Validez interna y externa en los diseños experimentales

Por F. J. TEJEDOR TEJEDOR

1. *Validez interna y externa de un diseño experimental*

El problema de la validez de un diseño experimental es una preocupación permanente en el contexto metodológico de las ciencias humanas. Si bien hoy no se han alterado las circunstancias que condicionan dicha validez cierto es también que puede resultar interesante un replanteamiento de dichas circunstancias y de los presupuestos de actuación experimental del investigador de las ciencias del hombre.

Así pues, el objetivo de las líneas que siguen a continuación es doble:

— Preocupación por las circunstancias experimentales que condicionan la validez de un diseño.

— Nuevas pautas de conducta aconsejables para la experimentación en el contexto de las ciencias humanas.

Tradicionalmente, y siguiendo la opinión de Campbell y Stanley (1), quienes sin duda más importancia han concedido al tema, se acostumbra a diferenciar dos tipos de validez exigibles a todo diseño experimental: validez interna y validez externa.

a) *Validez interna*

Nos determina hasta que punto el investigador puede atribuir la variación observada en la variable dependiente a la presencia de la variable independiente. Es el requisito mínimo imprescindible sin el cual es imposible interpretar el modelo: ¿Introducen, en realidad, una diferencia los tratamientos empíricos en este caso experimental concreto?

---

(1) Campbell, D. y Stanley, J. C.: **Diseños experimentales y cuasiexperimentales en la investigación social.** Buenos Aires, Amorrortu, 1978, pág. 16.

Pueden cifrarse en ocho las variables o factores que han de ser objeto de control por su incidencia en la validez interna de un diseño:

a.1. Historia: Acontecimientos ocurridos durante las diferentes mediciones.

a.2. Maduración: Evolución de los procesos internos de los participantes por el mero paso del tiempo (edad, hambre, cansancio...).

a.3. Administración de test (en términos generales, presentación de estímulos), cuya mera aplicación puede modificar los resultados de aplicaciones posteriores.

a.4. Instrumentación: Los cambios en los instrumentos de medición o en los observadores o calificadores participantes pueden producir variaciones en las mediciones.

a.5. Regresión estadística: Opera allí donde se han seleccionado los grupos sobre la base de puntajes extremos.

a.6. Sesgos: Resultantes de una selección diferencial de participantes para los grupos de comparación.

a.7. Mortalidad experimental: Pérdida de participantes en los grupos de comparación.

a.8. Interacción entre la selección y la maduración.

Analicemos más detenidamente la incidencia de cada una de estas ocho variables en la validez interna del diseño.

a.1. *El efecto historia*

Existen dos posibles momentos de influencia del efecto historia en el proceso de experimentación:

— En el desarrollo de la experimentación, que la propia naturaleza del experimento exige diferenciado en razón a los diversos tratamientos incorporados a estudio, y
— En el proceso final de medición de los resultados habidos en el experimento.

Me parece mucho más problemático el control del efecto historia en el primero de los momentos señalados pues, a menudo, los tratamientos son simultáneos y no es posible equiparar determinadas fuentes de sesgo, por ejemplo, la propia persona encargada de realizar el experimento. Por supuesto que se reducirá el influjo de este factor si existen instrucciones previas respecto al comportamiento del investigador en las diferentes sesiones y tratamientos. Otras fuentes de sesgo pueden igualmente controlarse: hora de la sesión experimental, día de la semana, circunstancias académicas, etc., lo que contribuirá igualmente a reducir la influencia del

factor historia en las posibles diferencias entre los tratamientos y, por tanto, a mejorar la validez interna del diseño.

En control de este factor se dificulta a medida que los tratamientos requieren aplicación individualizada, siendo preciso entonces extremar las condiciones de igualdad de aplicación de los tratamientos del proceso experimental.

Respecto al segundo momento mencionado, proceso final de medición, cabría indicar recomendaciones similares a las expuestas para el primer momento, que se facilitan si el experimento termina en un proceso de medición de la variable dependiente (en la investigación educacional generalmente alguna modalidad de rendimiento) que se lleve a cabo conjuntamente para los diferentes grupos de experimentación. De esta manera existirá una historia intrasesional única que suponemos afectará por igual a los sujetos experimentales, pudiendo entonces considerarse satisfecho el control del efecto historia para nuestro intento de mejorar la validez interna del diseño.

Parece muy claro, respecto a ambos momentos, que cuanto mayor sea el intervalo de tiempo experimental y el tiempo dedicado a la medición final, más difícil resultará controlar el efecto de la historia. Veremos que el tiempo es un problema importante para la investigación experimental ya que tiene repercusiones sobre varios de los efectos perturbadores mencionados.

Las soluciones de tipo estadístico para el control del efecto historia, tales como la aleatorización de las sesiones experimentales, la aleatorización de las circunstancias que rodean el experimento (experimentador, hora, etc.), si bien contribuyen a reducir los efectos perturbadores no parece que solucionen definitivamente el problema. Personalmente opino que respecto al efecto historia, además del control estadístico es conveniente plantearse un control discursivo, crítico, en los términos anteriormente mencionados.

a.2. *El efecto maduración*

Este efecto abarca todos aquellos procesos biológicos o psicológicos que varían de manera sistemática con el correr del tiempo e independientemente de determinados acontecimientos externos.

La maduración es a menudo el punto más crítico de los estudios experimentales en educación, debido a que entre las diferentes mediciones del esquema experimental clásico (pre-test, tratamiento, post-test) transcurría un determinado tiempo capaz muchas veces de explicar por sí mismo, o mejor por los cambios habidos en los sujetos experimentales durante ese tiempo, las diferencias encontradas entre las dos mediciones realizadas.

El problema de la enorme dificultad de control de la maduración de los

sujetos se traduce en un criterio importante para definir la calidad de un experimento, hasta el punto que determina la categorización del diseño como experimental, preexperimental, etc.

La solución a este problema en el contexto metodológico experimental viene dada por la incorporación al experimento del «grupo control» existiendo diferentes alternativas para su utilización lo que determina diseños diferenciados.

La suposición metodológica que conlleva la utilización del grupo control es que el efecto de maduración entre el pre-test y el post-test será similar en ambos grupos —experimental (GE) y control (GC)— pudiendo atribuirse entonces las diferencias encontradas entre los grupos en la evaluación post-test al efecto del tratamiento seguido, exclusivamente, por el grupo experimental

Es decir, en la primera evaluación debería producirse una igualdad de comportamiento en los grupos (sobre todo si responden a criterios de aleatorización): $GC_1 = GE_1$. Posteriormente el GE se somete al tratamiento experimental objeto de investigación. Realizada la segunda evaluación en ambos grupos deberíamos esperar que $GC_2 = GE_2$ si el tratamiento tiene efectos sobre la variable estudiada. Las diferencias entre GC_1 y GC_2 serían las que habría que atribuir a la maduración, en tanto que las habidas entre GC_2 y GE_2 serían las que podrían atribuirse al efecto del tratamiento experimental.

Indudablemente, la preocupación del investigador por el efecto de maduración debe ser mayor a medida que aumenta el intervalo de tiempo que transcurre entre la primera y la segunda evaluación o medición. Cuando este tiempo sea largo (y en no pocas investigaciones educacionales lo es) ni siquiera el establecimiento del grupo control es garantía suficiente para considerar eliminado el influjo del efecto maduración, ya que para ese largo tiempo es más difícil aceptar la suposición de que la maduración en los grupos de control y experimental es similar, pues a veces el propio tratamiento experimental prolongado modifica la evolución psicológica del sujeto: interés por una determinada actividad, vivencia diferenciada de sucesos ajenos al experimento pero con repercusión sobre el mismo, modificación de actitudes, motivación mejorada, etc.; factores en si mismos capaces de producir diferencias entre los grupos y que no pueden ser atribuidas a la influencia de los tratamientos, por ejemplo, a nuevos métodos de enseñanza, a diferentes técnicas de dirección grupal, a diferentes actitudes del profesor, etc.

En los diseños factoriales, en los que no necesariamente hay medida pre-tast ni grupo de control, sino varios grupos sometidos a diferentes tratamientos, no es posible diferenciar los efectos de la maduración propiamente dicha pero sí son considerados como integrantes del denominado error experimental, o variación debida a los sujetos, cuya consideración estadística en los modelos de análisis de varianza no sólo está justificada

sino analizada conceptual y matemáticamente, hasta el punto de convertirse en base estadística fundamental del análisis e interpretación de los datos (2).

No debe pensarse que la no utilización de medida pre-test ni de grupos de control en los diseños factoriales signifique un retroceso respecto a los diseños que sí los utilizan y que ya hemos señalado su razón de ser era precisamente el control del efecto maduración. Realmente cuando se trabaja con varios grupos experimentales cada uno de ellos juega el papel de grupos control respecto a los demás; es decir, suponemos que los efectos de maduración afectan por igual a los diferentes grupos de experimentación, por tanto si los tratamientos no suponen efectos importantes en la variable dependiente la medición proporcionaría resultados similares en los diferentes grupos experimentales.

Bajo este punto de vista, la posibilidad de controlar este efecto perturbador es esencialmente estadística, con doble consideración:

Por una parte, aleatorizar las unidades experimentales (los sujetos) asignados a los diferentes tratamientos.

Por otra, incluir en el modelo el término de error experimental.

Ambas circunstancias son consideradas por el análisis de varianza, hasta el punto, como hemos dicho, de utilizar la variación intragrupo (o variación debida al error experimental) como unidad de medida para la explicación de la variación entre los tratamientos (variación intergrupo). Recuérdese al respecto que el estadístico de contraste F viene definido por el cociente entre el cuadrado medio de los tratamientos y el cuadrado medio del error experimental.

La consideración de medida pre-test es igualmente posible en este tipo de diseños, generalmente definida y utilizada experimentalmente como «covariable», con el análisis de covarianza como respuesta estadística para el estudio comparativo de los resultados habidos.

Así pues, la aleatorización de las unidades experimentales y la utilización de diseños de análisis de covarianza son recursos estadísticos suficientes para garantizar el control del efecto de maduración y considerar satisfecha la exigencia de validez interna en los diseños factoriales.

a.3. *El efecto administración de test*

Este efecto perturbador surge en aquellos experimentos que requieren la realización de pre-test o algún tipo de entrenamiento previo a la medi-

(2) Sobre el término de «error experimental» y su inclusión en los modelos de análisis de varianza puede consultarse el artículo que me publicó la R. E. P. y cuya cita completa es: Tejedor, F. J.: **El término de error experimental en los modelos estadísticos de análisis de varianza. Condiciones subyacentes en el Anva referidas a la variable aleatoria** e. R. E. P., 145, julio-sept., 1979, págs. 97-111.

ción experimental (experiencia de aprendizaje). Parece claro que estas experiencias de aprendizaje pueden modificar las respuestas experimentales posteriores, se siga o no tratamiento experimental. Este efecto, denominado a menudo «efecto reactivo», se produce en mayor cuantía cuando el proceso de medición sea en sí un estímulo al cambio y no un mero registro de comportamiento. Ejemplos de este tipo de efecto serían la preocupación surgida en un sujeto ante un control de peso que, por sí mismo, puede significar un estímulo para adelgazar; la presencia de un micrófono que puede variar las pautas de interacción del grupo; la presencia de un observador extraño que puede condicionar las respuestas de los miembros del grupo... En general, cuanto más nuevo y motivante sea el estímulo, mayor será en influencia su efecto reactivo.

Como es lógico, este efecto no actúa si no existen pruebas de entrenamiento o si dichas pruebas no adquieren un carácter protocolario distinto a un registro normal de comportamiento. En una palabra, sería deseable conseguir que el sujeto no se sintiera protagonista de experimentación y la realización de mediciones no tuviera efectos especialmente reactivos. Indudablemente, cuanto más sofisticado sea el proceso experimental más difícil será conseguir evitar la influencia de los efectos reactivos de las pruebas.

En los diseños experimentales clásicos se controla este efecto con el establecimiento de grupos de control (podría considerarse este efecto como una manifestación específica del efecto maduración). En los diseños factoriales se controla, bien por no ser necesaria la utilización de pre-test o pruebas de entrenamiento previas, bien por la aleatorización de los grupos experimetales, recogiéndose su inflluencia, no eliminada pero sí compensada por el hecho de la aleatorización, en el término de error experimental.

a.4. *El efecto de instrumentación*

Se recogen aquí lo sefectos perturbadores producidos por los cambios habidos en los instrumentos de medición, en las personas encargadas de efectuar las evaluaciones, en los entrevistadores...

Este efecto se controla con facilidad cuando el experimento posibilita su realización en sesión única, con un único experimentador y con un único instrumento de medida, que no admita dudas de interpretación de las respuestas. Si no es posible la realización de esta sesión única habrá que asignar aleatoriamente los observadores a las diferentes sesiones, tanto en el grupo experimental como en el grupo control, procurando además que no sepa de que grupo se trata a fin de evitar que este conocimiento pudiera hacer variar sus registros.

La influencia de este efecto aumentará a medida que se utilicen pretest y cuando la tarea a desarrollar exija un grado elevado de entrenamien-

to o cuando el diseño requiera la utilización de medidas repetidas para un mismo sujeto.

Puesto que, de alguna manera, también este efecto podría incluirse en el efecto maduración, su control experimental requiere las mismas pautas de acción que las entonces mencionadas.

a.5. *El efecto de la regresión estadística*

Ocurre este efecto cuando en el experimento se incluyen grupos de sujetos con puntuaciones extremas en el pre-test. Por el efecto de la regresión los grupos con puntuaciones muy bajas (muy altas) en una variable obtendrán puntuaciones no tan bajas (no tan altas) en otra variable, sin que esas diferencias puedan atribuirse al efecto de los tratamientos ni a los efectos de historia, maduración, etc., sino simplemente al efecto de la correlación imperfecta entre las variables: cuanto menor sea esta correlación mayor será la regresión hacia la media. La falta de correlación perfecta puede deberse a error de medición y/o a fuentes sistemáticas de varianza específicas de una u otra medición.

Campbell (3), que describe y estudia este efecto con relativa profundidad, dice: «Los errores de inferencia ocasionados por no haber tomado en cuenta el efecto de la regresión han planteado tantos problemas en la investigación educacional porque muy a menudo se desconoce su verdadera naturaleza...». En otro momento (4) dice: «Los efectos de regresión son pues acompañamientos inevitables de la correlación imperfecta del test-retets para grupos seleccionados por su ubicación extrema. No son sin embargo concomitantes necesarios de puntajes extremos donde quieran que ellos se produzcan. Si un grupo seleccionado por razones independientes resulta poseer una media extrema, hay una menor expectación a priori de que la media grupal regresione en una segunda prueba, pues se ha permitido a las fuerzas aleatorias o externas de varianza que influyan sobre los puntajes iniciales en ambas direcciones...».

En los diseños factoriales este efecto aparece controlado ya que no suelen utilizarse mediciones pre-test, se procede aleatoriamente en la asignación de los sujetos a los diferentes grupos experimntales y en la medida en que el presupuesto de independencia entre las muestras, requisito básico para la aplicación de las técnicas de análisis de varianza, es empíricamente controlable.

Si en el diseño se incluyen mediciones pre-test, la solución estadística es, como sabemos, la aplicación del análisis de covarianza, que permite previamente controlar la hipótesis Ho : $\beta = 0$, o hipótesis de contraste sobre el coeficiente de regresión entre las variables. Si no se rechaza la hi-

(3) Campbell y Stanley. Obra citada, pág. 25.
(4) Campbell y Stanley. Obra citada, pág. 28.

pótesis nula (si β puede considerarse estadísticamente como cero), el análisis de covarianza no aporta nada nuevo ni diferente al resultado que se obtendría en un análisis de varianza. Pero si la hipótesis se rechaza, el análisis de covarianza tendrá en cuenta la correlación entre las variables ofreciendo una interpretación más coherente sobre la diferencia entre las medias de los grupos de los diferentes tratamientos.

Si en el diseño se utilizan muestras equiparadas, la respuesta estadística es el establecimiento de «bloqueo» entre las correspondientes unidades experimentales, teniendo presente que la equiparación no sustituye sino que completa al necesario proceso de aleatorización.

Sin embargo, aun en las condiciones experimentales expuestas, pueden producirse vacíos interpretativos a causa de los mecanismos de regresión sobre todo si en el diseño se asignan tratamientos diferenciados a los niveles cuantitativos de un factor. En este caso el análisis de las llamadas por Ostle (5) «curvas de respuesta» puede clarificar al experimentador las inferencias a realizar sobre los diversos tratamientos.

a.6. *El efecto de la selección de sujetos*

Este efecto reflejaría el sesgo resultante de una selección diferencial de participantes para los grupos de comparación.

Se controlará este efecto procediendo aleatoriamente en la asignación de los sujetos a los grupos de experimentación. En algún caso se ha pretendido sustituir este proceso de aleatorización por un proceso de equiparación de sujetos. Además de entender que la equiparación experimental cuando se trata de sujetos es muy difícil de conseguir, no puede interpretarse, como dijimos anteriormente, como proceso sustitutorio de la aleatorización sino, en todo caso, como proceso complementario.

La influencia de este efecto se reduce a medida que aumenta el tamaño de los grupos de experimentación.

La fuente de error debida a la diferencia entre los sujetos seleccionados viene recogida en el término de error experimental de los modelos estadísticos asociados a las técnicas del análisis de varianza. Recordemos hasta qué punto es esencial en dicho análisis esta fuente de error teniendo preente que constituye el paradigma respecto al cual se compara la variación encontrada entre los diversos tratamientos. El razonable deseo del investigador de reducir a sus justos límites la cuantía del error experimental puede lograrse, entre otros medios, homogeneizando las unidades experimentales lo que puede conseguirse estableciendo bloques (equiparaciones) entre las unidades experimentales de los diferentes tratamientos, lo que no elimina la conveniencia de proceder aleatoriamente en la asigna-

(5) Ostle, B.: **Estadística aplicada.** México, Limusa-Wiley, 1965, pág. 347.

ción de las unidades experimentales a los tratamientos. Solución estadística todavía más radical para evitar este efecto sería la utilización de diseños intrasujetos o de diseños factoriales mixtos (intersujetos-intrasujetos).

Así pues, si existe duda en el investigador respecto a la no reducción a límites razonables (control) del efecto selección con el proceso de aleatorización y la naturaleza del diseño lo permite, deberá procederse al establecimiento de bloques aleatorios, siendo necesario entonces aplicar las técnicas estadísticas adecuadas y que en términos generales constituyen el tipo de diseños denominados BCA o diseños en bloque completamente azarizados (aleatorizados).

a.7. *El efecto de la mortalidad experimental*

Este efecto, debido a las diferencias que puede implicar la pérdida de participantes en alguno de los grupos de experimentación, se posibilitará a medida que se incremente la duración del experimento. En términos generales, cabe aceptar que la «mortalidad» o abandono de los sujetos a las pruebas de experimentación no es proceso aleatorio sino que suele ir unido a una cierta indiferencia por la situación, lo que puede introducir sutiles diferencias en los grupos, que serán más importantes a medida que sean más reducidos.

El control de este efecto perturbador será efectivo a medida que se evite el abandono de los sujetos sometidos a experimentación o se disponga de sujetos reservas que hayan seguido todo el proceso experimental y que previamente hayan sido asignados al azar a los diferentes grupos de experimentación.

Desde el punto de vista estadístico existen técnicas de predicción de las puntuaciones que obtendrían los sujetos que abandonaron, pero no consideramos suficientemente idóneo este procedimiento ya que en dicha predicción estaríamos introduciendo nuevas fuentes de error.

Si el investigador había planificado un diseño equilibrado (igual número de sujetos en cada grupo de experimentación) el abandono de algún sujeto le llevaría a la utilización de diseños no equilibrados, que si bien están perfectamente estudiados estadísticamente, cierto es también que presentan nuevas dificultades de interpretación y comparación de las diferencias entre los tratamientos. Bajo todos los puntos de vista es recomendable, mientras sea posible, la utilización de diseños equilibrados y, por tanto, es recomendable una vez elegido un diseño equilibrado evitar la mortalidad experimental.

a.8. *El efecto de interacción entre maduración y selección*

Si bien hemos incluido este octavo factor de perturbación por mantener fiel la idea de Campbell y Stanley, a quienes estamos fundamentalmente

siguiendo en esta exposición, no creemos que aporte nada nuevo su camentario por considerarlo implícito en los efectos anteriormente expuestos; es decir, es posible la actuación conjunta, interactiva, de dos o más de los efectos de cada factor por separado. Y si entrásemos a discutir sobre los efectos de las interacciones tendríamos que considerar también otras como maduración-historia, selección-historia, selección-test, etc., etc., hasta agotar todas las combinaciones de primer orden (entre dos factores) y de órdenes sucesivos.

b) *Validez externa*

Plantea el interrogante de la posibilidad de generalización: ¿A qué poblaciones, situaciones, variables de tratamiento y variables de medición pueden generalizarse estos efectos?

Los factores que amenazan la validez externa o representatividad de un diseño serían:

b.1. Efecto reactivo o de interacción de las pruebas: Cuando la prueba modifica la calidad de la reacción del participante a la variable experimental, por lo que no sería legítimo generalizar a poblaciones que no han recibido el estímulo-prueba.

b.2. Interacción de los rasgos de selección y la variable experimental.

b.3. Efectos reactivos de los dispositivos experimentales: Impedirán hacer extensivo el efecto de la variable experimental a las personas expuestas a ella en una situación no experimental.

b.4. Interferencia de los tratamientos múltiples: Suelen persistir efectos de los tratamientos anteriores.

Procedemos a comentar un poco más detenidamente cada uno de ellos.

b.1. *Efecto de la interacción de las pruebas y X*

Este efecto ocurrirá cuando el pre-test aumente o disminuya la sensibilidad o calidad de la reacción del participante a la variable experimental, haciendo que los resultados obtenidos para una muestra con pre-test no sean representativos de lo sefectos de la variable experimental para poblaciones sin pre-test.

El efecto del pre-test sobre la variable dependiente X dependerá del grado en que las situaciones de medición experimental difieran notoriamente de las características del conjunto respecto del cual se pretende generalizar. Cuando se utilizan pruebas de experimentación poco o nada usuales para los sujetos participantes no resultará procedente la generalización a sujetos que no hayan vivido el proceso de experimentación.

Nótese que esta preocupación fue planteada en términos parecidos al

hablar de la validez interna. Entonces nos preocupamos de controlar la posible repercusión de la experiencia-aprendizaje de las pruebas experimentales sobre la propia medición; ahora nos referimos a la posibilidad de generalizar los resultados experimentales a sujetos que no han vivido el proceso de experimentación.

Así pues, se potenciará la validez externa de un diseño en lo referente a este efecto a medida que el proceso de experimentación no suponga una ruptura importante respecto a la conducta normal del sujeto; es decir, a medida que no implique estímulos capaces de provocar conductas reactivas.

b.2. *Efecto de la interacción entre la selección y X*

Este efecto hace básicamente referencia a los problemas de la selección de los sujetos participantes en un determinado experimento; es decir, a la representatividad de la muestra utilizada.

Si bien no es fácil lograr una muestra perfectamente representativa de la población a la que pretende representar (por supuesto, más difícil a medida que la población de referencia aumente), debe aspirarse a ello como un presupuesto básico de investigación. Lo que de ninguna manera es aceptable es despreocuparse del problema en la planificación del experimento y pretender posteriormente generalizar los resultados obtenidos.

El problema de la representatividad de la muestra no es inherente al propio experimento (validez interna) sino que está relacionado con la generalización de resultados (validez externa). Sería válida al respecto la pauta de comportamiento científico de «no generalizar en las conclusiones más allá de lo que los datos permitan». Y esta pauta viene marcada por los aspectos mencionados:

— Uno de carácter metodológico (control de variables, saltos en el vacío en el proceso inductivo, presunciones de leyes establecidas, etc.).
— Otro de carácter más restrictivo, estrictamente estadístico (no obtener conclusiones para poblaciones no representadas aleatoriamente en la muestra experimental utilizada).

La negativa de algunos centros a participar en investigaciones, la tendencia del experimentador a utilizar centros en los que «es bien recibido», lo costoso —económica y temporalmente— de diseñar un experimento incluyendo las diferentes tipologías de sujetos, el problema de los «voluntarios» en la experimentación, etc., son algunos de los factores que pueden contribuir a menoscabar la validez externa de un diseño.

Quizá sea oportuno tener presente a la hora de obtener conclusiones de un experimento que es mucho más peligroso científicamente la aceptación de falsos positivos que la de falsos negativos; si se prefiere, las fal-

sas afirmaciones generales que las dudas metódicas (en el contraste de hipótesis y en términos estrictamente probabilísticos, el error tipo II que el error tipo I).

b.3. *Efectos reactivos de los dispositivos experimentales*

No tenemos más remedio que insistir en la importancia que supone la obvia artificialidad de la situación experimental y la conciencia del sujeto de que está participando en un experimento, que a menudo son causas más que suficientes de carencia de representatividad y por tanto impedimentos para el proceso de generalización.

Ya hemos advertido reiteradamente que, si es posible, deben planificarse pruebas experimentales acordes con «la rutina» de los sujetos que integran la muestra experimental, por una parte, y de la población a la que representan por otra. A medida que no sea posible evitar situaciones experimentales novedosas (reactivas) habrá que restringir los límites de generalización de conclusiones.

Como un efecto reactivo importante es la presencia del experimentador, desconocido por los sujetos participantes si no es el psicólogo o uno de los profesores del centro, habrá que determinar qué es más importante, si dirigir el experimento personalmente con lo cual se potenciaría la validez interna y menoscabaría la validez externa (experimento más riguroso pero menos generalizable en sus conclusiones), o delegar la ejecución del experimento en personal del centro y pasar el experimentador a desempeñar un papel de coordinador y de análisis técnico de resultados (experimento de conclusiones más aplicables pero menos, quizá, riguroso). La solución óptima estaría más en la línea de una «investigación en la acción», dirigida por los propios psicólogos y educadores, aconsejados en la planificación del experimento por expertos en metodología de la investigación.

b.4. *Efectos de interferencia de tratamientos múltiples*

Este efecto tendrá lugar cuando se apliquen diferentes tratamientos a un mismo grupo de sujetos, ya que suelen persistir los efectos de los tratamientos anteriores.

Si, como venimos analizando, con un sólo tratamiento surgen posibles efectos que condicionan el proceso de generalización, éste se verá seriamente comprometido al realizarse respecto a experimentos que suponen tratamientos múltiples, ya que los efectos descritos anteriormente pueden aparecer en uno o varios de los tratamientos, adquiriendo en el conjunto, como es lógico, mayor importancia.

La solución metodológica será obviamente evitar la aplicación de tratamientos múltiples a los mismos grupos de experimentación. Puede lo-

grarse diseñando el experimento en etapas sucesivas (una etapa por tratamiento) con grupos de estudio diferenciados aleatoriamente. Podrá igualmente emplearse un diseño factorial completo, donde las unidades experimentales se asignen al azar a cada combinación de tratamientos (niveles de los factores).

Si no se dispone del suficiente número de sujetos o si se pretende eliminar el efecto de los respectivos tratamientos a grupos diferentes aplicando a todos los participantes la totalidad de los tratamientos, podrían utilizarse diseños «compensados» (denominados también «rotativos», «cruzados», «de conmutación»...) cuyo análisis estadístico mediante la técnica del análisis de varianza se concretiza en los diseños de «cuadrado latino» (doble bloqueo), «cuadrado grecolatino» (triple bloqueo) o en los «diseños intrasujeto», los más específicamente adecuados para esta situación, pudiendo así ser estudiada la influencia del efecto que nos ocupa.

2. *La validez de los diseños factoriales*

Aunque ya hemos hecho alguna alusión a los diseños factoriales, creo que pueden ser interesantes algunas precisiones al respecto.

A grandes rasgos, los diseños factoriales suponen metodológica y/o experimentalmente la posibilidad de incluir en el experimento más de una variable independiente viniendo a significar, en opinión de Campbell y Stanley, «el origen de la enorme brecha que separa las metodologías avanzadas de las tradicionales en el ámbito de la investigación educacional» (6).

La incorporación al experimento de más de una variable independiente supone, por una parte, una mayor aproximación al fenómeno estudiado, generalmente complejo, y por otra, la posibilidad de estudiar las interacciones, es decir, la influencia conjunta de dos o más variables en un determinado fenómeno.

No es nuestro objetivo en este trabajo describir este tipo de diseños, que pueden encontrarse en numerosos textos de estadística experimental (7), sino reconsiderar la problemática que presentan estos diseños respecto a su validez interna y externa.

En términos generales, cabría afirmar que los resultados experimenta-

---

(6) Campbell y Stanley. Obra citada, pág. 58.

(7) Citamos una somera bibliografía (en castellano) sobre diseños experimentales: Arnau, J.: Obra citada en (9).

Glass, G. V. y Stanley, J. C.: **Métodos estadísticos aplicados a las ciencias sociales.** Madrid, Prentice-Hall-Inter., 1974.

Li, Ch. Ch.: **Introducción a la estadística experimental.** Barcelona, Omega, 1969.

Ostle, B. Obra citada.

Ruiz Maya, L.: **Métodos estadísticos de investigación. Introducción al análisis de varianza.** Madrid, I. N. E., 1977.

Tejedor, F. J.: **La investigación en psicología y educación: Diseños experimentales con análisis de varianza.** Madrid, Alhambra, en prensa (aparición en dic. 80-enero 81).

les pueden estar afectados por los mismos efectos comentados para la generalidad de los diseños, dependiendo su posible comparecencia de la utilización o no de pre-test, del grado de aleatorización de las muestras utilizadas, del carácter reactivo de los tratamientos, etc.

Pero será la posibilidad comentada del estudio de las interacciones lo que nos invita a una reconsideración especial sobre la validez externa de estos diseños; es decir, la repercusión de las interacciones en los procesos de generalización de conclusiones, que puede formularse con carácter general en los términos siguientes:

— Un estudio detallado de las interacciones (estadístico y gráfico) puede favorecer el proceso de generalización acertada.

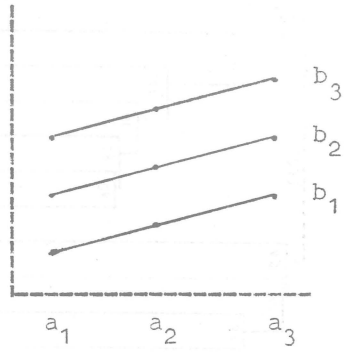
— Las interacciones no monótonas, estadísticamente significativas, limitan la posibilidad de generalizar, si bien una vez detectadas y analizadas permiten realizar inferencias más acertadas acerca de los factores principales (especificidad de efectos).

— Las interacciones monótonas, estadísticamente significativas, producen menos limitaciones, a veces ninguna, sobre la realización de inferencias.

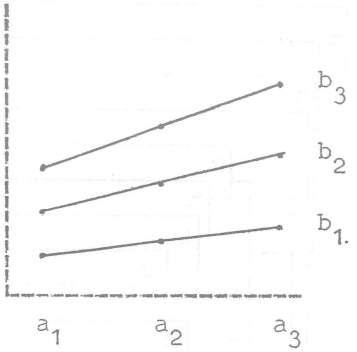
Repetimos la conveniencia de graficar las interacciones antes de proceder a la interpretación de los datos experimentales.

Trataremos de aclarar estas afirmaciones mediante algunos ejemplos, tomados de Campbell (8):

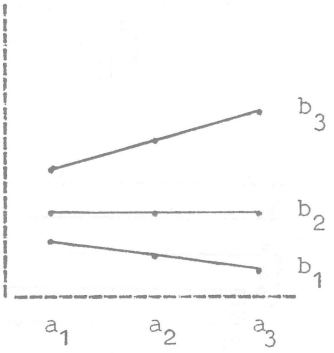
(8) Campbell y Stanley. Obra citada, pág. 60.


b3
b2
b1
a1
a2
a3

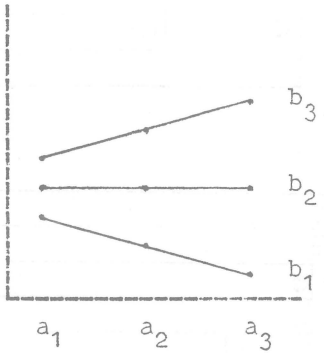
(a)


b3
b2
b1.
a1
a2
a3

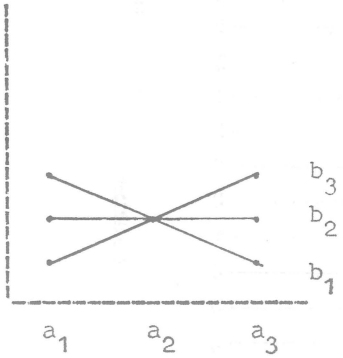
(b)


b3
b2
b1
a1
a2
a3

(c)


b3
b2
b1
a1
a2
a3

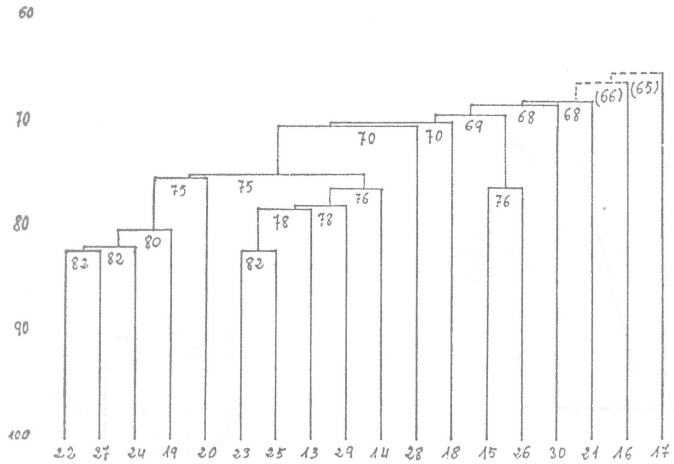
(d)


b3
b2
b1
a1
a2
a3

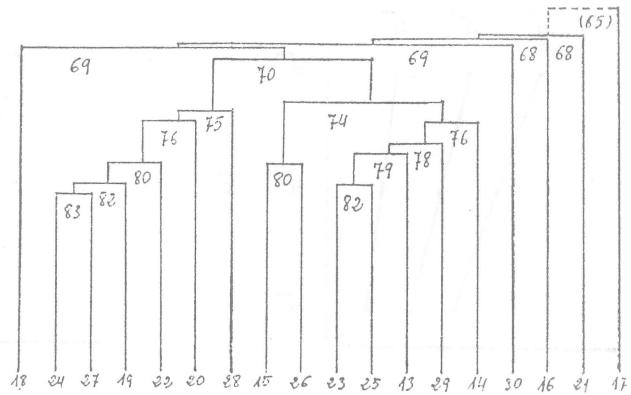
(e)

DENDOGRAMAS CORRESPONDIENTES A LA EDAD MASCULINA DE 15-16 AÑOS

Año 1975-6

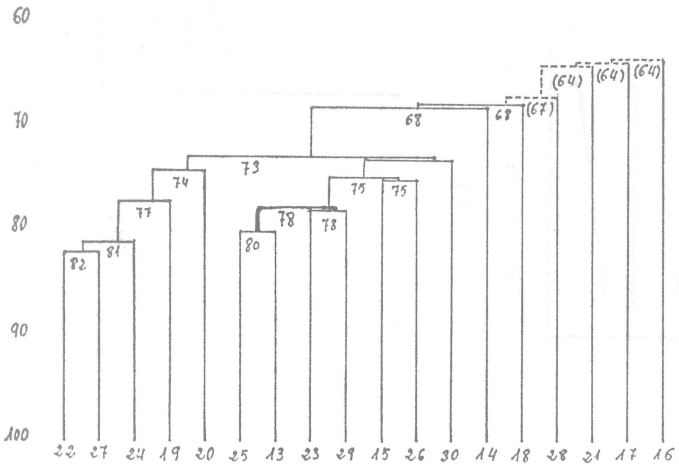


Año 1974-5

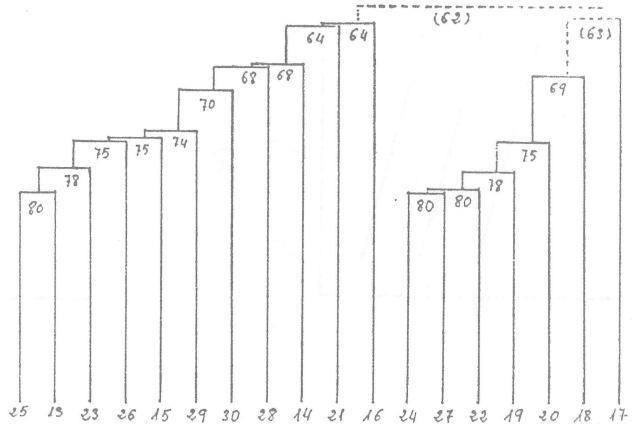


DENDOGRAMAS CORRESPONDIENTES A LA EDAD FEMENINO DE 15-16 AÑOS

Año 1975-6

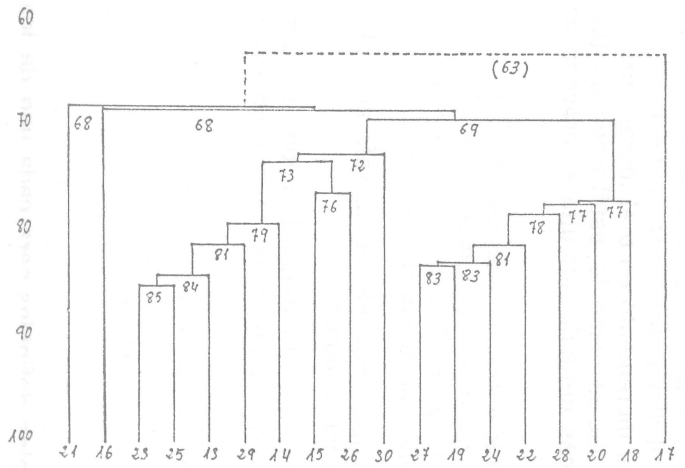


Año 1974-5

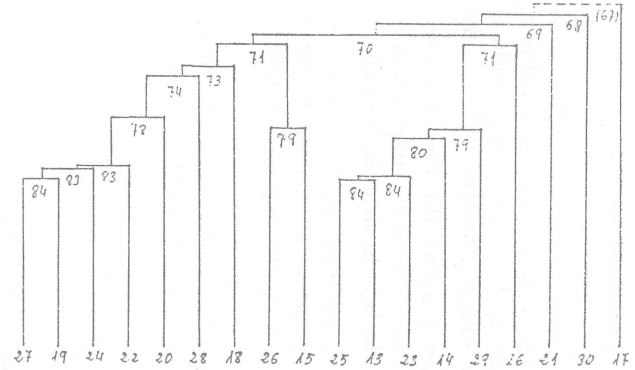


DENDOGRAMAS CORRESPONDIENTES A LA EDAD MASCULINA DE 16-17 AÑOS

Año 1975-6

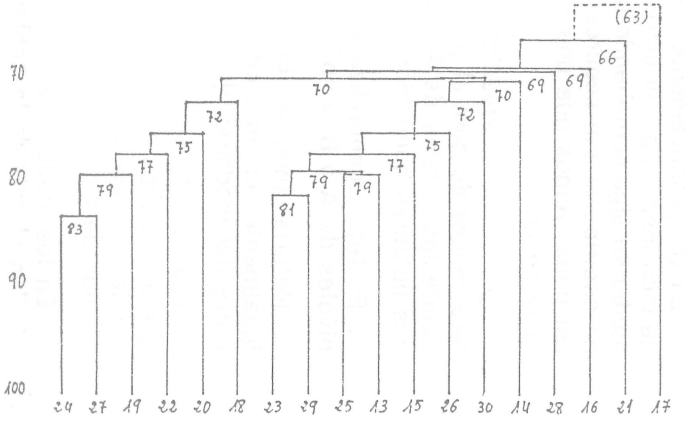


Año 1974-5

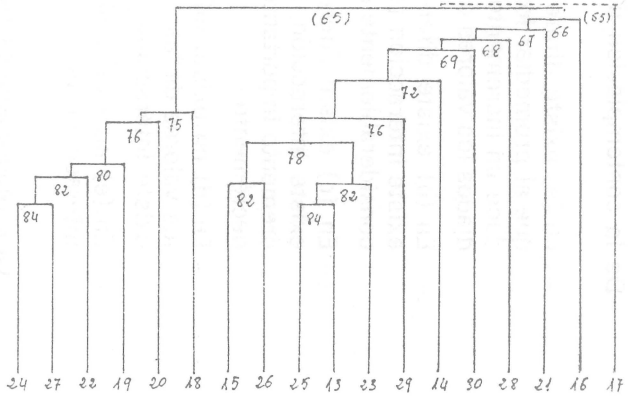


DENDOGRAMAS CORRESPONDIENTES A LA EDAD FEMENINA DE 16-17 AÑOS

Año 1975-6



Año 1974-5



De la contemplación de los gráficos podemos deducir:

— En (a) existe diferencia entre los niveles de los factores A y B ya que al promediar los valores de las tres *b* para cada valor *a* se produce un incremento en el valor de *a*; lo mismo para cada *b* promediados los valores de *a*; no existe interacción (paralelismo de líneas).

— En (b) existe diferencia entre los niveles de los factores A y B; existe interacción (el incremento sufrido en b_3 respecto a a_1 y a_3 es considerablemente mayor que el de b_1); le interacción es monótona.

— En (c) existe diferencia entre los niveles de los factores A y B; existe interacción, no monótona, pues mientras en b_3 se da un incremento importante respecto a a_1 y a_3, en b_2 y b_3 supone un ligero decremento.

— En (d) no existe diferencia entre los niveles de A (si se promedian los valores de las tres *b* para cada *a* resultaría una línea horizontal); existe interacción, no monótona.

— En (e) no existe diferencia entre los niveles de A ni de B; existe interacción muy fuerte no monótona.

¿Qué repercusiones tiene la existencia de interacciones en el proceso de generalización?

En (b) prácticamente no habría limitaciones al generalizar la relación de incremento en *b* al incrementar *a* (o al disminuir). En (c), aunque en promedio se siga cumpliendo la relación de incremento en *b* al incrementar *a*, no estaríamos tan seguros de dicha relación como lo estábamos en (b); así pues, vemos que una interacción más definida dificulta el proceso de generalización.

En (d) queda especialmente patente la influencia del estudio de las interacciones de cara al proceso generalizador; si se hubieran variado solamente los niveles de A, quedando B fijo en b_1, los resultados del diseño nos hubieran conducido a generalizaciones erróneas respecto a b_2 y b_3.

En (e) la generalización sólo es posible tomando como referencia los niveles de A y B, y no los factores en su conjunto.

Notoria importancia respecto al proceso de generalización presenta igualmente, en el contexto de los diseños factoriales, la diferenciación entre los modelos fijos y aleatorios.

Los modelos fijos se caracterizan por incorporar al experimento únicamente aquellos niveles de los factores por los que está interesado el investigador; las conclusiones habrán de referirse exclusivamente a esos niveles de esos factores.

En los modelos aleatorios se incorpora al experimento una muestra de niveles al azar de la población de niveles definidos para cada uno de los

factores; las conclusiones pueden generalizarse a toda la población de niveles de la que se han obtenido los niveles incorporados al experimento.

Cabe pensar también en los modelos mixtos: con dos factores, uno fijo y otro aleatorio; con tres factores, dos fijos y uno aleatorio o dos aleatorios y uno fijo, etc.

Parece claro que ofrecen más garantía las conclusiones obtenidas en los modelos fijos, quizá precisamente por su carácter generalizador más restringido; de hecho, la mayor parte de los diseños utilizados normalmente en la investigación psicopedagógica corresponden a modelos fijos.

3. *Validez interna vs validez externa*

Una de las características más notorias del diseño experimental, como señala Arnau (9), es que sea válido para inferir las hipótesis, es decir, que nos permita comprobar la supuesta relación entre los factores antecedentes y consecuentes. Se trata, por tanto, de un problema de eficacia, del cual dependerá fundamentalmente el logro de los objetivos de la investigación.

A todo diseño experimental debe exigírsele tanto validez interna como validez externa. A menudo las condiciones experimentales para alcanzar o aproximarnos a cada tipo de validez no son compatibles. Los rígidos controles para incrementar la validez interna limitan la validez externa de los resultados.

En general, la validez externa o posibilidad de generalización de los resultados de un experimento es más problemática en el contexto de las ciencias humanas que en otros sectores científicos, debido a la complejidad del ser humano.

Las amenazas a la validez externa suelen considerarse efectos de la interacción entre X y alguna que otra variable, que puede suponer causa de la modificación de X ajena a la propia incidencia del tratamiento específico.

Pensamos que la validez externa es un problema preocupante en toda la investigación por lo que entraña de proceso inductivo. Hume decía que la inducción nunca tiene una plena justificación lógica. De hecho los problemas relativos a la validez interna son susceptibles de solución dentro de los límites de la lógica de la estadística probabilística; por contra, los de validez externa no pueden resolverse con estricto rigor lógico de una forma nítida y concluyente. Aunque existen diferentes posturas muy extremas sobre la posibilidad o no de generalización lógica, quizá sea aceptable, como postura intermedia y conciliadora, el supuesto del «aglutinamien-

(9) Arnau, J.: **Psicología experimental. Un enfoque metodológico.** México, Trillas, 1978, pág. 349.

to»: cuanto más cercanos se hallan dos acontecimientos en tiempo, espacio y valor, más tienden a ajustarse a las mismas leyes.

No hay duda de que las interacciones complejas podrán confundir los intentos de generalización, aumentando esta posibilidad a medida que la situación experimental difiera respecto de la situación a la que pretende extenderse.

Hasta no hace mucho tiempo, y aun hoy en determinados círculos, se concedía mayor importancia a la validez interna del diseño que a su validez externa. Hoy, quizá por el notorio avance de las técnicas estadísticas de experimentación y su contribución a la mejora de la validez interna, se demanda a la investigación una mayor posibilidad de generalización en los resultados experimentales. El objetivo básico de la experimentación podría ahora formularse diciendo que consiste en lograr diseños que sin perder validez interna puedan contribuir a mejorar la validez externa.

Nuestra demanda de una mayor validez externa se refiere por tanto al necesario intento por parte del investigador de lograr una mayor similitud entre el experimento y las condiciones a las que pretende generalizarse, siempre que este presupuesto no signifique un deterioro definitivo de la validez interna.

Entiendo que en el campo de la investigación psicológica y educacional, ciencias cada día con mayor carácter de aplicadas, la preocupación por la validez externa del diseño es fundamental, ya que a menudo nos preguntamos por la aplicabilidad práctica de los resultados obtenidos en determinados experimentos; no olvidemos que la representatividad de resultados condicionará en gran medida los procesos de predicción.

Volviendo al problema central que ahora nos ocupa, validez interna vs validez externa, y recordando la incompatibilidad entre ambas que en determinadas situaciones experimentales puede producirse, no creo que hoy siga siendo válida la afirmación rotunda de que debe prevalecer la validez interna, considerando que es mejor para la ciencia probar hipótesis de carácter más restringido que sacrificarlas en aras de generalizaciones vagas e imprecisas.

Tampoco creo que sea aceptable una desconsideración respecto a las exigencias de validez interna y potenciar exclusivamente la posibilidad de generalización de resultados.

Ambas posturas quedan reflejadas en dos concepciones prácticamente irreconciliables del proceso de investigación: experimento de laboratorio y estudio de campo.

En los primeros, se potencia fundamentalmente la validez interna del experimento, limitándose notoriamente las posibilidades de generalización. La aportación de estos experimentos ha sido muy notoria y no creemos que deba abandonarse dicha práctica, aunque la permanente crítica de artificialidad de la situación experimental (los procesos observados en el

laboratorio no tienen por qué ser idénticos a la realidad) determina un replanteamiento profundo de los trabajos realizados en el laboratorio: replanteamiento que ya ha comenzado bajo nuevos presupuestos epistemológicos y que está suponiendo una profunda crisis en las ciencias humanas aplicadas que, por ejemplo en la psicología, está significando una sustitución del paradigma conductista por paradigmas evolutivo-cognotivistas.

En los estudios de campo se pretende estudiar a los sujetos en el medio en que se hallan normalmente y mientras realizan sus tareas habituales, lo que supone una renuncia explícita a los controles de validez interna y una aproximación al sujeto «tal como es», lo que no necesariamente debe entenderse como validez externa.

La solución para la investigación del presente y de un futuro no muy lejano en el contexto de las ciencias humanas nos parece que puede venir ofrecida por una triple vía:

— Los experimentos de campo (línea más cercana a la metodología clásica).
— Los nuevos diseños, resultado de nuevos modelos, utilizados en los estudios evolutivos (desarrollo y aprendizaje), que suponen fundamentalmente perfeccionar los métodos longitudinales y transversales y que, de alguna manera, son consecuencia de los nuevos planteamientos de acercamiento a los hechos psicológicos (modelos de Schaie, Baltes; diseños factoriales intersujetos-intrasujetos...).
— Las concreciones metodológicas que resulten de las nuevas corrientes epistemológicas y que todavía, lógicamente, no es posible aventurar.

El experimento de campo se encuentra un poco a caballo entre el rigor del experimento en el laboratorio y la conexión con la situación real que implica el estudio de campo. Si bien es cierto que en los experimentos de campo disminuyen las posibilidades de control eficaz de variables, sobre todo de los controles de tipo físico que se establecen en el laboratorio, también lo es que existe la posibilidad de establecer controles estadísticos que suplirían a aquéllos con garantía suficiente como para conceder credibilidad a los resultados del experimento; pero desde luego esta credibilidad sólo la entendemos como resultado de la aplicación rigurosa de diseños experimentalmente complejos.

Algunas de las ventajas que suponen los experimentos de campo serían las siguientes:

— El marco en que se realiza el experimento es real, interviniendo las variables correspondientes en condiciones naturales.
— Se amplía la posibilidad de estudiar fenómenos más complejos.
— Se generan nuevas ideas e hipótesis de trabajo y se proporcionan bases para investigaciones más restringidas.

— Son específicamente adecuados para la comprobación de hipótesis amplias y para la resolución de problemas prácticos.
— Se mejora enormemente la validez externa del diseño, aumentando las posibilidades de éxito en las generalizaciones realizadas, por existir entre ambos campos —el muestral y el poblacional— similitud de condiciones ambientales.

El principal inconveniente, ya señalado, sería la reducción de las posibilidades de control físico riguroso, pero la utilización de técnicas estadísticas adecuadas eliminarían este inconveniente y proporcionaría notorias aportaciones en el campo de la investigación psicopedagógica.

Los modelos preconizados hoy por las nuevas corrientes de la psicología del desarrollo pueden igualmente contribuir a mejorar el resultado de las investigaciones realizadas en el ámbito psicopedagógico. En esencia, estas aportaciones pueden resumirse, siguiendo a Oerter (10), en los siguientes términos:

Tradicionalmente se han estudiado las modificaciones corporales y las modificaciones de la conducta en relación con la edad, de forma que la edad se consideraba como variable independiente y los datos fisiológicos, psicológicos y educativos como variables dependientes.

Mientras se trata del crecimiento corporal puede establecerse una correspondencia bastante fija entre la edad cronológica y los cambios fisiológicos. Llama la atención de que la inteligencia, entendida como capacidad media de actividad intelectual, presenta un desarrollo análogo a las medidas corporales: casi puede superponerse la curva de evaluación de la inteligencia a la curva de variación del peso del cerebro. Esto ha inducido a determinados autores a establecer relaciones fijas entre la edad y el nivel de inteligencia, puesto que la curva de inteligencia no sólo puede obtenerse en base a promedios de grandes grupos sino en investigaciones longitudinales.

Parece evidente que esta postura no puede ser aceptada, ya que implica un menosprecio a la influencia del contexto socio-cultural en el desarrollo de la inteligencia o un presupuesto de homogeneidad de las influencias del medio en todas las personas, que realmente no se da.

Parece más bien que las curvas de inteligencia no son curvas de crecimiento puras ni tampoco curvas de aprendizaje puras, resultando probablemente de la combinación de una función de crecimiento con una función de aprendizaje.

Baltes, que se basa en las ideas de Schaie aunque las modifica ligeramente, propuso para el estudio de la dependencia de las variaciones de la

(10) Oerter, R.: **Moderna psicología del desarrollo.** Barcelona, Herder, 1975, pág. 423 y siguientes.

conducta con respecto a la edad modelos de secuencia en los que se tiene en cuenta la dependencia generativa (historicidad) del desarrollo. No se realiza únicamente una investigación transversal sino que se combina ésta con una segunda, y eventualmente una tercera, investigación de la misma clase en una nueva «generación», es decir, con los nacidos cinco años después, siete años después, etc. Se practican además investigaciones longitudinales en algunas «generaciones».

Los efectos de la edad, de la generación y de la repetición de la prueba (que suponen fuente de error en las investigaciones longitudinales), pueden estudiarse conjuntamente, combinando de esta manera la investigación longitudinal con la transversal.

Para que este método tenga éxito, utilizado con muestras independientes, es necesario homogeneizar las unidades experimentales, en lo referente a potenciales congénitos, estimulaciones recibidas del medio, estilo de educación, etc.

En estos esquemas de investigación se pone de manifiesto la importancia de fuentes de variación tales como el medio ambiente, las diferencias genéticas, etc. Por ejemplo, si se controla el proceso de aprendizaje las probabilidades de obtener diferencias significativas entre edad y generaciones disminuyen. Si se homogeneizan las muestras (control de condiciones genéticas e influencias ambientales) disminuye la probabilidad de encontrar diferencias significativas entre sujetos de la misma edad y la misma generación.

El cambio fundamental que suponen estos esquemas de investigación (pese a sus limitaciones) es pasar del estudio de campo (método etológico) en el contexto de los estudios del desarrollo a experimentos que implican métodos de simulación del medio ambiente. Se trata de simular los rasgos esenciales, provocando en determinadas condiciones experimentales y en un tiempo relativamente breve, modificaciones de la conducta, análogamente a como se presentan en el curso natural del proceso de desarrollo.

4. *A modo de síntesis*

A modo de síntesis personal sobre estos problemas y con la seguridad de que cada lector obtendrá sus propias conclusiones, diría que los fenómenos psicológicos y educacionales son el producto de muchos factores que operan simultáneamente, por lo que se hace necesario obtener información sobre la interacción de las variables que se produce en el proceso educacional, el cual es, por esencia, de carácter dinámico. Esto puede lograrse con el empleo de diseños factoriales, que hacen posible el empleo simultáneo de diversos sujetos, situaciones, tareas, estímulos...

No podemos olvidar que la búsqueda de conocimientos absolutamente ciertos excede las posibilidades del investigador, ya que los diseños ex-

perimentales son imperfectos y, en consecuencia, no pueden ofrecer resultados de validez universal. Las proposiciones experimentales son inferencias estadísticas, y por tanto, solamente susceptibles de alcanzar un cierto grado de probabilidad. Objetivo importante para el investigador es reducir el grado de incertidumbre, lo que puede conseguir utilizando diseños técnicamente adecuados y, sobre todo, repitiendo el experimento con otros sujetos en situaciones similares, analizando si la interpretación de las hipótesis se mantiene en los distintos experimentos. Sólo así podrán los educadores establecer hipótesis, teorías y leyes de alto nivel, ya que hasta ahora sólo han logrado generalizaciones de un bajo nivel teórico; reto de nuestro tiempo es tratar de integrar el conocimiento acumulado en niveles de generalización cada vez más amplios.

El tema, indudablemente, podría dar mucho más de sí, pero ante la no posibilidad —ni siquiera conveniencia— de establecer unas pautas rígidas para la realización de trabajos experimentales, estimo más conveniente aconsejar la reflexión personal de los investigadores sobre la problemática específica de los trabajos que han de emprender. Cada tema a investigar es específico, las circunstancias experimentales son específicas, las motivaciones del trabajo son específicas, los objetivos son específicos... Pero si queremos que el trabajo ofrezca credibilidad y contribuya a mejorar la praxis psicopedagógica, hemos de respetar unas normas técnicas de experimentación. Me inclino pues por argumentar en favor de una respuesta personal (resultado de un proceso de reflexión sobre el objeto a investigar y la metodología a utilizar) desarrollada en el marco de unas normas técnicas rigurosas desde el punto de vista estadístico-experimental.

BIBLIOGRAFIA

ARNAU, J.: **La investigación experimental en educación.** Barcelona, I. C. E. Univ. 1974, Informe n.° 10.

BUNGE, M.: **La investigación científica.** Barcelona, Ariel, 1969.

CAMPBELL, D. T.: **Factors relevant to the validity of experiments in social settings.** Psychol. Bull., 54, 1957, p. 297-312.

CASTELL, M. e IPOLA, E. de.: **Metodología y epistemología en las ciencias sociales.** Madrid, Ayuso, 1975.

COCHRAN, W. G. y COX, G. M.: **Experimental designs.** New York, Wiley, 1957.

EDWARDS, A. L.: **Experimental designs in psychological research.** New York, Rinehart, 1960.

FESTINGER, L. y KATZ, D.: **Los métodos de investigación en las ciencias sociales.** Buenos Aires, Paidos, 1972.

KERLINGER, F. N.: **La investigación del comportamiento. Técnicas y metodología.** Buenos Aires, Interamericana, 1975.

PIAGET, J.; MACKENZIE, W. J. y otros.: **Tendencias en la investigación en las ciencias sociales.** Madrid, Alianza, 1973.

POPPER, K. R.: **La lógica de la investigación científica.** Madrid, Tecnos, 1965.

SIDMAN, M.: **Tácticas de investigación científica.** Barcelona, Fontanella, 1973.

STANLEY, J. C.: **The influence of Fisher's The design of Experiments on educational research thirty years later.** Amer. Educ. Res. J., 3, 1966, p. 223-229.

VAN DALEN, D. B. y MEYER, W. J.: **Manual de técnica de la investigación educacional.** Buenos Aires, Paidos, 1978.