Revista de Investigación Educativa, 2000, Vol. 18, n.º 2, págs. 565-585

EVALUACIÓN DE PROGRAMAS PARA ALUMNOS DE ALTA CAPACIDAD: ALGUNOS PROBLEMAS METODOLÓGICOS

Javier Tourón

Departamento de Educación
Universidad de Navarra

«Evaluation is the tool of defensibility.
Where this tool is used skillfully, programs
for the gifted will survive» (SEELEY, 1986)

RESUMEN

Este trabajo parte de la premisa de que es necesaria una educación diferencia para los alumnos de alta capacidad, y por tanto el desarrollo de programas que atiendan a las necesidades peculiares de estas personas. La investigación, principalmente llevada a cabo en países de habla inglesa, ha venido mostrando evidencias abundantísimas de esta necesidad.

Los programas se enfrentan con una problemática evaluativa peculiar, pues deben constantemente mostrar que son eficaces para poder subsistir. La evaluación es la garantía para la defensa de estos programas allá donde existen.

Algunos de los problemas metodológicos con los que la evaluación se enfrenta han sido objeto de consideración. De modo particular los relacionados con la medida y el diseño. Problemas, por otra parte, que si bien tienen una peculiaridad propia en los programas dirigidos a una población particular, comparten muchos aspectos con la problemática metodológica general. La metodología de la evaluación, aunque presenta problemas que hemos tratado también es cierto que ofrece pautas razonablemente adecuadas para valorar la eficacia de tales programas.

ABSTRACT

This papar is based on the premise that highly gifted students require a differentiated education and thus the need for the development of programs geared toward the special needs of these

students. Research carried out principally in English-speaking countries offers abundant evidence in this regard.

Programmes are faced with a peculiar set of evaluative problems since, if they are to be continued, their effectiveness must be constantly demonstrated. Ongoing evaluation is a guarantee of the continued support of such programmes, wherever they might be in use.

Some of the methodological problems encountered in evaluation are considered, particularly those related to measurement and design. On the other hand, although problems of a unique kind are frequently found in those programmes designed for a specific population, they invariably have many characteristics in common with the general methodological issues. Although the methodology of evaluation presents problems which we have analysed, it is also true that it offers reasonably adequate guidelines for evaluating the effectiveness of such programmes.

1. INTRODUCCIÓN

Antes de entrar en la consideración de los principales problemas metodológicos que presenta la evaluación de programas dirigidos a personas de alta capacidad, quisiera plantear algunas cuestiones previas, comenzando por la necesidad misma de dichos programas. ¿Son realmente necesarios los programas específicos para estos alumnos? ¿Por qué?

La LOGSE plantea en sus artículos 36 y 37 la importancia de atender a las necesidades educativas permanentes o transitorias derivadas de las peculiaridades de los aprendices. El decreto 696/1995 plantea, por primera vez en la historia reciente de la legislación educativa, que existen necesidades educativas especiales asociadas a sobredotación intelectual y regula las condiciones y procedimientos para flexibilizar (aunque tímidamente) el sistema educativo. Por otra parte, uno de los pilares de la LOGSE es la atención a la diversidad, que lamentablemente se ha entendido con demasiada frecuencia como atención a los déficit de los escolares, pero nunca a los superávits. Cuando se analiza todo el entramado legislativo actual de nuestro país para la enseñanza no universitaria y los principios y características psicopedagógicas que lo animan, es fácil ver que se adapta, o pretende hacerlo, a las necesidades de todos los escolares. Sobre este particular hemos escrito extensamente en otro lugar (Cfr. Tourón y cols. 1998).

Sin embargo, no son pocos los mitos que planean sobre los alumnos superdotados (de alta capacidad), que actúan como freno para su adecuada atención en la escuela. Mitos y estereotipos que están bien lejos de la realidad y necesidades de estos alumnos (Cfr. Tourón y Reyero, 2000). La superdotación es un constructo complejo, multidimensional que ha de verse como una capacidad potencial que necesita de unas determinadas condiciones para que pueda desarrollarse de modo satisfactorio (Gagné, 1993). Están lejos los tiempos en los que superdotación se veía como algo fijo, dado e inmutable, para dar paso a un claro cambio de paradigma en el que la superdotación sólo llegará a desarrollarse si se dan las condiciones adecuadas (Reyero y Tourón, 2000). Como señalan con acierto Treffinger y Feldhusen (1996) «los talentos emergen y crecen evolutivamente, y para algunos no llegan a emerger porque no se produce una adecuada estimulación en la escuela y la familia. Es imperativo que todos los que tra-

bajan con jóvenes vean los talentos y potencialidades como algo educable y emergente, y no como algo fijo e inmutable».

La identificación de los alumnos de alta capacidad, paso previo para poder plantear estrategias educativas de intervención (programas), se hace precisa en el momento en el que los programas educativos regulares de la escuela no pueden responder a las demandas educativas de estos alumnos, poniendo en riesgo su desarrollo tanto cognitivo como afectivo. Siguiendo a Feldhusen (1986), hay tres premisas básicas que deben ser tenidas en cuenta para entender la necesidad de una educación diferenciada para los alumnos más capaces, en realidad para todos los alumnos: a) cada estudiante tiene derecho a una educación que sea adecuada a sus especiales características y necesidades; b) todo estudiante tiene derecho a unos servicios educativos que le ayuden a desarrollar sus habilidades potenciales al más alto nivel y c) debemos desarrollar los talentos de los jóvenes de modo que sirvan a las necesidades de su propia nación.

En suma que, a mi juicio, está fuera de duda que la escuela y los programas regulares que en ella se desarrollan, orientados al alumno medio, no pueden dar respuesta adecuada a las demandas psicoeducativas de los alumnos más capaces, por lo que es preciso tender a un sistema educativo más adaptativo que favorezca el desarrollo del potencial de cada escolar. Existen excelentes trabajos en los que se analizan con profusión las características y principios que deben seguirse en el desarrollo de programas diferenciados para los alumnos de alta capacidad, y que aunque éste no es el lugar para tratar (Cfr. Brennam, 1988; Kaplan, 1979; Maker, 1982, 1995; Renzulli, 1995; Van Tassel-Baska, 1984, por citar sólo unos pocos), podemos brevemente apuntar que deben seguir, de acuerdo con el *Leadership Training Institute* (ver Maker, 1986) siete grandes principios:

a) El contenido debe estar enfocado y organizado de modo que permita un estudio más elaborado, complejo y profundo de las principales ideas, problemas y temas que integran el conocimiento en los diversos sistemas de pensamiento.
b) Debe permitir el desarrollo y la aplicación de destrezas de pensamiento productivo que permitan a los estudiantes reconceptualizar el conocimiento existente o producir otro nuevo.
c) Debe permitir explorar los cambios constantes del conocimiento y la información y desarrollar la actitud de que es valioso seguir tales cambios en un mundo abierto.
d) Debe estimular el uso, la selección y exposición de recursos especializados.
e) Debe promover la iniciativa personal y el aprendizaje autodirijido.
f) Debe fomentar la comprensión de uno mismo y de nuestras relaciones con las personas, la sociedad, las instituciones, la naturaleza y la cultura.
g) La evaluación de los programas para alumnos de alta capacidad debe, de acuerdo con los principios anteriores, centrarse en destrezas de pensamiento de alto nivel, creatividad y excelencia en el rendimiento y los productos.

Así pues, el análisis de la investigación tanto teórica como experimental, nos permite señalar, sin muchas dudas, que los movimientos en contra de la superdotación que se aprecian en algunos países, el nuestro no es una excepción, son más producto

de posiciones ideológicas concretas que de argumentos educativos y resultados de investigación sólidos que los avalen.

2. LA EVALUACIÓN DE PROGRAMAS PARA ALUMNOS DE ALTA CAPACIDAD

Como señala Seeley (1986) «mientras es cierto que los programas para los superdotados tienen algunas características únicas, no hay necesidad de crear enfoques completamente nuevos para evaluarlos (...). Hay excelentes prácticas de evaluación educativa que son perfectamente adecuadas para los programas de superdotados. No hace falta reinventar la rueda. Lo que no significa que las características especiales de estos programas deban ignorarse» (p. 265). La evaluación de los programas para alumnos de alta capacidad puede llevarse a cabo desde los enfoques denominados tradicionales, vinculados con concepciones *positivistas* de la realidad, de la ciencia y de la evaluación, criticados por muchos como inservibles para captar la verdadera esencia de la acción educativa, o desde concepciones que se agrupan bajo la rúbrica *postpositivista*, son los modelos naturalistas (Cfr. Borland, 1990; Lincoln y Guba, 1985), que conciben la realidad como algo construido, múltiple, donde lo observado se ve en interacción con el observador; donde la generalización se ve como algo imposible y quizá indeseable; y donde el proceso mismo no pretende ser objetivo ni libre la influencia de los valores. Los diseños fijos dan paso a los emergentes, los instrumentos de medida rigurosamente validados dejan su lugar al investigador como principal instrumento de recogida de datos. Lo nomotético es sustituido por lo idiográfico, donde los datos no son *descubiertos*, como si estuviesen ahí fuera, sino que son literalmente *creados* (Guba y Lincoln, 1989). Son modelos que al oponerse a los de corte positivo se denominan alternativos (Cfr. Dinham y Udall, 1986; Callahan y Cadwell, 1986). Pero no es el propósito de este trabajo entrar en el análisis de los modelos posibles para evaluar programas para alumnos de alta capacidad, sino analizar algunos de los problemas metodológicos que en dicha evaluación se plantean. La discusión sobre los enfoques evaluativos ya ha sido objeto de tratamiento en otro lugar de este número monográfico.

Sí que es preciso decir que el análisis de los problemas metodológicos que más adelante vamos a realizar se alinea con una óptica cuantitativa de la evaluación, compatible por otra parte con enfoques y modelos muy diversos.

La evaluación de programas se ha visto vinculada al movimiento de rendición de cuentas nacido en los EEUU a mediados del siglo pasado (para una visión comprensiva general puede consultarse Tejedor, 1994; García Ramos, 1992) y si bien esta es una razonable función, no siempre es la que mayor impacto positivo tiene en la mejora de los mismos. Sin embargo, por los problemas metodológicos que analizaremos enseguida, y por otras razones, la evaluación ha sido con frecuencia vista como una amenaza de supresión de programas especiales, que se ven exigidos a mostrar unos resultados y una eficacia que rara vez se pide para los programas regulares. La evaluación se reduce a un carácter sumativo que, si bien es importante, es netamente incompleto. Como señala Borland (1997, p. 255), «la mejora de los programas es uno de los resultados más importantes de la evaluación y puede ser una de las razones más poderosas para llevar a cabo el proceso.

(...) Dicha mejora debe ser uno de nuestros imperativos, de modo que forme parte de nuestra concepción general de lo que deben ser los propósitos de la evaluación».

Según Renzulli (1975) la evaluación de programas para los superdotados tiene que cumplir cinco propósitos: a) Descubrir si los objetivos se han cumplido o no y en qué grado; b) descubrir consecuencias inesperadas y no planeadas derivadas de las prácticas del programa; c) determinar las políticas subyacentes y las actividades relacionadas que contribuyen al éxito o fracaso en áreas particulares; d) ofrecer un continuo *feedback* durante el proceso en etapas intermedias a lo largo del programa y d) sugerir cursos de acción alternativos, reales e ideales, para modificar el programa.

Callahan (1993) refiriéndose a la importancia de la evaluación de los programas señala seis aspectos clave que deben tenerse en cuenta si pretendemos tener procedimientos de intervención defendibles: a) la evaluación debe entenderse como una parte integrante del diseño y planificación del programa; b) los problemas que surgen en la evaluación de los programas no pueden ser causa que justifique los fallos de la evaluación; c) la evaluación como proceso está cambiando tanto en sus propósitos como en su amplitud; d) la evaluación no supone sólo determinar el valor de un programa; e) los nuevos desarrollos de la evaluación pueden ser de utilidad en la evaluación de los programas para superdotados y f) la evaluación acaba siendo lo que se quiere que sea.

No es posible, sin embargo llevar a cabo una adecuada evaluación de un programa sin una adecuada descripción del mismo, sin un adecuado establecimiento de los estándares, de los puntos de referencia con los que comparar, sin un plan operativo, sin una adecuada base para atribuir los resultados, sin una previsión de cómo actuar ante los posibles problemas que puedan surgir.

No obstante, la literatura especializada en este campo abunda en una serie de problemas y resistencias con las que se encuentra la evaluación de programas. Las principales dificultades se pueden agrupar en torno a nueve aspectos que hemos reelaborado, siguiendo a Callahan (1993), del siguiente modo:

1) La evaluación se ve como una amenaza
2) Con frecuencia los programas están mal definidos y descritos
3) Existen dificultades para determinar cuál es 'el programa' para poder aislar sus efectos
4) No siempre se formulan las preguntas de evaluación apropiadas ni se establecen adecuadamente las prioridades en la evaluación
5) La comparación de los efectos del programa con determinados estándares y el establecimiento de los grupos de control es difícil
6) El profesor como programa
7) Falta de atención a las posibles interacciones entre aptitud y tratamiento
8) Poca claridad en el establecimiento de los indicadores de éxito y problemas en la instrumentación
9) La utilización de la evaluación (sumativa, formativa, administrativa, etc).

Vamos a analizar ahora algunos de los principales problemas de carácter metodológico que surgen en la evaluación de estos programas, principalmente desde una óptica cuantitativa.

3. ALGUNOS PROBLEMAS METODOLÓGICOS EN LA EVALUACIÓN DE PROGRAMAS PARA ALUMNOS DE ALTA CAPACIDAD

Dadas las limitaciones de espacio disponible, vamos a seleccionar sólo algunos de los problemas que consideramos más importantes. Los vamos a organizar de acuerdo a cuatro apartados: a) problemas derivados de la concepción de superdotación; b) las metas y objetivos del programa; c) problemas de medida y d) problemas con los diseños de evaluación. Trataremos los dos primeros con más brevedad y nos extenderemos algo más en los dos últimos por tener una relación más directa con las cuestiones estrictamente metodológicas.

a) *La concepción de superdotación*

Son múltiples las concepciones tanto implícitas como explícitas que se han propuesto sobre la superdotación (Cfr. Sternberg y Davidson, 1986), algunas de las cuales pueden verse desarrolladas en Tourón y cols. (1998); así mismo, en los últimos años se ha venido produciendo una clara modificación del paradigma clásico hacia un nuevo paradigma más centrado en la identificación y desarrollo del talento (un tratamiento extenso puede verse en Reyero y Tourón, 2000). Pero estas diversas concepciones no son un problema *per se* respecto a la evaluación de los programas, ya que como señala Carter (1991), los evaluadores pueden operacionalizar un determinado concepto de superdotación y analizar los resultados del programa en función de lo que se espera a partir del concepto adoptado. Lo que realmente constituye un problema es que los responsables del programa (*stakeholders*) pueden tener diferentes concepciones de la superdotación y esperar resultados diversos del programa, con lo cual es difícil llegar a una evaluación de la bondad del mismo.

«Los evaluadores deben cerciorarse de que los *stakeholders* están trabajando desde el mismo marco de referencia que los evaluadores, tienen las mismas expectativas y están de acuerdo respecto al tipo de estudiantes a los que el programa está sirviendo» (Carter, 1991, p. 249), y por tanto, hay un marco común de referencia y un acuerdo previo sobre los resultados que se esperan como efecto del desarrollo del programa. No es necesario que el evaluador y los responsables del programa coincidan en el concepto de superdotación, lo que es preciso es que se pongan de acuerdo en el concepto que va a operar en un programa dado y de qué modo se operacionalizará. A partir de ahí será más fácil ponerse de acuerdo en las metas y objetivos en los que debe centrarse la evaluación.

b) *Las metas y objetivos*

Este es otro problema potencial de la evaluación. No es infrecuente que las metas y objetivos estén formulados de una manera vaga o ambigua. Y como señala Borland (1997, p. 257) «las metas y objetivos que no especifican claramente qué se espera que los alumnos ganen como resultado de la existencia del programa son de escasa utilidad en la evaluación». Algunos autores recomiendan que la evaluación se centre en

grandes metas como: el incremento de la creatividad, la capacidad de resolución de problemas, el fomento de estrategias de pensamiento, etc. Pero hay otro tipo de resultados más concretos y medibles que suelen venir reflejados en los objetivos de tipo curricular que el programa persigue. Algunos autores recomiendan que la evaluación se centre en las primeras y abogan por el desarrollo de diseños de evaluación que permitan hacerlo (Cfr. Gallagher, 1979), no obstante es fácil ver la complicación que entraña el determinar cuáles serán los indicadores que se considerarán válidos para metas tan genéricas. Por otro lado se puede producir un efecto negativo evidente y es que al existir una relativa distancia entre los indicadores y las metas, la información que se obtenga de la evaluación puede ser poco útil para mejorar el programa, o bien que no se pueda llevar a cabo una atribución razonable entre el programa y el efecto producido.

Por ello, y sin perder de vista la importancia de las grandes metas, al servicio de las cuales deben estar los objetivos más específicos (metas intermedias), la evaluación debe centrarse en buena parte sobre éstos. Los resultados específicos podrán ser utilizados con carácter formativo y podrán emplearse en la mejora y modificación del programa. La solución óptima es llegar a un adecuado compromiso entre la evaluación de las grandes metas y los objetivos curriculares específicos que se supone tienden a ellas.

c) *Problemas de medida*

La evaluación de programas, particularmente desde un enfoque cuantitativo entraña serios problemas relacionados con la medición, no siempre relacionada con el uso tests, aunque éstos son los que presentan los problemas más complejos, tanto si hablamos de la medición de los resultados como de variables de entrada, de contexto o de proceso. Vamos a apuntar algunos de estos problemas.

Quizá el primero de ellos sea la falta de instrumentos adecuados de la que muchos autores se hacen eco (Cfr. Borland, 1997). Este problema bastante generalizado en muchos contextos se manifiesta de modo particularmente grave en el nuestro, donde es difícil encontrar procesos sistemáticos de desarrollo y validación de instrumentos, así como de actualización de los existentes (Cfr. Tourón, Repáraz y Peralta, 1999).

Siguiendo a Feldhusen y Jarwan (1993), podemos señalar entre los criterios clásicos para la adecuada elección de los instrumentos: a) relevancia del test, b) fiabilidad, c) validez, d) baremación, e) sesgos posibles y f) efecto de techo (para un tratamiento de algunos de los señalados puede consultarse Martínez Arias, 1995 y Muñiz, 1996. Y naturalmente es obligada la lectura de los *Standards for Educational and Psychological Testing*, 1999).

Aunque casi todos ellos son muy obvios merecen un comentario en esta panorámica general, ya que si bien pueden considerarse problemas generales de cualquier tipo de evaluación, los problemas son mayores cuando hablamos de una población tan específica como la de los alumnos de alta capacidad, para la que la ausencia de instrumentos adecuados es casi general.

La relevancia del test se refiere a la adecuación entre el propósito para el cual ha sido diseñado y el uso que se pretende hacer de él. Por ejemplo, si pretendemos seleccionar

los candidatos más adecuados para un programa de desarrollo de la capacidad matemática, un test de inteligencia general no parece lo más adecuado; del mismo modo un test de habilidad matemática tiene poco sentido —aunque sea técnicamente correcto— si se pretende seleccionar sujetos para un programa de desarrollo de la creatividad en artes plásticas. Por tanto, al hablar de relevancia estamos refiriéndonos a la adecuación del test para el propósito específico para el que se va a utilizar. Naturalmente este problema está relacionado con la decisión que se tome de evaluar resultados generales (metas) o específicos (objetivos) en un determinado programa. Cierto es que se hace preciso no perder de vista otros procedimientos para abordar la estimación de los efectos de los programas sin el uso de test. Sería el caso de utilizar otras modalidades de evaluación (*assessment*) como el *consensual assessment* utilizado para el estudio de la creatividad, por citar sólo una alternativa (Cfr. Amabile, 1983).

La fiabilidad no precisa demasiados comentarios. Se trata de una condición esencial, aunque no suficiente, para que un test pueda ser empleado en un proceso de evaluación. Es importante valorar la información disponible sobre la fiabilidad de la prueba que pensemos utilizar: sobre qué muestras se ha obtenido, con qué procedimientos, hace cuánto tiempo, etc. Asimismo, relacionado con la fiabilidad, será importante hacer uso del error de medida, ya que permitirá realizar juicios más precisos sobre las puntuaciones individuales, el establecimiento de intervalos de confianza, puntos de corte, etc. Estos datos son importantes a la hora de tomar decisiones.

Ahora bien, todo lo señalado se refiere a una perspectiva de la medida desde la óptica de la teoría clásica, cuyas limitaciones son suficientes como para que se consideren otros abordajes más acordes con los desarrollos modernos de la misma. Nos referimos a la TRI (Teoría de Respuesta al Item) que permite superar muchas de las limitaciones de la teoría clásica. Lamentablemente no es posible extenderse en este punto ahora, pero baste señalar que la TRI debería tenerse más en cuenta en los procesos de evaluación (Cfr. Orden y cols., 1998; Tourón y Gaviria, 2000a y b), toda vez que nos permite aplicar modelos de tests adaptativos computerizados (o no) por ejemplo, de modo que se maximiza la información que se puede obtener de un sujeto con un 'gasto' mínimo de recursos, ya que los ítems que se le presentan se adaptan a su competencia. Se evita así que un sujeto se vea obligado a responder ítems demasiado fáciles o difíciles para él, lo que en el caso que nos ocupa es crucial. Este tipo de estrategias maximizan la información que se puede obtener y ofrecen un error específico para cada puntuación estimada, lo cual es bastante más plausible que calcular un error común para todas las puntuaciones como se hace en la teoría clásica.

Así pues, y dado que el tratamiento de este tema está fuera de las posibilidades de este breve trabajo, se puede señalar que, además de las aportaciones de tipo técnico que ofrecerá la TRI a la hora de construir tests y aplicarlos a situaciones concretas de evaluación, por ejemplo, «su gran contribución se centra en la posibilidad de obtener mediciones invariantes respecto de los instrumentos utilizados y de los sujetos implicados. En la TCT el resultado de la medición de una variable depende del test utilizado (...). En la Teoría Clásica la medición de una variable es inseparable del instrumento utilizado para medirla y ello constituye una seria limitación, pues inevitablemente se acabará definiendo operativamente la variable por el instrumento con que se mide (...).

Además, las propiedades del instrumento de medida, esto es, de los ítems y, por tanto, del test, están en función de los sujetos a los que se aplican (...). El acercamiento clásico se encontraba encerrado en esa incongruencia teórica: la medición depende del instrumento utilizado y las propiedades de éstos están en función de los objetos medidos, de los sujetos. El objetivo central de la TRI será solucionar este problema» (Muñiz, 1990).

La validez es la *condicio sine qua non*. Un modo clásico sencillo de referirse a la validez es decir que se trata de una apreciación del grado en el que un instrumento mide aquello que pretende. Más precisamente habría que decir que la validez no es tanto del instrumento aunque está implicado, —naturalmente— cuanto de las inferencias que pretendamos hacer a partir de las puntuaciones del mismo. Es conocido que clásicamente hemos distinguido entre diversos tipos de validez: de contenido, concurrente, predictiva, convergente, discriminante, etc., pero la concepción más inclusiva de todas ellas es la validez de constructo, que supone una inserción de la medida en la teoría, de modo que medir se convierte en una forma de validar una teoría, la estructura teórica del fenómeno medido. Pero como recomiendan los *Standards* de 1999, citados más arriba, es más correcto hablar de diversos tipos o fuentes de evidencia sobre la validez que de diferentes tipos de validez. «La validez es un concepto unitario. Es el grado en el que la evidencia acumulada apoya las interpretaciones pretendidas para el uso del test. Como los *Standards* de 1985 esta edición se refiere a tipos de evidencias respecto a la validez, más que a diferentes tipos de validez» (*Standards*, 1999).

Sin entrar en mayores tecnicismos ahora, podemos señalar que se trata aquí, para los propósitos que perseguimos, de responder a dos preguntas: a) ¿qué constructo queremos medir?, b) ¿qué evidencias muestra este instrumento de ser una medida adecuada de este constructo? No parece necesario insistir en la importancia de esta característica, sin la cual todas las demás son superfluas (Cfr. APA, 1986; Cronbach, 1970; Cronbach y Meehl, 1955; Tourón, 1989).

Los baremos son una pieza de información imprescindible para poder interpretar las puntuaciones de un determinado test. Para determinar el grado de excepcionalidad y rareza (Cfr. Sternberg, 1993; Sternberg y Zhang, 1995) de las competencias o talentos de una determinada persona es preciso compararla con sujetos comunes en alguna característica, generalmente la edad, el nivel escolar, etc. Pues bien, los baremos de un test lo que reflejan es el comportamiento típico de un grupo concreto en el test, es decir, su nivel de ejecución. Así pues, no será posible decir cuan excelente o rara es una determinada capacidad sin conocer qué es lo esperable en sujetos de esa edad, por ejemplo.

Su importancia es capital, ya que sin baremos adecuados no podremos, desde una perspectiva normativa, valorar el grado o nivel de ejecución de un sujeto en la prueba correspondiente y por tanto será difícil, sino imposible estimar los efectos del programa que queramos evaluar. Un ejemplo patente de este problema, y sus implicaciones en el proceso de identificación (extensible a la evaluación de programas) puede verse en Tourón, Repáraz y Peralta, (1999). Por eso abordar procesos de validación rigurosos que aporten baremos actualizados obtenidos sobre muestras actuales y suficientemente representativas es esencial (pueden consultarse a este respecto los trabajos que venimos realizando de baremación del SCAT en Navarra, por ejemplo, Tourón y cols., 2000; Tourón, 2000).

Los efectos de sesgo son otro de los criterios a tener en cuenta a la hora de seleccionar un instrumento de medida. Los sesgos se refieren, entre otras cosas, al hecho de que las puntuaciones obtenidas por los sujetos pueden ser inferiores o, en general, verse alteradas, por razón de su sexo, raza, situación cultural, religión, etc., lo que llevaría a una inadecuada valoración de los mismos. El sesgo, como señalan Feldhusen y Jarwan (1993), es —principalmente— un problema de fiabilidad del diagnóstico. La justicia (adecuación) del diagnóstico es una cuestión de validez. Por ejemplo, sería poco razonable someter a los alumnos españoles a un test de razonamiento verbal en el que muchos items incluyesen vocabulario perteneciente a algún deporte típicamente norteamericano, como el beisbol o el fútbol americano. Del mismo modo, sujetos que hayan vivido en el ámbito rural extremo durante toda su vida tendrán problemas para contestar a tests profundamente impregnados de cultura urbana. Todos estos efectos producen sesgos que llevan a los sujetos a obtener puntuaciones que no reflejan su habilidad o capacidad real en la variable medida. Por lo mismo, un test de inteligencia general excesivamente verbalizado producirá un sesgo claro en sujetos deficientemente escolarizados o que viven en un ámbito culturalmente deprivado.

Los problemas de sesgo se analizan modernamente a partir del estudio del funcionamiento diferencial de los ítems (*diferential item functioning*). Pero conviene no confundir ambos aspectos. En efecto, un ítem puede tener un funcionamiento diferencial, para digamos chicos y chicas y no tener sesgo. La definición de DIF aclarará este extremo.

«Se dice que un ítem funciona diferencialmente para dos o más grupos si la probabilidad de dar una respuesta correcta a un determinado ítem está asociada con la pertenencia de sujetos de la misma capacidad a uno de los grupos. Si el grado de DIF es significativo desde el punto de vista práctico y puede ser atribuido plausiblemente a una característica del ítem que es relevante para el constructo medido, entonces la presencia de este ítem en el test sesga la estimación de la habilidad de algunos individuos» (Holland y Wainer, 1993).

Así pues, para planificar el proceso de evaluación será necesario atender a la validez y equidad del test para la población específica para la que se va a emplear, al tiempo que se deben estudiar con cautela los baremos disponibles y todas las evidencias que el constructor del test pueda ofrecer respecto al uso e interpretación de las puntuaciones del mismo.

El efecto de techo es el último de los aspectos que queremos señalar en relación con la medida, pero en absoluto el menos importante. Más aún, es un aspecto crítico. «Se refiere, como es sabido, a la falta de un rango de dificultad adecuado en los ítems, lo que conduce a que los sujetos más capaces no puedan demostrar adecuadamente todo su potencial. Dicho en otros términos, el test pierde la capacidad de discriminar o distinguir las diferencias entre los sujetos a partir de determinado nivel. De este modo, cuando se produce el efecto de techo, sujetos muy distintos en su potencial aparecerán como iguales al obtener puntuaciones similares» (Tourón y cols. 1998).

Utilizar un test que no presenta una dificultad adecuada para los sujetos más competentes en un ámbito dado es como hacer una carrera de 100 metros lisos para descubrir corredores de fondo. Ciertamente todos llegarán a la meta, pero si detenemos ahí la

carrera, nunca sabremos a donde podrían haber llegado los corredores con mejor forma física, cuáles realmente son corredores de fondo. Y lo que es peor, consideraremos a todos como velocistas, cuando muchos de ellos no lo son. Esto es particularmente serio a la hora de valorar los efectos de un programa, pues si el test o los tests que empleemos no tienen suficiente *recorrido*, aparecerán como iguales alumnos de competencias muy diversas. Por otra parte, para complicar más las cosas, lo veremos al hablar del diseño, si estos instrumentos se utilizan como medidas pretest y posttest, los alumnos más aventajados tendrán pocas o ninguna posibilidad de mostrar sus ganancias como resultado del programa, pues ya tenderán a obtener puntuaciones muy altas en el pretest. Además, unido a esto está el conocido efecto de regresión por el que los sujetos que en un pretest toman posiciones muy altas tenderán a obtener puntuaciones más bajas en segundas medidas con el mismo test. Este efecto puede atenuar o llegar a cancelar efectos del programa que son reales. Este es un *artefacto* estadístico que debe tenerse muy presente y que afecta a la validez interna de los diseños (Campbell y Stanley, 1979) y que puede paliarse utilizando diferentes instrumentos para las medidas pre o postest, lo que vuelve poner en primer plano la problemática de la medida.

Este efecto será tanto más grave, lógicamente, cuanto más extremos sean los sujetos evaluados. Se considera que comienza a presentarse este efecto cuando la puntuación media de un grupo está por encima del 75% de la puntuación máxima del test, o cuando la distribución de las puntuaciones está muy sesgada negativamente.

Uno de los mejores sistemas para corregir el efecto de techo es utilizar el procedimiento denominado en el ámbito sajón «*out of level testing*», es decir, utilizar tests previstos para sujetos de mayor edad que la de aquéllos que van a ser evaluados (Cfr. Feldhusen, 1991). Este es un sistema utilizado con probado éxito en el estudio de la precocidad matemática a partir del modelo denominado Talent Search (Cfr. Benbow, 1991; Stanley, 1991; Tourón y Reyero, en prensa).

d) *Problemas con los diseños de evaluación*[1]

Así como los problemas tratados anteriormente pueden ser comunes a evaluaciones centradas en el contexto, el proceso, las variables de entrada, etc., los problemas de diseño que vamos a tratar brevemente se refieren principalmente a evaluaciones del producto, de resultados. Cualquier evaluación de programas está relacionada de un modo u otro con la evaluación de los resultados obtenidos por aquéllos alumnos que ha recibido dicho programa. Ordinariamente los resultados han de compararse con los de otro grupo de sujetos de las mismas características pero que no han estado sometidos al efecto del programa. Es, como se comprende, la estrategia clásica del diseño experimental en la que es ocioso entrar aquí. Lo que sí puede tener interés, por ser una problemática importante en la aplicación de este modelo, es el llamado problema del *grupo de comparación* y el problema del *control*. El primero relacionado,

1 En el anexo I incluimos un cuadro con los diseños pre-experimentales, quasiexperimentales y experimentales más frecuentes en la evaluación de programas, analizando las limitaciones a su validez interna y externa de acuerdo con la obra de Campbell y Stanley.

como es obvio, con la selección adecuada de un grupo de comparación para el que recibe el programa (tratamiento), el segundo relacionado con la compleja problemática del control de las variables dentro del diseño. Estos problemas y otros, que no son del caso, han llevado a muchos a preferir el abandono del modelo experimental por costoso e inadecuado a la realidad educativa y a optar por diseños (o paradigmas) alternativos. A nuestro juicio, la evidencia experimental y su aproximación a la causalidad no pueden ser aparcadas por razones más ideológicas que científicas. La capacidad probatoria del diseño está muy por encima de supuestas metodologías más *flexibles*, aunque su puesta en práctica pueda representar serios problemas, algunos de los cuales vamos a analizar.

Aunque sea sucintamente señalemos que los grupos de comparación son grupos de «control no equivalentes» (intactos) porque no se han formado por procedimientos aleatorios, pero que se consideran suficientemente equiparables a los grupos experimentales, y por tanto no son grupos de control en sentido estricto, según la terminología clásica del diseño (Winner, 1971). Cuando los procedimientos de formación de los grupos de comparación no son aleatorios se nos plantean una serie de problemas que vamos a analizar. Desde luego el mejor grupo de comparación para un grupo de niños de alta capacidad que reciben un programa de resolución de problemas, por ejemplo, sería aquél formado por niños de alta capacidad de su mismo entorno que no han recibido el programa. Esto plantea problemas bien obvios, tanto políticos como éticos, y pocos padres estarían dispuestos a que sus niños fuesen privados de una ayuda potencialmente beneficiosas para ellos. Veremos alternativas a este problema.

Desde la lógica del diseño y atendiendo a la validez interna del mismo (Campbell & Stanley, 1966), es difícil poder atribuir los efectos de un programa (tratamiento) sin una comparación estricta con un grupo de control formado aleatoriamente. Aunque irónicamente, según señala Carter (1991) citando a Snow (1974), un diseño que opere sobre grupos aleatorizados puede dejar de ser un *diseño representativo* del contexto del programa que pretendamos evaluar. Por eso Snow recomienda diseños que representen el contexto natural donde los escolares se desenvuelven normalmente y no los diseños artificiales que pueden llevar a los sujetos actuar de modo diferente a cómo lo harían de no estar sujetos a la *manipulación experimental*. El problema es ser capaces de establecer un *equilibrio aceptable* entre las exigencias de la evaluación del efecto producido por el programa y la naturalidad del contexto. Este es el dilema clásico entre la investigación de campo y la de laboratorio. Como señala Carter (1991, p. 262): «Esto es por lo que los evaluadores deben seleccionar diseños que estén lo más próximos posible al diseño ideal, mientras que se acomodan a las restricciones y circunstancias de la situación. Sea cual fuere el diseño que se elija debe permitir al evaluador responder a las preguntas clave de la evaluación en el tiempo asignado al proyecto».

En relación con los grupos de control no equivalentes, que hemos *llamado grupo de comparación*, se han propuesto algunas soluciones que no están exentas de problemas, pero que vamos a comentar brevemente. La primera de ellas es la *equiparación*. Se trataría de seleccionar escuelas o distritos escolares que pudiesen ser emparejados en determinadas variables consideradas relevantes para la igualación de los grupos. Este procedimiento que es viable en determinados contextos tiene problemas evidentes, no

obstante, ya que es difícil determinar en qué variables se deben equiparar los grupos, y más difícil todavía determinar que cualquier otra variable no considerada no pueda convertirse en una hipótesis rival alternativa al efecto del programa. Por otra parte es improbable que un distrito o escuela haya identificado a alumnos de alta capacidad y no haya establecido algún tipo de programa para ellos. Más aún, aun aceptando que es posible equiparar al grupo experimental con un grupo de control razonablemente igualado a él en algunas variables, muchas otras quedarán necesariamente fuera de control, con lo que cualquier atribución causal del efecto del programa estará comprometida.

Un diseño bastante interesante propuesto por Callahan (1983) para la evaluación de programas está inspirado en la técnica del contrabalanceo y está particularmente diseñado para resolver el problema del grupo de control. En el cuadro 1 reproducimos un esquema del mismo con una modificación importante introducida por Carter (1991). En este diseño se comparan alumnos de alta capacidad que han recibido el programa con otros que no, pero para ello se divide el currículo en unidades que se aplican de modo alterno a los grupos. Así, mientras en el momento 1 un grupo recibe la

CUADRO 1
DISEÑO PROPUESTO POR CALLAHAN Y MODIFICADO POR CARTER PARA LA EVALUACIÓN DEL CURRÍCULO PARA ALUMNOS DE ALTA CAPACIDAD

Grupos	**Evaluación de una Unidad curricular**	
	Tiempo 1	**Tiempo 2**
Grupo A (alta capacidad)		
G1	X	Y
G2	X	Y
G3	X	Y
Grupo B (alta capacidad, comparación)		
G4	Y	X
G5	Y	X
G6	Y	X
Grupo C (no alta capacidad)		
R1	X	Y
R2	X	Y
R3	X	Y
Grupo B (no alta capacidad, comparación)		
R4	Y	X
R5	Y	X
R6	Y	X

X: unidad curricular experimental; Y= otra unidad curricular

unidad X, el otro recibe la unidad Y, siendo la situación contraria en el momento 2. De este modo el grupo expuesto a la unidad X actúa de control para el grupo que ha recibido la unidad Y, y del mismo modo, el Y será control para el que ha recibido la unidad X. Obsérvese que se trata de alumnos de alta capacidad en ambos casos que han sido divididos (aleatoriamente de modo ideal) en dos grupos, con lo que la equivalencia está garantizada. Hay algunos problemas prácticos en los que no vamos a entrar, pero que están relacionados con las posibilidades de compartimentar el currículo de este modo o de alterar las secuencias de las unidades, por citar sólo dos casos evidentes.

Pero veamos la segunda parte del cuadro, que es análoga a la primera pero en este caso las unidades se alternan con grupos de sujetos no de alta capacidad. Los grupos C y D reciben las mismas unidades desarrolladas por los profesores de los grupos A y B. De este modo el grupo A es comparado con el C (ambos reciben la misma unidad X) y el grupo B es comparado con el D (ambos reciben la unidad Y). De esta forma, según Carter (1991) se puede estudiar el efecto de la educación diferenciada para los superdotados en variables como la tasa de aprendizaje, la amplitud y la profundidad, ya que se estarán comparando clases regulares con clases de alta capacidad. Esto permitirá a los evaluadores comprobar si las clases regulares pueden beneficiarse del currículo desarrollado para los superdotados en el mismo grado que éstos.

En ocasiones es imposible disponer de un grupo de comparación, por limitaciones de la propia realidad educativa en la que vamos a evaluar un programa o porque no es posible establecer un grupo comparable. Aún en este caso es importante poder establecer algún tipo de comparación que permita valorar el programa.

Un procedimiento que se ha descrito en la literatura es el *pretest retrospectivo*. En este caso los sujetos actúan como su propio control. Una vez que los sujetos han recibido la instrucción su rendimiento es evaluado con algún test o cuestionario que se considere apropiado. A continuación se les ofrece el mismo instrumento y se les pide que respondan al mismo como lo habrían hecho antes de recibir la instrucción. Los resultados se comparan para analizar las posibles diferencias. La debilidad del procedimiento reside en la confianza que se pueda conceder a la capacidad de los sujetos para autoevaluarse en este modo. Payne y Browne (1982) citado por Carter (1991) ofrecen resultados satisfactorios sobre todo en el campo de variables afectivas, aunque afirman haberlo utilizado con éxito también en el campo cognitivo.

Analicemos para terminar este, necesariamente rápido, repaso por algunos de los problemas más importantes en la evaluación de programas, la problemática del control cuando las limitaciones del contexto impiden la manipulación de variables, la formación aleatoria de grupos, etc. Hay tres grupos de diseños comúnmente utilizados: el causal comparativo, los diseños correlacionales y los diseños cuasiexperimentales. Veamos algunas particularidades de ellos.

Los *diseños causales comparativos* son diseños que se utilizan cuando la manipulación de variables es imposible. En su forma más simple se trata de dos grupos naturales (intactos), uno de los cuales ha recibido el programa y el otro no. La ausencia de control es total, pues ni siquiera es posible determinar aleatoriamente qué grupo recibirá el programa. El único control que puede ejercer el evaluador es la selección del

grupo de comparación, que deberá ser tan similar al que ha recibido el programa como sea posible. Aunque los análisis estadísticos que se llevan a cabo con estos diseños son iguales que los que se realizan con diseños experimentales, es preciso no cometer el error de hacer el mismo tipo de inferencias. Sin control sobre las variables toda inferencia causal será inadecuada, sólo será posible realizar afirmaciones relacionales. La debilidad de este diseño es patente, aunque ofrece un grupo de comparación que de ser probada su equivalencia en variables relevantes para el programa, aporta una cierta información.

Los *diseños correlacionales* son una alternativa interesante aunque poco utilizada por los evaluadores. Difieren de los anteriores en que sólo utilizan un grupo y los datos se analizan con técnicas correlacionales. Naturalmente estos diseños utilizan procedimientos que van más allá de la correlación bivariada, empleándose con frecuencia procedimientos de regresión múltiple, análisis discriminante o correlación canónica. Cuando se utiliza la regresión múltiple, la variable dependiente es el resultado que se pretende medir en el programa (simple o múltiple) y las variables dependientes son aquéllas cuya capacidad predictiva sobre tal resultado interesa analizar. De este modo es posible obtener información que, aunque no puede ser causal, es muy interesante para valorar distintas dimensiones del programa. Ni que decir tiene que las posibilidades de los modelos causales o los recientes desarrollos del análisis multinivel permitirán avances muy sustantivos en la evaluación de programas.

Para terminar esta sucinta exposición, podemos hacer mención a algunos diseños cuasi experimentales, quizá los que mejor se adaptan a las limitaciones y condicionamientos contextuales a los que debe atender la evaluación. Si bien es cierto que distan de los verdaderos experimentos en cuanto al control, no cabe duda de que ofrecen una información muy valiosa en la evaluación al tiempo que controlan muchas de las amenazas a la validez interna (ver anexo I).

Dos de estos diseños son: el diseño de *series cronológicas de grupo único* y el diseño de *series cronológicas grupo de control no equivalente*[2].

El primero de ellos utiliza un sólo grupo al que se mide a intervalos regulares antes del tratamiento (programa) y después de terminar la intervención. El disponer de medidas antes y después permitirá determinar el posible efecto del programa, para ello será preciso analizar la serie temporal de medidas antes y después y comprobar la tendencia y el cambio producido por el programa, lo que no puede hacerse de modo correcto en un diseño en el que sólo tuviésemos una medida antes y otra después. Las limitaciones de espacio no nos permiten analizar ejemplos ilustrativos de este diseño (puede verse la obra de Campbell y Stanley, 1966, o Fitz-Gibbon y Morris, 1987).

Finalmente, el diseño de series temporales con grupo de control no equivalente en su forma más simple consiste en un desarrollo del anterior pero utilizando dos grupos naturales (pueden verse las amenazas a la validez en el anexo I). El trata-

---

2 Un diseño muy interesante utilizado con éxito en la evaluación de programas es el de discontinuidad en la regresión, que aquí no vamos a tratar. Puede verse una descripción en Stanley y Robinson (1986) y una aplicación práctica en Robinson y Stanley (1989).

miento es aleatoriamente asignado a uno de los grupos actuando el otro como control (más bien como comparación, pues al no haber sido asignados los sujetos aleatoriamente, no es adecuado denominarlo de control). Como señala Carter (1991, p. 268), «la no equivalencia de los grupos es la mayor debilidad de este diseño. Para interpretar correctamente los efectos del tratamiento, los evaluadores deben detectar las diferencias entre los grupos antes del tratamiento a partir del los pretests y de comparaciones en otras variables que puedan ser relevantes para el programa. Si los análisis en las variables pretest no revelan diferencias significativas, los resultados del programa se puede analizar a partir de la comparación entre los posttests, pero si las hubiese, la utilización de un control estadístico, como el análisis de varianza sería necesario, en realidad es común y conveniente aplicar este control aunque no se encuentren diferencias significativas, siempre que las variables (covariantes) medidas antes del programa sean realmente relevantes y tengan un posible impacto en los resultados».

4. CONCLUSIONES

En las páginas precedentes hemos tratado de plantear la justificación a una educación diferenciada para los alumnos de alta capacidad. La investigación, principalmente llevada a cabo en países de habla inglesa, ha venido mostrando evidencias abundantísimas de esta necesidad. Es pueril pensar que los niños de alta capacidad se desarrollarán adecuadamente sin una intervención y ayuda adecuadas. Simplemente no es así, pero no es ésta una cuestión de opinión, sino de resultados de investigación convenientemente contrastados.

A pesar de ello, determinados prejuicios llevan a muchas autoridades educativas y a los educadores mismos a presentar resistencias más o menos fuertes en contra de una educación diferenciada, que no segregada (el principio de integración tampoco lo permitiría). Por esta razón, los programas se enfrentan con una problemática evaluativa peculiar, pues deben constantemente mostrar que son eficaces para poder subsistir. La evaluación, como señalamos al comienzo, es la garantía para la defensa de estos programas allá donde existen.

Algunos de los problemas metodológicos con los que la evaluación se enfrenta han sido objeto de consideración. De modo particular los relacionados con la medida y el diseño. Problemas, por otra parte, que si bien tienen una peculiaridad propia en los programas dirigidos a una población particular, comparten muchos aspectos con la problemática metodológica general.

La situación de nuestro país en lo que al desarrollo de programas para alumnos de alta capacidad se refiere es muy deficiente, por lo que los resultados de evaluación son prácticamente inexistentes, al menos hasta donde conocemos. O si existen no llegan a ser publicados en los canales ordinarios de divulgación.

Es preciso, a nuestro juicio, flexibilizar el sistema educativo y conseguir una escuela más adaptativa que ofrezca a cada alumno las ayudas que precisa para su desarrollo. Al mismo tiempo se hace patente la necesidad de diseñar programas específicos para atender a tales necesidades, más allá de lo que la escuela regular ofrece. La meto-

dología de la evaluación, aunque presenta problemas que hemos tratado también es cierto que ofrece pautas razonablemente adecuadas para valorar la eficacia de tales programas, al menos desde la óptica que aquí se ha adoptado.

BIBLIOGRAFÍA

Amabile, T.M. (1983). *The Social Psychology of Creativity*. New York: Springer-Verlag.

American Educational Research Association (1985). *Standards for Educational and Psychological Testing*. Washington: AERA, APA, NCME.

American Educational Research Association (1999). *Standards for Educational and Psychological Testing*. Washington: AERA, APA, NCME.

Benbow, C.P. (1991). Mathematical Talented Children: Can Acceleration meet their Educational Needs? En N. Colangelo. y G.A. Davis (Eds.). *Handbook of Gifted Education*.

Borland, J.H. (1990). Postpositivist Inquiry: Implications of the «New Philosophy of Science» for the Field of the Education of the Gifted. *Gifted Child Quarterly*, 34, 161-167.

Borland, J.H. (1997). Evaluating Gifted Programs. En N. Colangelo y G. A. Davis (Eds.). *Handbook of Gifted Education*. 2nd Ed. Boston: Allyn & Bacon.

Brennam, W. (1988). *El curriculo para niños con necesidades especiales*. Madrid: M.E.C. Siglo XXI.

Callahan, C.M. (1983). Issues in Evaluation Programs form the Gifted. *Gifted Child Quarterly*, 27, 33-37.

Callahan, C.M. (1993). Evaluation Programs and Procedures for Gifted Education: International Problems and Solutions. En K.A. Heller, F.J. Mönks y A.H. Passow (1993). *International Handbook of Research and Development of Giftedness and Talent* (pp. 605-618). Oxford: Pergamon Press.

Callahan, C.M. y Caldwell, M.S. (1986). Defensible Evaluation of Programs for the Gifted and Talented. En J. Maker. *Critical Issues in Gifted Education. Defensible Programs for the Gifted*. Volume I (pp. 277-296). Austin: Pro-Ed.

Campbell, D.T. y Stanley, J.C. (1966). Experimental and Quasi-experimental Designs for Research. *Chicago*: Rand McNally.

Campbell, D.T. y Stanley, J.C. (1973). *Diseños experimentales y quasi experimentales en la investigación social*. Buenos Aires: Amorrortu.

Carter, K.R. (1991). Evaluation of Gifted Programs. En N.K. Buchanan y J.F. Feldhusen (Eds.). *Conducting Research and Evaluation in Gifted Education. A Handbook of Methods and Applications (*pp. 245-272). New York: Teachers College Press.

Cronbach, L.J. (1970). Test Validation. En Thorndike, R. L. (Ed.). *Educational Measurement*. Washington: American Journal of Education.

Cronbach, L.J. y Meehl, P.E. (1955). Construct Validity in Psychological Test. *Psychological Bulletin*, 52, 281-302.

Dinham, S.M. y Udall, A.J. (1986). Evaluation for Gifted Education: Synthesis and Discussion. En J. Maker. *Critical Issues in Gifted Education. Defensible Programs for the Gifted*. Volume I (pp. 297-316). Austin: Pro-Ed.

Feldhusen, F.J. (1991). Identification of Gifted and Talented Youth. En Wang, M.C.; Reynolds, M.C. y Walberg, H.J. (Eds.). *Handbook of Special Education*. Vol 4. Oxford: Pergamon Press.

Feldhusen, J.F. (1986).Policies and Procedures for the Development of Defensible Programs for the Gifted. En J. Maker. *Critical Issues in Gifted Education. Defensible Programs for the Gifted*. Volume I (pp. 235-256). Austin: Pro-Ed.

Feldhusen, J.F. y Jarwan, F.A. (1993). Identification of Gifted and Talented Youth for Educational Programs. En K.A. Heller, F.J. Mönks y A.H. Passow (1993). *International Handbook of Research and Development of Giftedness and Talent* (pp. 233-251). Oxford: Pergamon Press.

Fitz-Gibbon, C.T. y Morris, L.L. (1987). How to design a program evaluation. Newbury Park: Sage.

Gagné, F. (1993). Constructs and Models Pertaining to Exceptional Human Abilities. En K.A. Heller, F.J. Mönks y A.H. Passow (1993). *International Handbook of Research and Development of Giftedness and Talent*. Oxford: Pergamon Press.

García Ramos, J.M. (1992). Recursos metodológicos en la evaluación de programas. *Bordón*, 43(4), 461-476.

Gallagher, J.J. (1979). Research Needs for the Education of the Gifted. En J.J. Gallagher, J.C. Gowan, A.H. Passow y E.P. Torrance (Eds.). *Issues in Gifted Education* (pp. 79-91). Ventura, CA: Ventura County Superintendent of Schools.

Guba, E.G. y Lincoln, Y.S. (1989). Fourth Generation Evaluation. Beverly Hills, CA: Sage.

Kaplan, S.N. (1979). *Insevice Training Manual: Activities for Developing Curriculum for the Gifted/Talented.* Los Angeles: Leadership Training Institute on the Gifted and Talented.

Lincoln, Y.S. y Guba, E.G. (1985). Naturalistic Inquairy. Beverly Hills, CA: Sage.

Maker, C.J., & Nielson, A.B. (1995). *Curriculum Development and Teaching Strategies for Gifted Learners*. Boston: Allyn and Bacon.

Maker, C.J. (1982). *Curriculum Development for the Gifted. Rockville*, MD: Aspen Systems.

Maker, C.J. (1986). Defensible Programs for Gifted Students: What are they? En J. Maker. *Critical Issues in Gifted Education. Defensible Programs for the Gifted*. Volume I (pp. 277-296). Austin: Pro-Ed.

Martínez Arias, R. (1995). Psicometría: Teoría de los tests psicológicos y educativos. Madrid: Síntesis.

Muñiz, J. (Coord.)(1996). *Psicometría*. Madrid: Editorial Universitas.

Muñiz, J. (1990). Teoría de Respuesta a los Items. Un nuevo enfoque en la evolución psicológica y educativa. Madrid: Pirámide.

Orden Hoz, A.; Bisquerra, R.; Gaviria, J.L.; Gil, G; Jornet, J. López, F. Sánchez, J.; Sánchez, M.C.; Sierra, J. y Tourón, J. (1998). *Los resultados escolares. Diagnóstico del Sistema Educativo 1997*. Madrid: INCE, Ministerio de Educación y Cultura.

Payne, D.A. y Brown, D.L. (1982). The use and Abuse of Control Groups in Program Evaluation. *Roeper Review*, 5, 11-14.

Renzulli, J.S. (1975). *A Guidebook for Evaluating Programs for the Gifted and Talented*. Ventura, CA: Office of the Ventura County Superintendent of Schools.

Renzulli, J.S. (1995). Intervenciones educativas para el desarrollo de la superdotación en los niños. Ponencia presentada en el II Congreso Internacional de Psicología y Educación. Madrid, 16-18 noviembre.

Reyero, M., y Tourón, J. (2000). Reflexiones en torno al concepto de superdotación: evolución de un paradigma. *Revista Española de Pedagogía,* 215, pp. 7-38.

Robinson, A. y Stanley, T.D. (1989). Teaching to Talent: Evaluating and Enrich and Accelerated Mathematics Program. *Journal for the Education of the Gifted*, 12(4), 253-267.

Seeley, K.R. (1986). Evaluation for Defensible Programs for the Gifted. En J. Maker. *Critical Issues in Gifted Education. Defensible Programs for the Gifted*. Volume I (pp. 265-277). Austin: Pro-Ed.

Snow, R.E. (1974). Representative and Quasi-representative Designs for Research on Teaching. *Review of Educational Research*, 44, 265-291.

Stanley, T.D. y Robinson, A. (1986). Regression Discontinuity: Integrating Research and Program Design in Programs for the Gifted. *Journal for the Education of the Gifted*, 9(3), 181-191.

Stanley, J.C. (1996). In the Beginning: The Study of Mathematical Precocious Youth. En C.P. Benbow y D. Lubisnki (Eds.). *Intellectual Talent. Psychometric and Social Issues* (pp. 225-235). Baltimore, MD: The Johns Hopkins University Press.

Sternberg, R.J. (1993). Procedures for Identifying Intellectual Potencial in the Gifted: A Perspective on Alternative «Metaphors of Mind». En Heller, K.A.; Mönks, F.J. y Passow, A.H. (Eds.). *International Handbook of Research and Development of Giftedness and Talent*. Oxford: Pergamon Press.

Sternberg, R.J. y Zhang, L. (1995). What do We Mean by Giftedness? A Pentagonal Implicit Theory. *Gifted Child Quarterly*, 39(2), 88-94.

Sternberg, R.J., & Davidson, J.E. (1986). *Conceptions of Giftedness*. Cambridge: Cambridge University Press.

Tejedor, F.J., García-Valcárcel, A. y Rodríguez Conde, N.J. (1994). Perspectivas metodológicas actuales de la evaluación de programas en el ámbito educativo. *Revista de Investigación Educativa*, 23, 93-127.

Tourón, J. (2000). Expanding the Talent Search in Spain. The Validation of the School and College Ability Test in Spain: Comparison of two Pilot Studies. Symposium paper presented at the 7th ECHA Conferencie. Debrece (Hungary). Agoust, 18-22.

Tourón, J. y Gaviria J.L. (2000a). *Evaluación de la educación primaria en la Comunidad Foral de Navarra*. Pamplona: Dirección General de Educación. Gobierno Foral.

Tourón, J. y Gaviria J.L. (2000b). *Evaluación de la educación primaria en la Comunidad de la Rioja*. Pamplona: Dirección General de Educación. Gobierno de la Rioja.

Tourón, J. y Reyero, M. (2000). Mitos y realidades en torno a la superdotación. En L. Almeida; E.P. Oliveira y A.S. Melo (Coords.). *Alunos sobredotados: contributos para a sua identificaçao e apoio* (pp. 19-27). Braga, Portugal: ANEIS.

Tourón, J. y Reyero, M. (en prensa). La identificación de alumnos de alta capacidad un reto pendiente para el sistema educativo. Madrid: XII Congreso de Pedagogía.

Tourón, J.; Peralta, F., y Repáraz, C. (1998). *La superdotación intelectual. Modelos, identificación y estrategias educativas*. Pamplona: EUNSA.

Tourón, J.; Repáraz, Ch. y Peralta, F. (1999). The Identification of High Ability Students: results of a detection process in Navarra (Spain). *High Ability Studies*, 10(2), 163-181.

Tourón, J.; Repáraz, C.; Peralta, F.; Gaviria, J.L.; Fernández, R.; Ramos, J.M. y Reyero, M. (2000). La validación del SCAT (*School and College Ability Test*) en Navarra: resultados del estudio piloto. En L. Almeida; E.P. Oliveira y A.S. Melo (Coords.). *Alunos sobredotados: contributos para a sua identificaçao e apoio* (pp. 81-97). Braga, Portugal: ANEIS.

Trefinger, D.J. y Feldhusen, J.F. (1996). Talent Recognition and Development: Succesor to gifted Education. *Journal for the Education of the Gifted, 19*(2), 181-193.

Van Tassel Baska, J. (1984). Appropriate Curriculum for the Gifted. En J.F. Feldhusen (Ed.). *Towards Excellence in Gifted Education*. Denver: Love.

Winner, B.J. (1971). *Statistical Principles in Experimental Designs*. New York: MacGraw Hill.

VALIDEZ INTERNA Y EXTERNA DE ALGUNOS DISEÑOS UTILIZADOS EN LA EVALUACIÓN DE PROGRAMAS

			EXPERIMENTAL		CUASIEXPERIMENTAL			PREEXPERIM.
			Diseño 1 Grupo de control pretest-posttest	Diseño 2 Grupo de control sólo posttest	Diseño 3 Pretest-posttest con grupo de control no equivalente	Diseño 4 Series cronológicas grupo único	Diseño 5 Series cronológicas grupo de control no equivalente	Diseño 6 Pretest-posttest de grupo único
			R O X O R O O	R X O R O	O X O - - - - - - - O O	OOO X OOO	OOO X OOO - - - - - - - - - - - - - OOO OOO	O X O
FUENTES DE INVALIDACION	INTERNA	Historia	+	+	+	—	+	—
		Maduración	+	+	+	+	+	—
		Administración de tests	+	+	+	+	+	—
		Instrumentación	+	+	+	?	+	—
		Regresión	+	+	?	+	+	?
		Selección	+	+	+	+	+	+
		Mortalidad	+	+	+	+	+	+
		Interacción de selecc. maduración, etc.	+	+	—	+	+	—
	EXTERNA	Interacción de admon. de tests y X	—	+	—	—	—	—
		Interacción de selección y X	?	?	?	?	—	—
		Efectos reactivos Interferencia de X múltiples	?	?	?	?	?	?
			(4)	(6)	(10)	(7)	(14)	(2)

Los números corresponden a la ordenación de Campbell y Stanley