hubiera ayudado a contolar variables extrañas que pudieron afectar a la relación causal entre satisfacción y tratamiento, haciendo así más inequívoca la relación causal inferida en el análisis de los resultados.

Por otra parte, teniendo en cuenta que la población diana de la intervención fue el educador, al que el formador indicó cómo debía llevar a cabo la intervención, y la población objetivo fueron los alumnos del curso, es posible que otra fuente de error surgiera al no llevar a cabo correctamente el educador las instrucciones dadas por el formador, siendo lo idóneo que éste realizara un control sistemático del trabajo de aquél para corregir en la medida de lo posible este problema.

Por último comentar algunos aspectos de la *historia* de los sujetos que pudieron influir en su percepción de satisfacción y, por tanto, en la validez de sus respuestas, ya que, en caso de ocurrir, afectaron a gran parte si no a la totalidad de la muestra. Entre estos cabe destacar posibles cambios estructurales del centro docente del P.A.S. o posibles cambios en las condiciones de trabajo de los sujetos.

Validez de constructo

Consideramos a la *validez de constructo,* que afecta tanto a la validez interna como externa, como el grado de confianza con que pueden realizarse generalizaciones a constructos de orden superior a partir de variables manifiestas específicas, tanto manipuladas como de medida (Ato, 1991). Así, pieza clave de este tipo de validez es el modo en que se han operacionalizado cada una de las variables de la investigación, siendo esto de vital importancia para la generabilidad y utilidad de los resultados de la misma.

Por tanto, atendiendo a la VD, para que ésta tenga validez de constructo debe medir realmente aquello que nos interesa, debe tener *validez convergente*, así como aquello no debe ser medido por otros constructos que no hayamos tenido en cuenta al diseñar nuestra investigación, debe tener *validez discriminante*. Para ello, debe ser operacionalizada de varias formas,

para evitar que subrepresente al constructo teórico de referencia y disminuya su validez. Es decir, debemos realizar un completo análisis conceptual del constructo variable de medida (VD) y tener en cuenta sus características esenciales, para posteriormente emplear diferentes indicadores que tengan en cuenta dichas características, con lo que, finalmente, estaríamos tomando datos de diferentes variables dependientes que pueden ser buenas representantes del constructo teórico, algo que no es demasiado costoso.

Tras esto debemos comentar que nosotros no somos responsables de la operacionalización del constructo VD de nuestro estudio, la *satisfacción*, ya que la encuesta utilizada como instrumento de medida no fue realizada ni aplicada por nosotros, y tampoco hemos tenido acceso al modo en que fue operacionalizado, por lo que no estamos en condiciones de evaluar en qué medida fue o no correcto, y no podemos conocer el grado en que las condiciones de validez convergente y validez discriminante se tuvieron en cuenta.

Pero la validez de constructo no sólo se refiere a la VD, sino que también debemos tener en cuenta la validez de constructo de causas, es decir, el grado en que el tratamiento (VI) afecta al cambio de la variable que queremos medir (*validez convergente*) y no al de otras variables ajenas a nuestro interés (*validez discriminante*). Además son también importantes la validez de constructo sujeto, contexto y momento histórico en que se enmarca la investigación (Ato, 1991)

Es conveniente, al igual que en el caso de las variables dependientes, realizar varias operacionalizaciones del constructo tratamiento para mejorar su validez, pero esto es costoso, ya que aumentar el número de tratamientos hace lo propio con el tamaño de la muestra y el esfuerzo invertido, además de suponer una amenaza para la validez interna del estudio, debido a que es muy posible que existan variables extrañas que cambien a la hora de aplicar los diferentes tratamientos. Aún así, en nuestro caso, lo interesante al medir la variable satisfacción fue comprobar si ésta se veía afectada de alguna forma, en concreto de forma positiva, al aplicar una supuesta mejora en el estilo educativo, es decir, analizar el valor pedagógico del tratamiento, y en ningún

caso se trató de realizar un experimento para estudiar qué factores influyen en la validez. Por tanto consideramos válido el uso de un solo tratamiento.

Además de los errores ya comentados que pueden surgir al realizar una única operacionalización de los constructos, tanto la validez de constructo de la causa, como de sujeto y contexto, pueden verse amenazadas por algunos sesgos, denominados *artefactos*, que producen confusión, ya que covarían con la VI e impiden identificar inequívocamente la causa del fenómeno que estudiamos, de los cuales comentaremos aquellos que nos parecen pudieron afectar al estudio que nos ocupa.

Así, creemos que pudieron producirse errores debido a la denominada *adivinación de hipótesis*, es decir, que los sujetos al contestar la encuesta intuyan o deduzcan los propósitos de ésta y sus respuestas estén condicionadas por este conocimiento. Es más, el mero hecho de sentirse participantes de un estudio puede alterarlas. Concretamente, y siguiendo la "*teoría del rol*", que supone que gran parte de la conducta humana está guiada por los deseos y la conducta de los demás (Ato, 1991), los sujetos podrían haber condicionado sus respuestas de diferentes formas de las cuales comentaremos dos por ser relevantes en nuestro caso: pudieron comportarse como "*el buen sujeto*", esto es, que sus respuestas se acomodasen a las expectativas que supusieron tenía el investigador sobre el resultado se las encuestas. O, por el contrario, pudieron hacerlo asumiendo el rol del *"sujeto negativista"*, respondiendo de modo contrario a las hipótesis supuestas, conducta resultante de verse forzados a participar en la investigación (Miller, 1976, cit. en Ato, 1991).

Por el contrario, efectos como el que las *expectativas del experimentador* pueden tener sobre el estudio no lo consideramos en nuestro caso una amenaza, ya que no se trataba de probar hipótesis, sino, como ya se ha dicho, introducir mejoras en el estilo pedagógico y comprobar si éstas tenían un efecto positivo en la satisfacción de los usuarios, por lo que es normal y adecuado que tanto el formador como el educador realizaran su trabajo siguiendo sus expectativas de mejorar la satisfacción de éstos. Además, teniendo en cuenta que el educador fue el mismo en todos los casos, las características biopsicosociales de éste se mantienen constantes, algo que no

hubiera ocurrido en el caso de que al introducir el tratamiento se hiciera con un educador distinto, pudiendo afectar esto a la reactividad de los sujetos.

Otro sesgo importante que pudo afectar fue que se produjera *interacción entre la administración de las encuestas y el tratamiento*, condicionando la administración de los pretest la recepción del tratamiento, ya que la muestra utilizada fue siempre la misma. Esta es la denominada "sensibilización al pretest" (Lana, 1969, cit. en Ato, 1991). La solución para controlar esta amenaza pasa por utilizar grupos experimentales distintos para administrar los pretests y los postests, pero el caso que nos ocupa es un DSTI, que implica utilizar siempre la misma muestra.

En cuanto al contexto en que se impartieron las encuestas y se introdujo el tratamiento, nos parece el adecuado ya que se realizó en las mismas aulas formativas.

Validez externa

Por todo lo comentado hasta ahora, creemos que la relación causa-efecto que analizaremos en el apartado del análisis estadístico, posee cierta validez externa aunque limitada y con reservas. Por supuesto no podemos generalizar los resultados a contextos diferentes a los cursos formativos del P.A.S., pero tampoco creemos que sea posible hacerlo a otros cursos diferentes a los de ofimática, ya que entre otros aspectos, la muestra de sujetos fue exclusivamente formada por aquellos que recibieron el curso de dicha materia. Por tanto, la única generalización que podemos realizar y con las ya mencionadas reservas, es hacia los futuros cursos de ofimática.

Conclusión

Para finalizar, tras lo comentado en el análisis de la validez y en el análisis estadístico, concretamos que la intervención tuvo un efecto positivo y significativo sobre la satisfacción percibida por los sujetos. En cuanto a la generalización de estos resultados, nos aventuramos a hacerlo hacia futuros cursos de ofimática, aunque con las reservas ya comentadas debido a ciertas amenazas para la validez de la investigación que hemos observado.

Además, sería interesante aplicar el tratamiento a otros tipos de cursos del P.A.S. y realizar el consiguiente estudio para poder aumentar la validez externa de la intervención. Incluso, en el caso de que las nuevas técnicas implantadas fueran novedosas, algo que desconocemos, sería apropiado aplicarlas en otros sujetos, contextos y situaciones y comprobar si los resultados son generalizables a una población distinta a la del P.A.S.

Bibliografía

-ARNAU, J. (1995). ***Diseños longitudinales aplicados a las Ciencias Sociales y Comportamentales***. México: Limusa.

-ARNAU, J. ***Diseños de series temporales: técnicas de análisis***. Barcelona: Edicions Universitat de Barcelona.

-ARNAU, J (1995a). ***Métodos de Investigación en psicología***. Madrid: Síntesis.

-ATO, M. (1991). ***Investigación en las ciencias del comortamiento. I: Fundamentos***. Barcelona: P.P.U.

-BARLOW, D.H. & HERSEN, M. (1988). ***Diseños experimentales de caso único***. Barcelona: Martínez Roca.

Preguntas del caso práctico

1.- Queremos elevar la autoestima a un grupo de escolares con niveles de autoestima muy bajos. El diseño que más nos interesa al plantear una intervención será:

a) AB por razones éticas
b) Diseños de tratamientos alternos
c) AB por razones éticas.

2.- En relación al efecto de la dependencia serial en el análisis de datos de los DST:

a) el análisis visual evita el problema de la dependencia serial.
b) Pruebas t o F pueden ser utilizadas indistintamente
c) Ninguna es correcta.

3.- Tenemos que tras la estrategia 1 de Tryon en la aplicación del estadístico C, hay tendencia en la línea base. Entonces:

a) pasamos a la estrategia 2
b) pasamos a la estrategia 3
c) concluimos que el tratamiento es significativo.

4.- Al plantearnos el análisis estadístico en DST:

a) si hay autocorrelación no podremos usar las pruebas t y F
b) para series temporales cortas, podemos usar los análisis ARIMA
c) a y b son correctas.

5.- La autocorrelación o dependencia serial se define como:

a) la correlación entre los datos de la conducta objetivo con otras conductas.

b) La estrecha relación entre las respuestas emitidas por un sujeto en un determinado momento con otras respuestas emitidas en un tiempo pasado de la misma serie.
c) La dependencia que tienen los datos con los procedimientos de recogida de datos.

6.- Para comenzar el diseño de un estudio debemos operacionalizar los constructos variables, de lo que sabemos que:

a) Es conveniente realizar operacionalizaciones múltiples del constructo VI pero no así del constructo VD, ya que esto último crea confusión.
b) No podemos utilizar diferentes indicadores de un mismo constructo.
c) a y b son incorrectas.

7.- En un estudio sobre la satisfacción de los consumidores de un determinado servicio tendremos validez discriminante de la variable dependiente (satisfacción) si:

a) La operacionalización del constructo *satisfacción* nos permite medir realmente la satisfacción de los usuarios.
b) Tras la operacionalización de las variables no podemos encontrar otros constructos que midan aquello que nos interesa y que no se hayan tenido en cuenta.
c) El constructo *satisfacción* es selectivo con respecto a los demás constructos utilizados

8.- Una intervención fue aplicada a una muestra aleatoria de una determinada población y los resultados fueron replicados en otra muestra también aleatoria de la misma población tras lo que se concluyó que tiene validez externa, por tanto:

a) No sabemos si tiene validez interna
b) Tiene validez interna
c) Tiene validez interna y validez convergente

9.- En una investigación las respuestas de un sujeto se acomodaron a las expectativas del investigador. Según la teoría del rol, se comportó como:

a) El buen sujeto
b) El sujeto fiel
c) El sujeto aprensivo

10.- Queremos comprobar la eficacia de dos tratamientos novedosos en sujetos fumadores. Nuestro diseño será:

a) ABABA
b) Diseños de línea base múltiple (con dos líneas base)
c) Ni a ni b.