EVALUACIÓN A GRAN ESCALA: SU DESARROLLO E INCONVENIENTES DE SU MODELO TEÓRICO

José Alfonso Jiménez Moreno
alpsic@gmail.com

Resumen

La evaluación a gran escala es una de las formas de trabajo en educación con mayor desarrollo y trascendencia, ello debido a la posibilidad que representa su uso para conocer tendencias educativas y, con ello, establecer políticas educativas, así como facilitar la posibilidad de identificar avances educativos para efecto de rendición de cuentas. Este tipo de evaluación tradicionalmente se ha desarrollado bajo el modelo de evaluación iniciado por Tyler en los años cincuenta, ya que se centra en la valoración de los logros de los estudiantes como indicador de avance educativo. La consideración de un modelo único para generar este tipo de evaluaciones ha promovido una serie de problemáticas respecto a la generación de sus resultados, así como de su uso por medio de los diversos actores relacionados con su quehacer. Frente a ello se plantea la necesidad de reflexionar respecto a la posibilidad de desarrollar evaluaciones a gran escala bajo otro modelo teórico de evaluación que delimite sus objetivos y, por ende, su metodología.

Palabras clave: Evaluación a gran escala, Modelos de evaluación.

I. Evaluación a gran escala

La evaluación a gran escala es una de las formas de evaluación educativa de mayor uso en los sistemas educativos y su característica principal radica en la observación de procesos o resultados educativos en una gran cantidad de personas. En la práctica normalmente sus resultados son utilizados como herramienta de decisión en materia de política educativa y, en algunos casos, como proveedora de información para toma de decisiones que involucran el ingreso o egreso de estudiantes a un nivel educativo en particular.

La evaluación a gran escala basa su quehacer en la teoría de la medición psicológica (Martínez, 2001), desarrollada desde el S. XVIII por Laplace y Gauss, aunque su aplicación en educación comenzó en el S. XIX en países como Alemania, Inglaterra, Estados Unidos y Francia. Es a partir del S. XX que se desarrolla la Teoría Clásica de los Tests (TCT), centrada en elementos estadísticos como confiabilidad y error de medida, lo cual impulsó la utilización de la evaluación a gran escala. Los autores principales de la TCT son Spearman y Thorndike (Martínez Arias, 1995, en Martínez, 2001). Es a partir de los años 40 del S. XX que comienza a desarrollarse fuertemente esta línea de trabajo. En 1946 Stevens realizó la clasificación de niveles de medición de variables (nominal, ordinal, intervalo y razón), lo cual potenció la implementación de métodos cuantitativos para la medición de rasgos psicológicos en las personas, así como la determinación de medición de variables sociales para el desarrollo de la estadística.

El uso de pruebas psicométricas centradas en la educación inició bajo la intención de medir el aprendizaje como rasgo psicológico, ya que éste, como se comentará más adelante, era considerado como el elemento que mejor podía reflejar los avances educativos. La evaluación a gran escala que comenzó a desarrollarse desde el S. XX, particularmente en Estados Unidos, tenía como interés identificar si los logros de los estudiantes eran acordes con los mínimos aceptables (Martínez y Santos, 2009). De ahí que la evaluación a gran escala consista en la aplicación estandarizada de pruebas a una gran cantidad de alumnos, ello con la finalidad de identificar el nivel de aprendizaje logrado por un país, región o distrito (Martínez, 2008).

La aplicación estandarizada de pruebas destinadas a medir el logro del aprendizaje bajo el respaldo de técnicas estadísticas ha favorecido la obtención de evidencia empírica objetiva, debido a la confianza que el uso de métodos objetivos trae para la toma de decisiones; por lo que el uso de pruebas de lápiz y papel[1] ha facilitado el alcance de este objetivo, ya que permiten la obtención de información y su eficiente codificación con fines de simplificación en el procesamiento de datos. En otras palabras, el uso de técnicas

[1] Se hace uso del término "pruebas de lápiz y papel" para referir a pruebas objetivas; sin embargo, en la actualidad existe una tendencia creciente del uso de pruebas objetivas estandarizadas mediante medios electrónicos. El presente trabajo no hace énfasis en esta diferencia, ya que ambas formas de presentación de las pruebas se rigen bajo los mismos estándares, salvo en elementos de aplicación y calificación, donde las pruebas computarizadas permiten acelerar estos procesos, e incluso desarrollar pruebas adaptativas.

psicométricas y estadísticas ha facilitado el desarrollo de la evaluación, ya que conjuntan un proceso objetivo con una forma eficiente de procesamiento de datos. En particular, la TCT fue de gran importancia para la medición psicológica, pero fue superada por la *Teoría de Generalización* de Cronbach, y por la *Teoría de Respuesta al Ítem* (TRI) de Lazarsfeld y Lord, la cual permite estimar las habilidades de una persona en función de un manejo probabilístico de los datos, a diferencia de la TCT, que se basaba únicamente en respuestas observadas (Martínez, 2001). Actualmente, gracias a estos trabajos ha habido un gran desarrollo técnico respecto a cómo desarrollar este tipo de evaluaciones, ya sea considerando reactivos con respuestas construidas (comúnmente llamados "reactivos de respuesta abierta"[2]), pruebas adaptativas, diseños matriciales, entre otros.

Adicional al importante desarrollo técnico en este tipo de pruebas, su uso se ha favorecido por el interés de las instituciones educativas de estandarización de contenidos y la verificación de logros alcanzados por los alumnos. Incluso, la masificación de la educación ha traído consigo la necesidad de contar con elementos objetivos y eficientes para el ingreso a diferentes niveles educativos, y es que las IES requieren de medios para asegurar que los estudiantes de recién ingreso cuentan con los conocimientos o habilidades mínimas requeridas para poder enfrentar las demandas que el nivel educativo implica, además de la necesidad de contar con mecanismos que permitan la regulación de los espacios disponibles para el estudio, cuando la solicitud de ingreso alcanza niveles altos cada año[3]. Adicionalmente, en el caso de la educación superior, las instituciones requieren de medios que favorezcan la regulación del egreso debido a la baja eficiencia de los medios tradicionales, como la tesis o los informes de investigación.

Considerando los elementos que se han descrito, los aspectos definitorios de la evaluación a gran escala están supeditados por cuestiones técnicas definidas por la psicometría y usos que los *stakeholders* dan a

[2] Cabe mencionar que el uso de reactivos de respuesta construida no implica que la evaluación a gran escala prescinda de su elemento objetivo y sistemático definitorio. Las pruebas que hacen uso de este tipo de formato de reactivos se codifican por medio de rúbricas, cuyo uso para la codificación de respuestas construidas de los estudiantes permite mantener la sistematicidad y objetividad buscada en las pruebas a gran escala

[3] Por ejemplo, los datos de solicitud de ingreso a la Universidad Nacional Autónoma de México disponibles en: http://www.escolar.unam.mx permiten conocer la enorme demanda de este tipo de servicio educativo, al punto que aproximadamente 10% de los aspirantes serán quienes alcancen un lugar para realizar sus estudios profesionales en cada uno de los ciclos en que este proceso se lleva cabo al año. Esta situación se repite en nivel medio superior en diferentes instituciones educativas y entidades federativas.

los resultados. Ahora bien, cabe aclarar una precisión terminológica; en el idioma inglés a la evaluación a gran escala se le conoce como *large scale assessment*, sin embargo, en dicho idioma existe una diferencia conceptual entre *assessment* y *evaluation*, lo cual permite diferenciar al *assesment* como forma de medición constante de un fenómeno en particular, de la *evaluation* como elemento de retroalimentación a partir de las mediciones observadas. En el castellano, se hace uso de término *evaluación* sin diferencia respecto al *assessment*, ya que no existe una traducción fiel para el término. Sin embargo, más allá del término utilizado para el fenómeno a gran escala en castellano (*large scale assessment* vs *evaluación a gran escala*), la diferenciación terminológica pareciera no explicar que la base teórico-conceptual de este tipo de estudios se basa en la mejora técnica y no en un modelo que defina su quehacer, ya que la evaluación a gran escala asume su base psicométrica y su ejercicio en pro de las decisiones que se quieran tomar con los resultados debido a la objetividad que persigue.

II. El modelo de la evaluación educativa a gran escala

El interés respecto al conocimiento del impacto que tienen las acciones de intervención sobre los avances educativos ha tenido una larga historia, pero es después de la Segunda Guerra Mundial que tuvo un auge importante, de forma particular en los Estados Unidos. El uso de tests estandarizados permitía contar con evidencia bajo una política de búsqueda de rendición de cuentas, ello gracias a la fuerte inversión que el gobierno de dicho país había hecho sobre programas educativos como consecuencia de los problemas acarreados por la Gran depresión. El enorme interés del gobierno estadounidense de tener una educación de calidad promovió que los académicos centraran su atención en buscar maneras de verificar que los programas educativos eran exitosos (Chiva, Perales y Pérez, 2009; Stufflebeam y Shinkfield, 1993).

Bajo este contexto, Ralph Tyler en su obra *Principios básicos del currículo* de 1950, desarrolló una propuesta en la que manifestaba que la evaluación debía observar la congruencia entre los objetivos de un programa educativo y los logros obtenidos. De esta forma, la evaluación de la educación comenzó a entenderse como un proceso que determina hasta qué punto los objetivos de un programa de intervención educativa se habían alcanzado, y, dada la tendencia empirista vívida en esos momentos, la verificación de logros educativos se podía obtener a través de la observación del aprovechamiento de los estudiantes.

La propuesta de Tyler fue considerada una opción viable y relevante para la identificación de logros académicos de los diversos programas educativos durante diversos años y se le comenzó a considerar un modelo de evaluación conforme inició el desarrollo de otras perspectivas evaluativas en educación. El llamado *Modelo centrado en objetivos* implica inicialmente establecer metas u objetivos educativos por alcanzar dentro de un programa o intervención educativa (Stufflebeam y Shinkfield, 1993). Cuando esto se logra lo consecuente es ordenar los objetivos en amplias clasificaciones, lo cual permite su organización, además de establecer las dimensiones que consideran el espectro de evaluación. Posteriormente es necesario definir los objetivos en términos de comportamiento, lo cual permite su mensurabilidad, dado que se operacionalizan con fines de medición. Ahora, con fines de complementar la operacionalización de los aprendizajes, se deben establecer situaciones y condiciones según las cuales puede ser demostrada la consecución de objetivos, lo cual permite estandarizar su implementación (elemento que posteriormente facilita su medición y registro). Ahora bien, para la continuación de esta labor, es necesario explicar los propósitos de la estrategia más importante en las situaciones más adecuadas a los involucrados en el programa educativo, con fines de sistematizar las acciones de intervención. Ello lleva a escoger o desarrollar las apropiadas medidas técnicas de intervención o aplicación, posteriormente, recopilar los datos de trabajo y finalmente comparar los datos obtenidos con los objetivos de comportamiento inicialmente establecidos. Es este último punto lo que cierra el proceso y permite valorar los avances educativos a la luz de la sistematización planeada, esperando así entonces que los resultados sean una especie de *espejo* respecto a los objetivos planteados.

Esta perspectiva –que le valió a Tyler el título del *padre de la evaluación educativa*– implicaba entonces considerar al aprendizaje como una modificación de conducta resultante de un programa educativo, y, por tal, involucraba la definición de objetivos curriculares en términos de comportamiento observable, establecer situaciones y condiciones según las cuales puede ser demostrada la consecución de objetivos y buscar recopilar los datos de trabajo para su comparación con lo establecido inicialmente. La propuesta de Tyler se pensó originalmente con fines de retroalimentación y no como herramienta de decisiones sumativas que respondieran a la rendición de cuentas, o bien, ligadas al otorgamiento de recursos o para el establecimiento de un ranking de desempeño; irónicamente, en la práctica se ha utilizado ampliamente con el objeto de valorar los logros finales con fines sumativos, como sucede con evaluaciones a gran

escala de ingreso a la educación media superior y superior. Incluso, fue esta perspectiva formativa lo que marcó un parte aguas dentro de la tradición de medición psicológica que le precedió a Tyler, ya que si bien la medición era un sinónimo de la evaluación, de ahora en adelante ya se pretendía también una mejora de lo analizado.

Como se mencionó previamente, dado que esta perspectiva concebía al aprendizaje como un cambio conductual producto de los programas educativos, los objetivos educativos siempre debían estar establecidos en términos operativos. Un panorama de estas características representó entonces una gran oportunidad para el desarrollo de evaluaciones, dada su relativa facilidad de implementación, lo cual propició un importante desarrollo técnico en lo que a métodos cuantitativos refiere. Sin embargo, a pesar del avance que este modelo representó, también se considera una de sus grandes desventajas, ya que los elementos relacionados con el aprendizaje que no necesariamente son operacionalizables (como la reflexión sobre un tema o el juicio individual) quedan fuera de toda valoración para fines de retroalimentación.

La tendencia centrada en medición y estadística del modelo tyleriano permitió un avance técnico importante respecto a las formas de medir el aprendizaje como evidencia de la calidad de los programas educativos, sin embargo, ya en la época de los sesenta, Cronbach (Chiva, Perales y Pérez, 2009) planteaba la necesidad de ampliar los métodos de la evaluación educativa. De acuerdo con Cronbach, la evaluación debe hacerse a lo largo de todo el proceso educativo, no solo para atender resultados, ya que es así como se puede orientar de mejora manera la práctica educativa. Asimismo, afirma que la evaluación no debe basarse en los objetivos, sino en los resultados que pretende informar. Sin duda, las críticas de Cronbach representan una innovación para la evaluación, al considerar la necesidad de ampliar las formas metodológicas en su implementación, pero, sobre todo, invita a virar el enfoque de la evaluación centrada en objetivos hacia la verificación de distintos procesos, así como una retroalimentación mucho más amplia de lo que la perspectiva tyleriana puede ofertar.

Es frente a esta creciente amplitud de los enfoques de evaluación y de diversificación metodológica que comienza una proliferación de modelos de evaluación educativa, generados básicamente por la necesidad

de retroalimentar los programas educativos en Estados Unidos y posteriormente retomados en otros países. Adicional al *Modelo centrado en objetivos* de Tyler, la literatura nos ofrece la posibilidad de considerar el *Modelo científico* de Suchman; el *Modelo CIPP (Contexto, Insumo, Procesos y Productos)* de Stufflebeam; la *Evaluación Iluminativa* propuesto por Parlett y Hamilton; la *Evaluación centrada en el consumidor* de Scriven; así como la *Evaluación comprensiva* de Stake. Para mayor referencia de estos modelos puede consultarse Stufflebeam y Shinkfield (1993). Si bien el propósito de este texto no es mostrar a detalle las características de cada uno de los modelos, baste por ahora mencionar que los distintos modelos de evaluación representan una diversidad de enfoques para llevar a cabo evaluaciones en el ámbito de la educación. En el caso de la evaluación de programas educativos ha existido una oportunidad de adoptar diversos enfoques de evaluación para su quehacer, sin embargo, en el hecho de la evaluación a gran escala pareciera haber una tendencia implícita a continuar el trabajo mediante un enfoque tyleriano, ya que su desarrollo se ha centrado únicamente en la mejora de los métodos de medición de los rasgos latentes de aprendizaje, más que en la diversificación de su evaluando como elemento de valoración del avance educativo, en este caso, el logro cognitivo de los estudiantes.

III. El uso de la perspectiva tyleriana en evaluación a gran escala

Los usos de la evaluación a gran escala por parte de la sociedad en general y por las IES pueden ser diversos, y es que considerar a la evaluación a gran escala únicamente bajo la perspectiva de sus avances en la precisión técnica de medida de constructos psicológicos puede potencializar su uso solo como un aspecto de medición. Un ejemplo de este tipo de uso es el *rankeo*, y es que hay una tendencia actual en jerarquizar a las instituciones o programas educativos en función de los resultados de los estudiantes en evaluaciones objetivas. Normalmente este tipo de juicios son descontextualizados de la enseñanza (Martínez, 2009) y no favorecen una mejora institucional, al no dar una retroalimentación de utilidad para las IES, los programas educativos o las mismas instituciones. Con tan solo enlistar los resultados de una prueba no se lleva a cabo un proceso de evaluación, sino solo una jerarquización de los resultados de una medición que, al igual que otros métodos, tiene limitaciones en la observación de un fenómeno en particular. Más aun, resulta peligroso juzgar a todo un programa educativo o institución a partir de los resultados de algunos de sus estudiantes.

Otro de los problemas que enfrenta la evaluación del egreso a gran escala es la llamada *enseñanza para la prueba* (DePascale, 2003), la cual consiste en que los educadores centran su enseñanza en los temas que se incluyen en los reactivos o temas de las pruebas estandarizadas. Una de las razones de ese problema es el hecho de que no exista un currículum unívoco para el nivel educativo la prueba a gran escala, por ende, el contenido de la prueba se vuelve un contenido de facto para la enseñanza, si se suma un énfasis particular en la rendición de cuentas o en la intención de rankeo como objetivo primario, el problema se incrementa. Este problema está cada vez más en aumento en México, debido a la importancia sumativa que han tenido las evaluaciones a gran escala, sobre todo en cuanto a ingreso o egreso se refiere, baste con revisar la oferta de cursos de preparación para pruebas a gran escala (en especial para el ingreso a instituciones de educación superior, pero también para pruebas con fines de acreditación de conocimientos). Éste es un problema más que se genera debido a la consideración de la evaluación como elemento de castigo/recompensa y no como factor de retroalimentación/mejora.

En este mismo sentido, Volante (2006) presenta un resumen de las mayores críticas de la evaluación a gran escala en los últimos años; entre ellas se encuentran las que a continuación se mencionan. Kane (2002) y Popham (2001) afirman que las pruebas de este tipo pueden tener limitaciones psicométricas importantes y aun así ser la base de políticas educativas; y es que hay que reafirmar el hecho de que no hay prueba perfecta, siempre habrá grados de error de medida, errores en el levantamiento de datos o incluso en calificación. Según la Canadian Teachers Federation (2009, en Volante, 2006), la evaluación a gran escala actúa como una forma de control que va en contra de la autonomía de los educadores, ya que al estar alejada de la enseñanza la orilla a centrarse únicamente en los contenidos de las pruebas (como en el fenómeno de enseñanza para la prueba). Por su parte, la American Education Research Association (2000) y Darling-Hammond (2004), ambos en Volante, (2006) critican el hecho de que la evaluación a gran escala puede orillar a que se mantenga a los docentes como responsables de los resultados, ello a partir de considerar el resultado de los estudiantes como único indicador de la calidad de programas educativos, y como consecuencia, a contribuir a que tengan sentimientos de baja eficacia. McNeil (2000) y Valencia y Villareal (2003), ambos en Volante, (2006), aseguran que la evaluación a gran escala aumenta brechas entre clases sociales ignorando elementos importantes que tienen gran influencia dentro de resultados educativos, como el nivel socioeconómico, lengua de origen, así como su estado físico o emocional,

tratando las condiciones de todos los estudiantes de la misma manera. Por otra parte, un pensamiento que comparten Anderson y Postl (2000), Glaser, Linn y Bhornstedt (1997), Hess y Bringham (2000), así como Wideen, O'Shea, Pye e Ivany (1997), todos ellos en Volante, (2006) es que este tipo de evaluación limita a la educación penalizando el pensamiento crítico y creativo, así como el trabajo intelectual en general, dado que los estándares de evaluación los excluyen como parte de los elementos de análisis. También puede considerarse que la evaluación a gran escala modifica el currículum provocando que se enseñen estrategias orientadas solo a responder adecuadamente las pruebas, lo cual limita el tiempo a otras habilidades que no son evaluadas, en particular cuando los resultados son de alto impacto o de índole únicamente sumativa (Kohn, 2002; Smith y Fey, 2000 y Volante, 2004; en Volante, 2006). Asimismo, si la evaluación a gran escala solo se utiliza con fines de rankeo o de certificación, la información que proveen resultará de poca utilidad para la práctica en el aula (Stiggins, 2002 y Falk, 1998; en Volante, 2006) e incluso dañar el autoconcepto en los evaluados (Meaghan y Casas, 2001 y Miller y Tovey, 1996; en Volante, 2006).

Todas estas perspectivas ponen de manifiesto la existencia una fuerte tendencia de crítica a la evaluación a gran escala, y es que cuando se concibe al aprendizaje solamente como un constructo que se puede decantar en mínimas expresiones, la única opción viable para su observación es el uso de pruebas objetivas (Martínez, Cepeda, Fuentes y Burgos, 1995). La evaluación debe considerar la vinculación entre el discurso científico, el objeto y la matriz social. Debe entonces de buscar conocer las características del aprendizaje y dar una explicación del por qué se da de esa forma. La cuestión se agrava cuando incorporamos estas prácticas a la evaluación del egreso de la formación profesional, y es que "al volver de alto impacto pruebas concebidas con propósitos fundamentalmente formativos se corre un riesgo muy grande de que se produzcan consecuencias indeseables, que ya han comenzado a presentarse, como el que los maestros enseñen para la pruebas o que algunas escuelas usen los resultados para campañas mercadotécnicas sin sustento sólido" (Martínez y Santos, 2009, pág. 273). Debe cuidarse que el contenido de las pruebas a gran escala sirva de eje para la actividad docente con fines de obtener buenos resultados en ellas.

En este mismo sentido, la apreciación de la AERA (2000) y de Darling-Hammond (2004), ambos citados en Volante (2006), es que uno de los mayores problemas de la evaluación a gran escala pareciera ser la imposición de repercusiones a partir de los resultados obtenidos, ya que puede tener un efecto adverso para la educación, problemática que se denota en las evaluaciones del egreso en México, las cuales condicionan la terminación de estudios profesionales y en algunos casos son utilizadas como forma de titulación.

Conclusiones

Los problemas descritos, así como la gama de modelos de evaluación que pueden dar sustento a las metodologías que se desarrollan con el fin de retroalimentar la práctica educativa mostrados con anterioridad, permiten valorar la necesidad de repensar el modelo de evaluación a gran escala que se ha utilizado; y es que los problemas actuales de este tipo de evaluación pueden resumirse de la siguiente manera: 1) la falta de identificación de las formas de evaluación a gran escala del egreso respecto al modelo de evaluación que sustenta su desarrollo y uso. 2) La consideración única del modelo tyleriano como forma para evaluar a gran escala el egreso de la formación profesional, lo que conlleva a retroalimentar a los planes de estudios profesionales considerando solo los resultados de los estudiantes en una prueba como único elemento de calidad. 3) Hacer un uso excesivo de la evaluación a gran escala del egreso como herramienta de toma de decisiones de alto impacto contribuye a modificar el sentido tyleriano de retroalimentación de la formación para convertir a este tipo de evaluación en un premio o ranking. 4) Limitar a que la evaluación del egreso se centre solamente en la revisión de logros disciplinares y no a elementos que no sean producto del análisis, creatividad y reflexión de los estudiantes. 5) Que este tipo de evaluación, en potencia de gran relevancia para la retroalimentación de la formación terciaria, se confunda con solo llevar a cabo mediciones del aprendizaje.

Frente a este problema se presenta la necesidad de valorar la posibilidad de realizar la evaluación a gran escala bajo un modelo o enfoque diferente. No se debe perder de vista que las evaluaciones a gran escala surgieron con un objetivo de rendición de cuentas de un sistema o estado, lo cual no es lo mismo que dar seguimiento al desempeño de estudiantes en lo individual (DePascale, 2003). Se debe considerar un modelo de trabajo mayormente enfocado en la retroalimentación que permita entender de mejor manera el

sistema con los fines de: mejorar la educación, determinar el logro de estudiantes y dar retroalimentación relevante a los *stakeholders* (estudiantes, profesores, tomadores de decisiones y la comunidad en general) (DePascale, 2003).

Referencias bibliográficas

- Chiva, I., Perales, M. y Pérez, A. (2009). Historia de la evaluación educativa. En: Jornet, J. y Leyva, Y. (Comp, 2009). *Conceptos, metodología y profesionalización de la evaluación educativa*. México: INITE.
- DePascale, (2003). The ideal role of large-scale testing in a comprehensive assessment system. *National Center for the Improvement of Educational Assessment*. 5. (1).
- Martínez, F. (2001). Evaluación educativa y pruebas estandarizadas. Elementos para enriquecer el debate. *Revista de la educación superior*. 30 (4). 71-85.
- Martínez, F. (2008). Evaluación a gran escala y evaluación en el aula. *Los usos de Excale para la mejora educativa*. México: INEE.
- Martínez, F. (2009). Evaluación formativa en el aula y evaluación a gran escala: hacia un sistema más equilibrado. *Revista electrónica de investigación educativa*. 11 (2).
- Martínez, F. Cepeda, B. Fuentes, G. Burgos, R. (1995). Evaluación del aprendizaje. En: Rueda, M. (Coord.). *Procesos de enseñanza y aprendizaje I.* México: COMIE.
- Martínez, F. y Santos, A. (2009). Evaluación e investigación educativa. En: Alba, A. (Ed.) *¿Qué dice la investigación educativa?* México: COMIE.
- Stufflebeam, D. y Shinkfield, A. (1993). *Evaluación sistemática. Guía teórica y práctica*. Madrid: Paidós.
- Volante, L. (2006). An alternative vision for large-scale assessment in Canada. *Journal of teaching and learning* 4. (1).