



מבחני הערכה בין-לאומיים בחינוך - קריאה לבדיקה והצעות לפעולה

מאת ג'ודית ד' זינגר והנרי ' בראון

1 הטקסט שלהלן תורגם ועובד לקראת מפגש למידה עם פרופ' הנרי בראון בנושא מדידה והערכה ככלי לשיפור החינוך. לפרסום המלא ראו:

Singer, J.D. & Braun, H. I. (2018). [Testing international education assessments](#). *Science* 360 (6384), 38-40.

Reprinted with permission from The American Association for the Advancement of Science (AAAS). This translation is not an official translation by AAAS staff, nor is it endorsed by AAAS as accurate. In crucial matters, please refer to the official English-language version originally published by AAAS.

אומנם דירוגי המבחנים זוכים לכותרות בעיתונים, אך לעיתים קרובות הם מטעים. במאמר זה מוצעות דרכים לשימוש מושכל בתוצאות המבחנים למען שיפור מערכת החינוך

תבות בעיתונים המדווחות על מבחני הערכה בין-לאומיים רחבי היקף בחינוך (International Large Scale Education Assessments, להלן ILSA), דוגמת מבחן פיז"ה, נוטות להדגיש את ביצועי המדינה שבה הן מתפרסמות. בתוך כך הן נוהגות לערוך השוואות בינה לבין המדינות בעלות הציונים הטובים ביותר (ובשנים האחרונות מדובר, בדרך כלל, במדינות מזרח אסיה). דירוגים נמוכים או במגמת ירידה לפעמים מדאיגים כל כך את קובעי המדיניות עד שהם קופצים לפתרונות בזק - לעיתים קרובות חסרי יסוד - המתבססים על "הרכיב הסודי" שאחראי לדעתם לביצועים הטובים ביותר. כסטטיסטיקאים החוקרים את השיטות להערכה בין-לאומיים רחבת היקף בחינוך ואת השימוש בה בקביעת מדיניות,² אנו מאמינים שהמרדף אחר הדירוגים הללו - והניסיונות הבלתי נמנעים לחקות מאפיינים ייחודיים של המערכות בעלות הביצועים הטובים ביותר - אינו רק מטעה, אלא גם מסיט את תשומת הלב מהשימושים המועילים יותר בנתוני המבחנים. בהמשך נדגיש את הסכנות הטמונות בהתוויית כללי מדיניות נוקשים על סמך נתונים מצרפיים (aggregated data); נדגים את היתרונות שבבחינה מעמיקה יותר של נתוני ההישגים במבחני ILSA - הן בין מדינות שונות, הן בין קבוצות שונות בכל מדינה ומדינה; ונציע כמה הצעות קונקרטיות לשיפור מבחני ILSA בעתיד.

למרות הביקורת שאנו מותחים על המבחנים, העלויות הגבוהות של ביצוע המבחן והמסקנות השגויות לפרקים של קובעי המדיניות, איננו חושבים שיש לזנוח את מבחני הערכה הבין-לאומיים בחינוך, דוגמת מבחן פיז"ה, מבחן TIMSS למתמטיקה ומדעים ומבחן PIRLS לבדיקת אוריינות הקריאה. מבחני ILSA מספקים מסגרת התייחסות ייחודית להבנה של התוצאות ושל הקשרים במדינות השונות. על כן יש להם ערך רב בגיוס המניע הפוליטי של מדינות להשקיע משאבים בתחום החינוך. כמו כן נתוני ILSA שימשו בסיס למאות ניתוחים שניוניים (secondary analyses) שנגעו בשאלות חשובות על מדיניות חינוכית. עם זאת, כדי לממש את ההבטחה הגלומה במבחנים, יש לערוך שינויים בפרשנות, בהפצה ובניתוח שלהם, וכן באסטרטגיות המשמשות לתכנון ההערכות העתידיות.

מנתחים את ההצלחה המזרח אסייתית

בשנת 2012 זכו שבעה תחומי שיפוט לציונים הממוצעים הגבוהים ביותר במתמטיקה במבחני פיז"ה: שנגחאי, סינגפור, הונג קונג, טיוואן, דרום קוריא, מקאו ויפן. לפני שנסיק כי אימוץ כל אחד מהמאפיינים של החינוך במזרח אסיה - כגון ספרי לימוד ברמה גבוהה מאוד או "פתרון השעה" שאימצה בריטניה לאחרונה³ - יוביל לשיפור ציוני המבחן של תלמידים בכל מקום אחר, כדאי לחשוב על הדברים הבאים.

המדגמים של מבחני ILSA אינם מייצגים בהכרח את קבוצת הגיל (או השכבה) הרלוונטית בכל תחום שיפוט. אוכלוסיית היעד של מבחן פיז"ה מוגדרת כ"בני 15 הרשומים במוסדות החינוך ללימודים מלאים". עד לאחרונה התקיימה בסין מערכת רישום אוכלוסין פנימית, שאסרה על מהגרים מהאזורים הכפריים להירשם לבתי ספר עירוניים. בשנת 2014 הודה הארגון לשיתוף פעולה ולפיתוח כלכלי (OECD), העורך את מבחני פיז"ה, כי המדגם בשנגחאי לשנת 2012 החריג 27% מבני ה-15 במדינה.

2 Singer, J.D., Braun H. I. & Chudowsky, N. (2018). *Methods and Policy Uses of International Large-Scale Assessments: Report of a Workshop*. Washington, DC: National Academy of Education.

3 <https://www.gov.uk/government/news/south-asian-method-of-teaching-maths-to-be-rolled-out-in-schools>.

לשם השוואה, בארה"ב מדובר על כ-11% בלבד.⁴ המדינות המפותחות פחות ב-OECD, דוגמת מקסיקו וטורקיה, נתקלות בבעיות דומות כיוון שעד הגיעם לגיל 15, כ-40% מהילדים כבר נושרים מבית הספר. לא ידוע מלוא השפעתן של החרגות אלה מן הממוצעים ברמה הארצית, אך סביר להניח שהתלמידים המוחרגים מגיעים מהאחוזונים הנמוכים בהתפלגות ההישגים.

התוצאות מערים בודדות (כגון שנגחאי), מערי מדינה (כגון סינגפור) או ממדינות בעלות מערכות חינוך לאומיות (כגון צרפת) אינן ניתנות להשוואה עם מדינות שבהן קיימות מערכות חינוך מבוזרות (דוגמת ארה"ב, קנדה וגרמניה). הסיכומים ברמה הארצית חסרי משמעות כמעט למערכות מבוזרות, כיוון שהם מסתירים את הרבגוניות הגדולה של כללי המדיניות והפרקטיקות בתוך המדינה. בשנת 2015, למשל, השתתפה מדינת מסצ'וסטס במבחן פיז"ה כתחום שיפוט נפרד. לא נמצא הבדל מובהק בין ציון הקריאה הממוצע במדינה לציונים במדינות מזרח אסיה בעלות הביצועים הטובים ביותר; ציון המתמטיקה הממוצע היה בינוני יותר, אך אם המדינה, שהיא חלק מהפדרציה האמריקאית, הייתה מדורגת ככל מדינה עצמאית אחרת, היא הייתה מגיעה למקום ה-12.⁵

מספר הגורמים שביכולתם לנבא באופן מהימן את ציוני המבחנים של תלמידים הוא גדול מאוד. אף על פי שפריסת מבחני ILSA מתרחבת, רובם נערכים ב-50 עד 75 מדינות או תחומי שיפוט. מומחים בתחום החינוך מציעים 50 עד 75 גורמים מנבאים מהימנים לציוני המבחן ברמה הארצית: ספרי לימוד ברמה גבוהה מאוד? הכשרת מורים? שיטות הוראה? השפעת העמיתים לספסל הלימודים? בגלל ריבוי הגורמים המנבאים האפשריים בכל מדינה ובכל ניתוח, אין זה אפשרי להסיק אם מאפיין מסוים של מערכת חינוך - אפילו מאפיין בעל מתאם גבוה לציוני המבחנים הממוצעים ברמה הארצית - מסביר באופן מוחלט את ההבדלים בביצועי התלמידים.

ציוני המבחנים מושפעים גם מגורמים רבים מחוץ לכותלי בית הספר, ולכן תהיה זו טעות להתייחס לציוני המבחנים כמחוננים (אינדיקטורים) בעלי תוקף מוחלט לאיכות מערכת חינוך כלשהי. לדוגמה, במזרח אסיה, כבכל מקום אחר, נפוץ השימוש במורים פרטיים. קוריאה היא הדוגמה הבולטת ביותר לכך: בשנת 2012 דיווחו כמחצית ממשתתפי מבחן פיז"ה כי קיבלו שיעורים פרטיים שהתמקדו לעיתים קרובות בהכנה למבחן. בסך הכול, ההוצאה על שיעורים פרטיים הייתה 2.6% מהתמ"ג, שנוספו ל-3.5% שתרמה הממשלה לתחום החינוך.⁶ אחד הפירושים הסבירים לתוצאות של קוריאה ותחומי שיפוט אחרים במזרח אסיה במבחני ILSA הוא שלא מדובר בתוצאות של מערכות החינוך הציבוריות, אלא בתוצאותיה של השקעה פרטית ניכרת.

המוטיבציה של התלמידים להשתתף במבחני הערכה בסיכון נמוך משתנים ממדינה למדינה. לציונים במבחני ILSA אין כל השפעה על אנשים בודדים, ולכן יש לצייד את התלמידים במוטיבציה להצליח ככל יכולתם. סביר להניח שחלק מהסיבה לכך שתלמידים מתרבויות מזרח אסיה מקבלים ציונים גבוהים יותר היא שהם מורגלים להשיג את הביצועים הטובים ביותר בכל מבחן שהוא, גם במבחנים בעלי סיכון נמוך.

הדירוגים נגזרים מהממוצעים ברמה הארצית, ורווחי הסמך (confidence intervals) המתאימים רחבים, כך שיתכן שלא יימצא הבדל מובהק בין מדינות בעלות דירוגים שונים באופן ניכר. לדוגמה, במבחני פיז"ה לשנת 2015 הגיעה קנדה למקום העשירי במתמטיקה, אך רווח הסמך שלה, העומד על 95%, חופף לזה של קוריאה (שדורגה במקום השביעי) ושל גרמניה (שדורגה במקום ה-16).⁷ עם הזמן, הדירוגים מושפעים גם מתחומי השיפוט המשתתפים באחד ממבחני ILSA, ולכן דירוג המדינה עשוי להשתנות גם אם הביצועים נותרים על כנם.

4 Strauss, V. (March 20th, 2014). *So how overblown were No. 1 Shanghai's PISA results?*. *The Washington Post*. Retrieved from: <https://www.washingtonpost.com/>

5 Schleicher, A. (2016). *Programme for International Student Assessment (PISA) Results from PISA 2015 – Massachusetts*. OECD.

6 Park, H., Buchmann, C., Choi, J., & Merry, J. J. (2016). Learning beyond the school walls: Trends and implications. *Annual Review of Sociology*, 42, 231-252.

7 National Center for Education Statistics, Program for International Student Assessment (PISA): Mathematics Literacy: Average Scores. Retrieved from: https://nces.ed.gov/surveys/pisa/pisa2015/pisa2015highlights_5.asp

גם אם אף אחת מהביקורות שלעיל לא יכולה לעמעם את החשיבות של הממצאים ממדינות מזרח אסיה, עדיין מערך החששות הזה מעמיד בסימן שאלה כל מסקנה פשטנית המבוססת על הדירוגים. אולם כל עוד תוצאות מבחני ILSA מדווחות בעיקר בטבלאות של הישגים בין-לאומיים, קובעי המדיניות במדינות בעלות הביצועים הגרועים או המידרדרים נוטים לאמץ "פתרונות קסם" אחידים. פתרונות אלו מבוססים על נתונים מצרפיים, לרוב בגלל תערוכת של לאומנות, חששות בדבר תחרותיות גלובלית והטבע האנושי.

עוד הזדמנויות מבטיחות לחקירה מעמיקה

הבטחה רבה גלומה בניתוחים של נתוני ILSA המפורקים לרמות נמוכות יותר מהרמה הארצית, אם מדובר באזור הגאוגרפי, במחוז או במדינה, או אם מדובר במאפייני בית הספר או התלמיד. הניתוחים הטובים ביותר משלבים מקור נתונים שני, הנפוץ כעת במבחני ILSA: שאלוני רקע שממלאים התלמידים, ההורים, המורים או המנהלים.

ג'ריס (Jerrim)⁸ למשל, השתמש בנתוני מבחן פיז"ה לשנת 2012 מאוסטרליה כדי לחקור את סיפורי ההצלחה של מדינות מזרח אסיה. ג'ריס מצא כי תוצאות המבחנים של תת-הקבוצה הכוללת את בני הדור השני למהגרים ממדינות מזרח אסיה היו דומות לתוצאות של בני מזרח אסיה בארצות המוצא שלהם. זאת אף על פי שבני המהגרים למדו בבתי ספר באוסטרליה, שזכו באופן כללי לצינונים נמוכים יותר. למרות ההכרה בכך שמהגרים אינם מהווים מדגם אקראי של אינדוידואלים ממדינת המקור, הניתוח מראה כי לגורמים משפחתיים יש השפעה רבה על ההבדלים בין התוצאות במבחני ILSA, וממחיש כמה תובנות הנובעות מניתוחים נוספים שנערכו בתוך המדינה.

כדי לערוך ניתוח בתוך המדינה, כמובן, אין צורך בנתוני ILSA. ההבטחה הגלומה בנתוני ILSA נובעת מהשוואת ניתוחים תוך-מדינתיים - בין מדינות שונות. לדוגמה, שמידט ומקנייט (Schmidt and McKnight)⁹ מצאו כי מדינות שכוללות כמות קטנה יותר של חומר בתוכנית הלימודים שלהן משיגות ביצועים נמוכים יותר במבחני TIMSS בנושאים אלה מאשר מדינות שבהן תוכנית הלימודים כוללת תוכן נרחב יותר. כך למשל, ארצות הברית אינה שמה דגש על מדעים מדויקים בתוכנית הלימודים לפני הכניסה לתיכון, והשיגה ביצועים נמוכים יותר מקוריאה, אשר דיווחה על תוכנית לימודים רחבה יותר. ניתוחים שניוניים כאלה אומנם שופעים מידע, אך לעיתים נדירות בלבד הם זוכים להשפעה הדומה לזו של טבלאות הישגים הבין-לאומיים. זאת משום שתוצאותיהם מתפרסמות בדרך כלל זמן רב לאחר פרסום הנתונים הראשוניים - לפעמים אפילו אחרי שפורסמו דירוגי הסיבוב הבא של מבחני ILSA.

עוד אתגר אנליטי עיקש הוא הצורך לבנות מחוונים של מאפייני רקע שהמהימנות והתוקף שלהם אינם מושפעים מההבדלים בין מדינות ותרבויות. יצירת מדדים מורכבים בני השוואה - כגון רקע סוציו-אקונומי - היא אפילו קשה יותר. מחוון הרקע הסוציו-אקונומי במבחן פיז"ה, המדד הסוציו-אקונומי הבין-לאומי למצב תעסוקתי (International Socio-Economic Index of Occupational Status), מתמודד עם מכשולים מרובים. ביניהם ניתן למצוא תפיסות מגוונות של המצב התעסוקתי של ההורים (כפי שהמצב היחסי של מהנדסים, רופאים ומורים משתנה בין מדינות), והשאלה האם הפריטים בו (למשל, בעלות על שולחן כתיבה בבית או טלפון סלולרי) באמת מתחברים למשתנה בסיסי המשותף לכל המדינות.¹⁰

8 Jerrim, J. (2015). Why do East Asian children perform so well in PISA? An investigation of Western-born children of East Asian descent. *Oxford Review of Education*, 41, 310-333.

9 Schmidt, W. H. & McKnight, C. C. (1998). What can we really learn from TIMSS? *Science*, 282, 1830.

10 Singer, J.D., Braun H. I. & Chudowsky, N. (2018). *Methods and Policy Uses of International Large-Scale Assessments: Report of a Workshop*. Washington, DC: National Academy of Education.

למרות האתגרים בהשוואה בין תרבויות, מצאנו שלושה סוגים של ניתוחים תוך-מדינתיים הניתנים להשוואה בין מדינות. הם טומנים בחובם הבטחה גדולה - גם כשאנם מלמדים על סיבתיות, ולרוב אינם עושים זאת - כיוון שהם מעלים היפותזה מעניינת שראויה למחקר מעמיק יותר.

מידוד (benchmarking) קשרים תוך-מדינתיים בין מדינות בעלות תרבויות דומות.

שימו לב לשתי השוואות ממחישות של ציוני המתמטיקה במבחן פיז"ה לשנת 2012: (1) הציונים הממוצעים של הונג קונג וטייוואן דומים מאוד, אולם חוזק הקשר בין תוצאות התלמידים לבין הרקע הסוציאו-אקונומי גבוה פי שלושה בטייוואן מאשר בהונג קונג; (2) הממוצע בקנדה גבוה ב-37 נקודות מהממוצע בארה"ב, אולם הקשר בין התוצאות לבין הרקע הסוציאו-אקונומי חלש הרבה יותר בקנדה מאשר בארה"ב. אנו מאמינים שלמידוד ניתוחים כאלה יש סיכויים טובים יותר להוביל מדינות שמערכותיהן הוגנות פחות (במקרה זה, טייוואן וארה"ב) להתנסות עם האסטרטגיות ששכנותיהן משתמשות בהן כדי לשפר את ההוגנות בחינוך, וזאת לעומת טבלאות ההישגים הבין-לאומיים [המתפרסמות לאחר מבחני ILSA].

שימת דגש על התפלגות הישגים בתוך המדינות.

בשל המגבלות שבהשוואת ממוצעים, נוספו לטבלאות ההישגים הבין-לאומיים של מבחני ILSA שיעורי התלמידים שקיבלו ציונים גבוהים במיוחד או נמוכים במיוחד, בכל תחום שיפוט. לדוגמה, ציוניהם של 9% בלבד מהתלמידים בארה"ב הגיעו לשתי הקטגוריות העליונות במתמטיקה במבחן פיז"ה לשנת 2012. לשם השוואה, חמש מדינות השייכות ל-OECD ושישה תחומי שיפוט אחרים שהשתתפו במבחנים רשמו שיעור של יותר מפי שניים.¹¹ אף שהשיעורים הללו אינם מדויקים (כפי שהסברנו קודם לכן, בנוגע לממוצעים ברמה הארצית), הבדלים בהיקף כזה עשויים להוביל את ארה"ב לשקול התנסות בגישות שונות כדי להציע לתלמידים בעלי הישגים גבוהים הזדמנויות למידה מאתגרות באופן הראוי להם.

בחינת הניסויים הטבעיים (natural experiments) שעושה כל מדינה בהשוואה

למדינות אחרות. ניסויים טבעיים מספקים בסיס מהימן להיסק על סיבתיות כאשר החוקרים יכולים לטעון באפקטיביות שהקצאת האנשים האינדיווידואליים ל"טיפולים" היא אקראית בקירוב. הניסויים הטבעיים המשתמשים בנתוני ILSA הולכים צעד קדימה ובוחרים - באמצעות תוצאה משותפת - אם ההשפעות של טיפולים דומים ניכרות בתחומי שיפוט מרובים. בדארד ודיואי (Bedard and Dhuey),¹² לדוגמה, בחנו את ההשפעה של גיל כניסת הילדים לגן על הישגי התלמידים. התלמידים הצעירים יותר בשכבת הגיל - בכיתות ד' ו-ח' כאחד - השיגו ציונים נמוכים באופן ניכר, בממוצע, מעמיתיהם הבוגרים יותר. תוצאות אלה ניכרו במגוון רחב של מדינות והראו כי למרות התקווה שההבדלים הנובעים מגיל הילדים בשכבה ייעלמו עם הזמן, הם ממשיכים גם בגיל ההתבגרות. לפיכך יש לחקור לעומק את ההשפעות לטווח ארוך של מדיניות הכניסה לבית הספר על הישגי התלמידים, וייתכן שיש להתנסות במדיניות חליפית. אפשר למשל לאפשר להורים להמתין עוד שנה לפני רישום ילד לבית הספר, שלפי תאריך לידתו עלול להיות הצעיר בשכבה.

חמש הצעות לשיפור

מבחני ILSA השתפרו מאוד מאז מבחן המתמטיקה הבין-לאומי הראשון שנערך בשנת 1964: המבחנים טובים יותר; האקוויוולנטיות (שקילות) בין תרבויות זוכה להכרה ועולה - גם אם לא לגמרי - בסדר העדיפויות; ניהול המבחנים ושיטות הציון (scoring) התעדכנו; ואסטרטגיות הדגימה והגישות האנליטיות הוגברו. אולם עדיין יש הזדמנויות לשיפור. אנו מציעים חמש הצעות קונקרטיות, שלדעתנו צפויות להניב חוזרים גבוהים יחסית לעלויות הגבוהות שקשורות לכל הצעה.

11 OECD (2013). *PISA 2012 Results: Excellence through Equity: Giving Every Student the Chance to Succeed (Volume II)*. Paris: OECD Publishing.

12 K. Bedard, E. Dhuey, Quart. J. Econ.121, 1437 (2006).

4. הוספת רכיבי אורך למערכים של מחקרי החתך הקיימים תועיל גם היא. למידה

היא תהליך, לא מצב קיים. כדי להשוות בין ביצועי תלמידים ולזהות באופן מהימן את הגורמים המנבאים אותם, הן בתוך מערכת חינוך אחת הן בין מערכות חינוך שונות, חובה להשתמש במחקרי אורך.¹⁸ מדינות רבות, שגילו את יתרונותיהם של נתוני אורך, החלו לעקוב אחר ביצועי תלמידיהן: חלקן מתבססות על מדגמים מתוך מבחני ILSA (לדוגמה, דנמרק ושווייץ עוקבות אחר משתתפי מבחן פיז"ה), ואחרות מתמקדות בנתוני אורך מקומיים (כגון מחקר האורך הנוגע לגיל הרך, המתקיים בארה"ב, ומחקר עוקבה לאומי בתחום החינוך, המתקיים בגרמניה). אף שניתן, בעיקרון, לבדוק את הקשר בין מבחנים ייחודיים למדינה לבין מבחני ILSA במדינות אחרות כדי לקבל נתונים ספציפיים הנוגעים לגיל או לציונים, הקשרים התגלו כחלשים מכדי לתמוך בניתוח חוצה מדינות של נתוני אורך תוך-מדינותיים. כמובן, יש כמה התנגדויות מתקבלות על הדעת להצעה זו: (1) מעקב אורכי הוא יקר; (2) ישנה בעיה רצינית של נשירה;¹⁹ (3) לא בטוח שהידע המתקבל מכך מצדיק את המחיר. אולם הדרך היחידה לטפל בטענות אלה היא לנסות כמה תוכניות הרצה (פיילוט) ולהיעזר בניסיון שמופק מהן כדי לאמוד טוב יותר את העלויות לעומת היתרונות. אנו מאמינים שהמעבר למבחנים ממוחשבים עשוי להקל על חוקרים לבצע מחקרי אורך ולהפוך אותם למשתלמים יותר.

אחת החלופות שזכתה להצלחה מסוימת היא מעקב אחר שכבת גיל ספציפית באמצעות דגימה אקראית (משתנה) מתוך שכבת הגיל לאורך זמן, כדי לבנות נתונים "דמויי עוקבה". לדוגמה, שכבת הגיל של בני ה-15 שנבחנו במבחני פיז"ה נכללת כעבור עשור בתוך אוכלוסיית PIAAC של בני ה-20 עד 25. גישה דומה לכך היא להשתמש בשיטת "הפרש הפרשים" (Differences-in-differences) כדי לאתר סיבתיות משוערת בנתונים כגון אלה. הרעיון המרכזי הוא לקשר ברמה הארצית בין השינויים לאורך זמן בגורמים המסבירים לבין השינויים המתאימים בביצועי המבחנים.

5. יש להשתמש בממצאים מניתוחים של נתוני ILSA כדי לעודד ניסויי שדה אקראיים

הבוחנים את השפעותיהן של תוכניות התערבות ספציפיות; לא כדי לשפר את מבחני ILSA, אלא כדי לבחון את הצעדים שיש לנקוט לאחר שניתוח נתוני ILSA מעלה את הצורך בהתערבויות פוטנציאליות במדיניות החינוכית. מבין חמש ההצעות שלנו, אנו מאמינים כי הצעה זו צפויה להניב את השיפור הגדול ביותר במדיניות החינוך. ה-OECD והארגון הבין-לאומי להערכת הישגים בחינוך (IEA) ערכו ניסויי שדה אקראיים וחקרו את השפעותיהם של מאפיינים שונים של מבחני ILSA (כגון אופני היבחנות). עם זאת, למיטב ידיעתנו, טרם עלתה ההצעה להשתמש בתוצאות של ניתוחי ILSA - שבמקרה הטוב, מעלות כמה שאלות מעניינות - כבסיס לניסויי שדה אקראיים, במטרה לבחון בקפידה את ההשפעות המשוערות של סוגי מדיניות שזוהו באמצעות ניתוח נתוני ILSA.

למרות שלל בעיותיהם, מבחני ILSA אינם צפויים להיעלם, וטוב שכן. הם יכולים לשמש כלי עוצמתי לשינוי, אף שטבעם ומידת השפעתם משתנים בכל מדינה ומדינה ולאורך זמן.²⁰ כדי לממש את ההבטחה הגלומה בהם, טוב תעשה קהילת ILSA אם תתחיל לתכנן מחקרים וניסויים שייעזרו באסטרטגיות שהצענו ובאסטרטגיות נוספות אחרות.²¹

תרגום מאנגלית: לירון רובינס | עריכת תרגום: מיכל ויזל | עיצוב גרפי: אמונה כרמל

18 Singer, J. D. & Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford University Press.

19 במקור: attrition. המונח מתייחס לחשש שנבחים בנקודת זמן אחת לא יבחנו מסיבות שונות גם בנקודת הזמן הבאה.

20 Ritzen, J. (2013). International Large-Scale Assessments as Change Agents. In: M. Von Davier, E. Gonzalez, I. Kirsch & K. Yamamoto (eds), *The Role of International Large-Scale Assessments: Perspectives from Technology, Economy, and Educational Research*. Dordrecht: Springer.

21 Singer, J.D., Braun H. I. & Chudowsky, N. (2018). *Methods and Policy Uses of International Large-Scale Assessments: Report of a Workshop*. Washington, DC: National Academy of Education.