

## קביעת סטנדרטים במערכת החינוך

צור קרליץ

פברואר 2012



מרכז ארצי לבחינות ולהערכה (ע"ר)  
NATIONAL INSTITUTE FOR TESTING & EVALUATION  
المركز القطري للامتحانات والتقييم  
מיסודן של האוניברסיטאות בישראל

**דוח מרכז 378**  
**ISBN:978-965-502-162-2**

## **קביעת סטנדרטים במערכת החינוך**

**ד"ר צור קרליץ, מרכז ארצי לבחינות ולהערכה**

נכתב עבור קורס 14021 "מדידה בחינוך : מושגי יסוד" של האוניברסיטה הפתוחה

## **תמצית**

סטנדרטים הינם אמות מידה להערכה של איכות. במערכות חינוכיות משתמשים בסטנדרטים משיקולי מדיניות ואחריותיות ציבורית. הסטנדרטים משמשים להערכת מרכיבי המערכת בכוונה לשפר את איכות החינוך ולסייע בקבלת החלטות בנוגע לתוכנית לימודים, משאבי בית הספר, הישגי התלמידים, יעילות המורים וכדומה.

קיימים סוגים שונים של סטנדרטים כגון סטנדרטים של תוכן המגדירים את הרכב תוכנית הלימודים וסטנדרטים של ביצוע המגדירים קריטריונים להצלחה בלימודים כפי שהם באים לידי ביטוי בביצוע במבחן. למעשה, סטנדרט ביצוע הוא הציון שתלמיד נדרש לקבל במבחן על מנת להראות רמה מסוימת של שליטה בחומר ובכישורים הנלמדים.

דוח זה עוסק בשימוש סטנדרטים במערכת החינוך ומתמקד בעיקר בשיטות לקביעת סטנדרטים לביצוע במבחנים. תחילה נדון בסטנדרטים באופן כללי ולאחר מכן נסקור את סוגי הסטנדרטים הקיימים במערכות חינוכיות. בהמשך, נדון בהגדרת תהליך קביעת הסטנדרטים לביצוע ומושגי יסוד רלוונטיים ונציג שיטות שונות לקביעת ציוני חתך במבחנים. לסיכום נדון בבקרת איכות ותיקוף של סטנדרטים. כמו כן, בסוף הדוח מופיעות מספר שאלות לתרגול עצמי.

## תוכן עניינים

1. מהם סטנדרטים?
  2. סטנדרטים במערכות חינוך
    - 2.1 חשיבות הסטנדרטים בחינוך
    - 2.2 סוגי סטנדרטים בחינוך
      - 2.2.1 סטנדרטים של תוכן
      - 2.2.2 סטנדרטים של ביצוע – קביעה של רמות ביצוע וציוני חתך
  3. קביעת סטנדרטים לביצוע – הגדרת התהליך
  4. מושגי יסוד לקביעה של ציוני חתך
  5. שיטות לקביעה של ציוני חתך
    - 5.1 שיטות מבוססות מבחן
      - 5.1.1 שיטת נדלסקי
      - 5.1.2 שיטת אבל
      - 5.1.3 שיטת אנגוף
      - 5.1.4 שיטת הסימנייה
    - 5.2 שיטות מבוססות-נבחן
      - 5.2.1 שיטת הקבוצות המנוגדות
      - 5.2.2 שיטת קבוצת הגבול
      - 5.2.3 חסרונות של שיטות מבוססות-נבחן
  - 5.3 שיטות הוליסטיות – שיטת אוסף עבודות
  - 5.4 שיטות פשרה – שיטת הופסקי
  6. בקרת איכות ותיקוף סטנדרטים
  7. סיכום
  8. שאלות תרגול
- ביבליוגרפיה
- תשובות לשאלות



## קביעת סטנדרטים במערכת החינוך

### **1. מהם סטנדרטים?**

ליה ובוועז הם תלמידים בחטיבת ביניים שלימודי האנגלית בה נעשים בהקבצות. בתום כיתה ז בבית ספרם ניתן לתלמידים מבחן מסכם המסייע לקבוע לאיזו רמת לימוד באנגלית - מתחילים או מתקדמים - ישובץ כל תלמיד. ציון החתך במבחן, 70 נקודות מתוך 100, נקבע על סמך הידע והיכולות שתלמיד צריך לגלות על מנת ללמוד בכיתה המתקדמים. ציונה של ליה במבחן היה 81, ואילו ציונו של בועז היה רק 65. על סמך ציון החתך נקבע עליה תלמד בהקבצת המתקדמים, ובוועז ילמד בהקבצת המתחילים. בדוגמה זו ציון החתך במבחן המסכם מייצג סטנדרט או רמה של ידע ושל יכולת הנדרשים מתלמיד. השוואה של ציוני התלמידים במבחן לציון החתך מאפשרת לסגל המורים לקבל החלטות באשר לכל תלמיד ותלמיד. בפרק זה נתמקד בצורך בסטנדרטים במסגרות חינוך ובשימוש בהם, ובעיקר בשיטות השונות לקביעת סטנדרטים לבצוע על ידי הגדרה של ציוני חתך במבחנים.

**סטנדרטים** - בעברית: תקנים - הם אמות מידה להערכה של מאפייני איכות של מוצרים, של תהליכים, של הישגים, של תכנים וכדומה. מכון התקנים הישראלי מגדיר תקן כך: מסמך "שמפורטות בו הדרישות החלות על מוצר או שירות כדי שיתאים ליעודו", והתהליך של קביעת התקן נקרא "תקינה". ארגונים שונים, כגון ISO-International Standards Organization, מספקים סטנדרטים בין-לאומיים בתחומים שונים. הסטנדרטים מייצגים מוסכמות המבוססות על דעת מומחים ובעלי עניין בתחום שהסטנדרט משתייך אליו; לדוגמה, מהנדסים ומומחי בטיחות עשויים לקבוע סטנדרטים בנושא בטיחות רכב. המוסכמות שהסטנדרט מייצג מאפשרות למשתמשים להעריך את איכותם של מוצרים או של תהליכים שונים באמצעות השוואתם לאמות מידה מקובלות. בדרך זו הסטנדרטים משמשים גם בסיס להשוואה בין מוצרים שונים (לדוגמה, מוצר א עומד בתקן ISO, ואילו מוצר ב אינו עומד בתקן זה) וגם אמצעי לשיפור האיכות (הגרסה הקודמת של המוצר לא עמדה בתקן ISO, אולם הגרסה העכשווית עומדת בתקן).

בעצם קביעת סטנדרט למאפיין כלשהו נקבעת הרמה שהמוצר צריך להגיע אליה על מנת להיחשב למוצר איכותי או למוצר שאיכותו מספקת. לעתים הסטנדרט מורכב מכמה רמות, ולפיכך ככל שרמת הסטנדרט גבוהה יותר, כך היא מתאימה למאפייניו של מוצר איכותי יותר. לדוגמה, כלי רכב מקבלים ניקוד על-פי פליטת ה- $CO_2$  שלהם. הניקוד מאפשר לדרג כל כלי רכב לפי אחת מכמה רמות של זיהום אוויר, והצרכנים משתמשים בדירוגים אלו בבואם לקבל החלטה בעניין רכישת רכב חדש. באופן זה הסטנדרטים הופכים חלק בלתי נפרד מתהליך של קבלת ההחלטות היום-יומי שלנו.

### **2. סטנדרטים במערכות חינוך**

#### **2.1 חשיבות הסטנדרטים בחינוך**

השימוש בסטנדרטים בחינוך הוא פועל יוצא של שיקולי מדיניות ואחריותיות (accountability) ציבורית, והם משפיעים על תהליכים ועל תוצרים חינוכיים (Hambleton & Pitoniak, 2006). בדומה לשימוש הכללי בסטנדרטים, גם כאן המטרה היא ביסוס של אמות מידה להערכה של מרכיבי מערכת החינוך בכוונה לשפר את איכות החינוך ולסייע בקבלת החלטות. בהקשר זה הסטנדרטים מתייחסים למגוון רחב של מוצרים ותהליכים, כגון: תוכנית לימודים, משאבי בית הספר, הישגי התלמידים, יעילות המורים וכדומה.

החשיבות של הסטנדרטים במערכות חינוכיות נובעת במיוחד מהצורך לקבל החלטות ברמת הפרט: החלטות על קבלה ללימודים, על שיבוץ להקבצות, על זכאות לתעודה ועוד או ברמת המערכת: השוואה בין הישגים של בתי ספר, הערכה של תוכניות לימודים, קביעת מדיניות ועוד. הסטנדרטים מסייעים לבצע החלטות אלו בהגינות, בדייקנות, במהירות, ביעילות ובשקיפות.

בדוגמה שלפניכם מובאות שתי תשובות של מורים לשאלה של הורה.

**הורה:** דני קיבל 75 במבחן המסכם. מה זה אומר?

**מורה 1:** דני מאוד משתדל, אבל הוא עדיין לא ברמת הביצוע הגבוהה. 75 זה ציון בינוני. בערך 35% מהתלמידים נמצאים ברמה הזאת, ובערך 10% מהתלמידים נמצאים מתחת לרמה הזאת.

**מורה 2:** הציון הזה אומר שדני נמצא ברמה בינונית של ביצוע. הוא יודע להשלים מילות תוכן בטקסט קצר ללא שימוש במחסן מילים. הוא יודע לענות על שאלות של אוצר מילים, כאשר אפשר להיעזר בהקשר של המשפט. עם זאת, הוא עדיין עונה רק באופן חלקי על שאלות שדורשות דיוק באזכור מידע מהטקסט ומסיק מסקנות רק ברמה מקומית ולא יוצר הכללות נכונות.

המורה הראשון מתייחס להתפלגות הציונים, ואילו המורה השני מתייחס לסטנדרטים.

זאת ועוד, הסטנדרטים הם אמצעי לשיפור מערכת החינוך והישגי התלמידים מהבחינה האקדמית, ההתנהגותית והערכית. קביעת הסטנדרטים מגדירה יעדים בני-השגה לכל המעורבים בתהליך. יעדים אלו מייצגים לא רק את מה שנחשב למתאים וחשוב ללמוד במערכת חינוך כלשהי, אלא גם את מה שמתאים וחשוב ללמוד על מנת לעבור לשלב הבא בחיים - בית ספר תיכון, שירות צבאי, השכלה גבוהה, שוק העבודה - ולפעול בו בהצלחה. הסטנדרטים משמשים גם ערוץ תקשורת ברור וסקוף בין נושאי התפקיד השונים בתהליך החינוכי: התלמיד, המורה, המנהל, ההורים והקהילה כולה. באמצעות הסטנדרטים אפשר לבנות ציפיות באשר למהות הלימודים והרכבם, לנטר את תהליך הלמידה ולהעריך הישגים ורמת ביצוע. על כן חשוב להתחיל בקביעת הסטנדרטים בשלבים המוקדמים של פיתוח תוכניות לימודים ומבחנים על מנת להגביר את ההלימה בין מרכיבים אלו (Cizek & Bunch, 2007).

## 2.2 סוגי סטנדרטים בחינוך



במערכות חינוך משתמשים במונח סטנדרט בארבעה הקשרים. נציג כל סוג סטנדרטים בקצרה ולאחר מכן נרחיב בעניין שני הסוגים הראשונים.

**סטנדרטים של תוכן (content standards)** - מגדירים מטרות המתייחסות לחומר ולכישורים הנלמדים בתוכנית הלימודים: מה המורים צריכים ללמד וכיצד, ומה התלמידים צריכים ללמוד. סטנדרטים אלו משמשים לצרכים רבים במערכות חינוך, כגון: הרכבה של תוכניות לימוד, קביעת מטרות לימודיות וציוני דרך, בניית מבחנים וסיוע בפירוש תוצאותיהם.

**סטנדרטים של ביצוע/הישג (performance or achievement standards)** - מגדירים קריטריונים לביצוע המסייעים להעריך את התקדמות התלמידים לקראת שליטה בחומר ובכישורים הנלמדים. סטנדרט ביצוע הוא הציון שתלמיד נדרש לקבל במבחן. ציון זה משקף את הידע ואת הכישורים הנדרשים על-פי הסטנדרטים של התוכן הרלוונטיים למבחן. בספרות המקצועית המונח "קביעת סטנדרטים לביצוע" מקביל למונח "**קביעת ציוני מעבר/חתך**" (**setting cutscores or passing scores**). כאמור, בהמשך הפרק נרחיב קצת בנושא סטנדרטים של תוכן, אבל עיקר הפרק יתמקד בקביעה של ציוני חתך. שימו לב שסטנדרטים של תוכן הם תמיד תיאורים מילוליים, ואילו סטנדרטים של ביצוע הם ציונים על מבחנים שונים שמיועדים לבדוק את העמידה בסטנדרטים של התוכן.

**סטנדרטים של הענקת הזדמנויות ללמידה (opportunity to learn standards / delivery standards)** - מגדירים זמינות של משאבים הניתנים לתלמידים ומאפשרים להם לעמוד בסטנדרטים של התוכן והביצוע. לדוגמה, האם קיימות תוכניות לימודים, מה מידת הזמינות של חומרי לימוד, האם יש למורים זמן לתכנן, ללמוד ולהתפתח באופן מקצועי, מה מידת הזמינות של משאבים מהקהילה ועוד. סטנדרטים של הזדמנויות ללמידה מסייעים למערכות חינוך לוודא שהתלמידים במוסדות השונים נמצאים במסגרת מתאימה המאפשרת להם ללמוד ולהתפתח (Ravitch, 1995).

**סטנדרטים של הערכה (assessment standards)** – מגדירים קריטריונים לשיפוט של איכות ההערכה במסגרת חינוכית; לדוגמה, מהימנות, תוקף והוגנות של שיטות הערכה בבית הספר. סטנדרטים של הערכה מסייעים למערכות חינוך לוודא שתלמידים נמדדים ומוערכים בשיטות ובכלים איכותיים. מידע נוסף בנושא זה אפשר למצוא במסמך הסטנדרטים למבחנים חינוכיים ופסיכולוגיים של AERA, APA & NCME (1999).

### 2.2.1 סטנדרטים של תוכן

סטנדרטים של תוכן מוגדרים לרוב על-פי נושאי הלימוד: לשון, מתמטיקה, מדעים וכן הלאה, ובכל נושא – על-פי תחומים ותת-תחומים. לרוב הסטנדרטים מוצגים באופן היררכי המסייע להבין את הקשרים ביניהם. ברמה הגבוהה ביותר מוגדר מספר מצומצם של סטנדרטי-על המציינים באופן כללי ורחב את מאפייני הידע והיכולות. מתחת לכל סטנדרט-על מפורטים

סטנדרטים נוספים, ומתחתם ניתנים ציוני דרך (benchmarks) המגדירים ביתר פירוט את מרכיבי הידע והיכולת הנדרשים. להלן דוגמה לאשכול סטנדרטים של תוכן במדעי החומר (מתוך "הסטנדרטים וציוני הדרך במדע ובטכנולוגיה לבית הספר היסודי, 2004"<sup>1</sup>):

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><u>סטנדרט תוכן 1.2</u>: התלמידים יכירו ויבינו תכונות חומרים, שינויים בחומרים, שימוש בחומרים, והשפעת ניצולם על החברה ועל הסביבה.</p> <p><u>סטנדרט משנה 1.2 א (מתוך ארבעה)</u>: התלמידים יכירו תכונות חומרים מעולם החי ומעולם הדומם, ימיינו ויאפיינו אותם על פי תכונותיהם ויקפידו על כללי בטיחות בשימוש בחומרים.</p> <p><u>ציון דרך לסטנדרט משנה 1.2 א לכיתות ה-ו (מתוך שלושה)</u>: התלמידים יזהו מתכות על פי תכונותיהם (קשיות, ברק, ניתנות לריקוע, הולכת חום, הולכת חשמל) וישוו ביניהן.</p> |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

לאחר קביעת הסטנדרטים של התוכן נהוג לשאול באיזו מידה עמדו התלמידים בסטנדרטים הנדרשים. אפשר לבחון שאלה זו ברמות שונות: בכיתה ספציפית, אצל מורה מסוים, בבית ספר אחד, במחוז או בארץ כולה - בהתאם להחלטות שיש לקבל. על מנת לענות על השאלות האלה מקבלי ההחלטות משתמשים בנתונים על הישגי התלמידים במבחנים המיועדים לבדוק את תחום התוכן הספציפי. לשם כך נקבעים סטנדרטים של ביצוע המוגדרים במונחים של ציונים המשקפים שליטה ברמה שנקבעה על-פי סטנדרט התוכן.

### 2.2.2 סטנדרטים של ביצוע – קביעה של רמות ביצוע וציוני חתך

קביעת סטנדרטים של ביצוע בהערכה חינוכית היא התהליך שמוביל להגדרה של ציון חתך - אחד או יותר - במבחן. המטרה העיקרית בקביעה של ציון חתך היא לקשר בין ביצועיו של תלמיד במבחן ובין השליטה שלו יחסית לסטנדרטים של התוכן הרלוונטיים למבחן. קביעת סטנדרטים של ביצוע נעשית בשני שלבים. ראשית יש להגדיר את רמות הביצוע האפשריות במבחן, ושנית יש לקבוע מהו ציון החתך המאפיין כל רמת ביצוע. שימו לב שקביעת אופי המבחן ותוכנו אינה חלק מהליך של קביעת הסטנדרטים של ביצוע. אנו מניחים שקביעה זו בוצעה בשלב מוקדם יותר כחלק מבניית תוכנית הלימודים וההערכה.

מרבית הפרק יעסוק בשיטות לקביעה של ציוני חתך ובעקרונות המנחים אותן, ויובאו הרחבה ודוגמאות בנושא. ואולם תחילה נפרט קצת יותר על השלב המקדים - קביעה של רמות הביצוע.

בעבר היה נהוג להשתמש רק בציון חתך אחד המבחין בין "נכשלים" לבין "עוברים", והוא נקרא "ציון עובר" או "ציון מעבר". ציון זה הוא הציון המינימלי המאפשר לנבחן להיות מסווג כבעל יכולת, לקבל תעודה, להתקבל למוסד מסוים, להתקדם במסגרת לימודית ועוד. כיום נהוג להשתמש בציון מעבר יחיד בעיקר במבחני הסמכה ורישוי, שבהם הנבחנים נדרשים לגלות רמה מסוימת של ידע ויכולת על מנת שיוכלו לעסוק במקצוע המבוקש.

<sup>1</sup> <http://tinyurl.com/3f9mrpo>

במערכות חינוך, כגון בתי ספר יסודיים ותיכונים, נהוג להשתמש בכמה ציוני חתך המסווגים נבחנים לכמה רמות ביצוע/הישג/יכולת או שליטה; לדוגמה, מתחיל, בקיא, מתקדם. לכל רמת ביצוע נקבעת **תווית** (PLL: performance level label), שהיא מילה או ביטוי קצר המאפיינים את רמת הביצוע, אך אין בהם תיאור מפורט מספיק על מנת להעריך את הביצוע בפועל. לדוגמה, במבחן המיצ"ב בעברית לכיתה ב יש שלוש רמות: נמוכה, בינונית/בסיסית, טובה ומעלה - ולכל רמה יש גם **תיאור ארוך יותר** (PLD: performance level descriptor) של כמה משפטים המפרטים את מאפייני הביצוע ברמה זו. לדוגמה, הרמה הנמוכה במיצ"ב בעברית מתוארת כך (לקוח מתוך דוח מיצ"ב תשע"א, אתר ראמ"ה):

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><u><b>רמה נמוכה:</b> הבנת הנקרא - [התלמידים יוכלו...]</u></p> <p>1. להשלים באופן מלא מילות תוכן ומילות תפקוד מתוך ההקשר במשפט ובטקסט קצר תוך שימוש במילים נתונות (מחסן מילים).</p> <p>2. להשלים באופן חלקי מילות תוכן במשפט ובטקסט קצר ללא שימוש במילים נתונות (מחסן מילים).</p> <p>3. לענות באופן מלא על שאלות איתור מידע ברמת מורכבות נמוכה בשתי הסוגות: השאלות מתייחסות לתוכן קונקרטי גלוי ומפורש, המידע הוא מקומי (ברמת המשפט או ברמה של שניים-שלושה משפטים רצופים), ונעשה שימוש באותן מילים בשאלה ובטקסט.</p> <p>4. לענות באופן חלקי על שאלות ברמת מורכבות בינונית, הדורשות איתור ואזכור של מספר פרטי מידע, והן מתייחסות לתוכן קונקרטי וגם לתוכן מופשט.</p> |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

קביעה של מספר רמות הביצוע ופיתוח התיאור המפורט שלהן הם תהליך מורכב המושפע משיקולים חינוכיים, פדגוגיים, פוליטיים וטכניים שונים. מאפייני הרמות תלויים בראש ובראשונה בתוכן המבחן ובמטרה שלשמה נוצר. את הרמות, את מספרן ואת מהותן קובע צוות של מומחים ובעלי עניין. התהליך של פיתוח רמות הביצוע נעשה בכמה סבבים, ובמהלכם הרמות עשויות להשתנות כמה פעמים לפני שאפשר לעשות בהן שימוש.

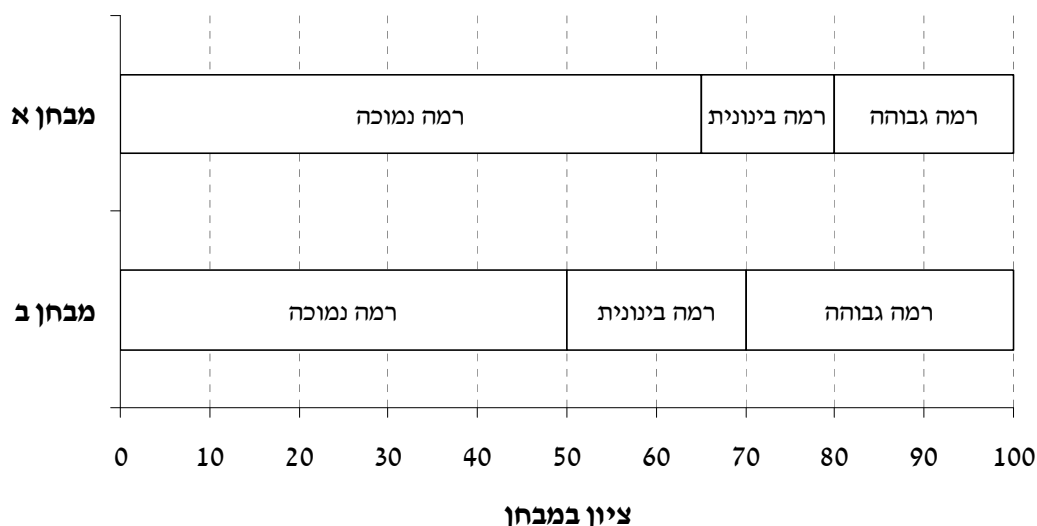
החלוקה לכמה רמות מסייעת בתהליך של פירוש הציונים ומאפשרת להתאים טוב יותר את משאבי המערכת לצורכי התלמידים. לא די לדעת שהישגיו של תלמיד מסוים במבחן טובים פחות מהישגיהם של תלמידים אחרים. חשוב גם להבין אילו מרכיבים של הידע ושל היכולת השיג התלמיד, ואילו לא השיג. רמות הביצוע מסייעות אפוא לפרש את הציון במבחן עבור התלמידים, ההורים, המורים ומקבלי החלטות אחרים.

לאחר שנקבע מספר רמות הביצוע, וסופקו להן תוויות ותיאורים, אפשר להתחיל בתהליך של קביעה של ציוני החתך. לשם כך דרוש צוות מומחים מיוחד, והוא קובע את ציוני החתך על-פי שיטה שנבחרה מבעוד מועד. תיאור ברור של רמות הביצוע מסייע לתהליך של קביעה של ציוני החתך משום שהוא מקטין אי-הבנות ואי-הסכמה בין חברים בצוות המומחים.

הקשר בין רמות הביצוע לציוני החתך מודגם בתרשים 1. עבור אותו סולם ציונים ועבור אותו רמות ביצוע אפשר להגדיר ציוני חתך שונים על מבחנים שונים. לדוגמה, במבחן א רמת הביצוע הנמוכה מוגדרת בטווח הציונים בין 0 ל-65, ואילו במבחן ב היא מוגדרת בטווח הציונים בין 0 ל-50. כיוון שבמבחן ב ציון החתך בין רמה נמוכה לבינונית הוא נמוך יותר מהציון המקביל במבחן

א, אפשר להסיק שמבחן ב הוא קשה יותר. תלמיד שמקבל ציון 60 במבחן א, ייחשב כבעל רמת ביצוע נמוכה, בעוד ציון זה במבחן ב היה מקנה לו ביצוע ברמה בינונית.

תרשים 1. הדגמה של הקשר בין רמות ביצוע לציוני חתך בשני מבחנים שונים



### 3. קביעת סטנדרטים לביצוע – הגדרת התהליך

ציזיק (Cizek, 1993) מגדיר את התהליך של קביעת הסטנדרטים לביצוע כך: תהליך מוקדם של החלטה על מספר המבחין בין שתי רמות ביצוע או יותר – באמצעות כללים ותהליכים לוגיים. ציזיק מדגיש את החשיבות של תהליך תקין – רציונלי, אחיד, ברור ושיטתי. לעומתו, קיין (Kane, 1994) מגדיר את התהליך כפועל יוצא של פירוש הציון במבחן. קיין טוען כי סטנדרט הביצוע הוא הפן המושגי של רמת יכולת רצויה; כלומר, זהו תיאור של הרמה המינימלית של ידע או של יכולת הנדרשת למטרה מסוימת. לדוגמה, הכישורים הנדרשים על מנת להיות מוסמך בראיית חשבון. לעומת זאת, ציון החתך הוא הפן התפעולי של רמת היכולת, כלומר, מיקום מסוים על סולם הציונים, כגון 80 מתוך 100 נקודות, המתאים לרמת היכולת הרצויה.

בתהליך טיפוסי של קביעת סטנדרטים לביצוע יש לקבוע תחילה את רמות הביצוע הרצויות. לאחר מכן יחל תהליך של קביעת ציוני החתך. תחילה יש להחליט על השיטה הרצויה לקביעה של ציוני החתך. כאמור, בהמשך הפרק נתאר כמה שיטות בנושא. לאחר מכן תיבחר קבוצה של מומחים שיעודה הוא קביעה של ציוני החתך. השופטים המומחים יקבלו אימון והדרכה בשיטה שנבחרה, ולאחר מכן ישפטו בשיטה זו בכמה סבבי שיפוט. אופי השיפוט – מה שופטים, כיצד ובאילו קריטריונים – הוא אחד ההבדלים העיקריים בין השיטות השונות לקביעת הסטנדרטים. על סמך שיפוטם אלו אפשר, בסופו של דבר, לחשב את ציון החתך המתאים לכל רמת יכולת. השיטות

נבדלות זו מזו גם באופן שבו שיפוטי המומחים מתורגמים לציוני חתך. בהמשך הפרק יש דיון בשיטות השונות לקביעה של ציוני חתך; דהיינו, לאחר שנקבעו הסטנדרטים של התוכן, נבחר המבחן המתאים, ונקבעו רמות הביצוע הרצויות, וכל מה שנותר הוא לקבוע את הערכים בסולם המדידה (הציונים במבחן) המתאימים לכל רמת ביצוע.

#### 4. מושגי יסוד לקביעה של ציוני חתך

עם הזמן הוצעו שיטות רבות לקביעה של ציוני חתך, ונצברו ידע וניסיון רבים בעקבות השימוש הנרחב שנעשה בהן. אף שכל שיטה מבוססת על הנחות שונות, יש בה הוראות שונות, ותוצאותיה - כלומר ציוני החתך - עשויות להיות שונות, כמה מושגי יסוד משותפים לכל השיטות או למרביתן:

א. שופטים: קביעה של ציוני החתך נעשית בידי צוות (פאנל) של מומחים ובעלי עניין - "השופטים" - אשר באים יחד לידי החלטה. בצוות השופטים עשויים להיות מורים, הורים, מנהלים, מפקחים, מומחים פדגוגיים, חוקרים בתחום התוכן, מומחי הערכה וכדומה. ככל האפשר מומלץ לבחור צוות שופטים מאוזן ומייצג מבחינת משתני רקע, כגון: מין, מגזר ועוד. לרוב השלב הראשון בתהליך של קביעה של ציוני החתך הוא אימון צוות השופטים. אימון זה נועד לחזק את ההתאמה בין השופטים ולהגביר את היכרותם עם השיטה שנבחרה לקביעה של ציוני החתך. בפועל עבודת צוות השופטים מסתכמת לרוב בהמלצות לגוף אחר אשר בידיו הסמכות המעשית לקבוע את הסטנדרט וליישם אותו, לדוגמה משרד החינוך.

ב. ניהול התהליך: מנהל התהליך הוא אדם, או קבוצה, האחראי/אחראית לבחירה של צוות המומחים ולזימונם, לניהול התהליך של קביעת הסטנדרטים ולאספקה של חומרי העבודה עבור השופטים. כמו כן בשיתוף השופטים המנהל אחראי לתיעוד התהליך של קביעת הסטנדרטים, השיקולים המנחים, הבעיות שהתעוררו וההבנות שהושגו.

ג. קבוצת התייחסות: לצורך קביעה של ציון חתך בין רמה אחת לרמה אחרת השופטים מתבקשים להתייחס לנבחן או לקבוצת נבחנים גבוליים (borderline examinees) או נבחנים שיכולתם מינימלית (minimally competent examinees). הכוונה היא להתמקד באוכלוסייה שביצועיה נמצאים על קו התפר בין ביצוע אופייני לבעלי יכולת ברמה הנדונה ובין ביצוע שרמתו נמוכה מרמה זו. על ידי מיקוד השופטים בקבוצה הגבולית אפשר להגדיר טוב יותר את מאפייניה של נקודת המעבר בין רמות ביצוע שונות. בדרך כלל חלק ניכר משלב האימון של צוות השופטים מתמקד בהבניית תפיסה משותפת של נבחנים גבוליים. ככל שתפיסתם של חברי צוות השופטים את אוכלוסיית המטרה הומוגנית יותר, כך יהיו שיפוטיהם דומים יותר זה לזה.

ד. מספר נקודות החתך: כפי שהוזכר קודם לכן, מטרת התהליך היא לקבוע ציון חתך אחד או יותר למבחן. מרבית השיטות מיועדות לקביעה של כמה נקודות חתך, ומיעוטן מיועדות לקביעה של ציון חתך יחיד, ציון מעבר. במקצת השיטות דרושות התאמות קלות על מנת לאפשר קביעה של כמה נקודות חתך.

ה. הסתברות תגובה: במקצת השיטות השופטים מתבקשים לקבוע ערך בין 0 ל-1 המייצג את ההסתברות לתגובה נכונה, RP (Response Probability). לדוגמה, השופטים עשויים להעריך שהסיכוי שנבחן גבולי יענה נכון על שאלה מס' 1 הוא 80%. באופן דומה אפשר לשאול לאילו נבחנים יש סיכוי של 50% לענות נכון על שאלה מסוימת. השימוש ב-RP מסייע לקשר את התהליך של קביעה של ציון החתך לפן ההסתברותי של תגובה לפריטים.

ו. סבבים ומשובים: שלב ההערכות נעשה לרוב בכמה סבבים, ובין סבב לסבב השופטים מקבלים משוב על עבודתם. הציפייה היא שלאחר כמה סבבים תקטן השונות בין הערכות השופטים; כלומר, שהערכים שיציעו השופטים יתכנסו סביב מספר אחד או סביב טווח מצומצם של מספרים. סוג אחד של משוב שאפשר לספק לשופטים מבוסס על מידע נורמטיבי. במצב זה ניתנת לשופטים טבלה המראה את הערכות כל השופטים בסבב הקודם: הערכות של הסתברויות לתשובה נכונה, הערכות של קושי הפריטים ועוד. באופן זה כל שופט מקבל משוב ולומד עד כמה הערכותיו דומות להערכות של שאר השופטים או שונות מהן. אנונימיות השופטים נשמרת על ידי תרגום שמם לקודים. סוג אחר של משוב מבוסס על מידע מציאותי, כלומר, על נתונים אמפיריים שהשופטים יכולים לבחון את הערכותיהם ביחס אליהם. לדוגמה, אם בשיטה כלשהי השופטים נדרשים להעריך את הקושי של כל פריט - כמשוב אפשר לספק להם ממצאים בנוגע לקושי הפריטים: אחוז העונים נכון או פרמטר הקושי מניתוח IRT. הציפייה היא שעם כל סבב ייעשו הערכות השופטים דומות יותר לקושי האמפירי של הפריטים. סוג שלישי של משוב מבוסס על מידע תוצאתי, כלומר, על אחוז התלמידים מתוך כל הנבחנים שישווו בכל רמת ביצוע על בסיס ציוני החתך שקבעו השופטים בסבב האחרון.

ז. סיכום ציונים: בסיום הסבבים נעשה סיכום של הערכות השופטים, ועל פיו נקבעים ציוני החתך הסופיים. ציון החתך הסופי עשוי להיות הממוצע או החציון של ציוני החתך שקבעו השופטים השונים. בתיאור השיטות השונות בהמשך נתייחס תמיד לציוני השופטים לאחר שסוכמו בתום הסבבים והדיונים השונים.

ח. עיגול ציונים: לעתים הערך שהשופטים הגיעו אליו אינו מספר שלם. במקרים אלו יש לעגל את הציון כדי לספק קריטריון ברור ובר-שימוש המתאים לציונים המתקבלים במבחן. הגוף האחראי לפיתוח הסטנדרטים וליישומם צריך לקבוע מדיניות של עיגול הציונים המתאימה לצרכים ולמטרות שלשמן נוצרו הסטנדרטים. שלוש האפשרויות הן אלה: לעגל תמיד כלפי מטה - גישה מקלה; לעגל תמיד כלפי מעלה - גישה מחמירה; או לעגל כלפי מעלה, כאשר השבר גדול מחצי או שווה לו, ולעגל כלפי מטה בכל שאר המקרים - גישה פשרנית.

ט. עלותן של החלטות שגויות: כל שיטה לקביעת סטנדרטים לביצוע מועדת לטעויות הנובעות מהחלטות המחלקות משתנה רציף (ציון במבחן) לקטגוריות (רמות ביצוע). מכיוון שבכל מבחן יש טעות מדידה כלשהי, סביר לצפות שקביעה של ציון חתך תביא לידי החלטות שגויות. לדוגמה, תלמיד שיכולתו מספיקה עשוי לקבל ציון נמוך מציון החתך משום שענה על המבחן בזמן שהיה חולה - **דחייה שגויה** (false negative),

ותלמיד שיכולתו אינה מספיקה עשוי לעבור את המבחן בזכות כישורי ההעתקה שלו - **קבלה שגויה** (false positive). אף שברור לנו שטעויות מסוג זה יתרחשו, איננו יכולים לדעת אם עבור נבחן ספציפי ההחלטה נכונה או שגויה. כמו כן עלות הטעות שונה בכל סוג של טעות. לדוגמה, במבחני הסמכה ברפואה קבלה שגויה חמורה הרבה יותר מדחייה שגויה: במקרה אחד נאפשר למישהו שאינו כשיר מספיק לטפל בחולים, ואילו במקרה אחר נמנע, או לכל הפחות נעכב, אדם כשיר מלעסוק ברפואה. על כן בעת קביעה של ציון החתך במבחן יש לשקול את משמעות ההחלטה מבחינת העלויות, הסיכונים וההשלכות של החלטות שגויות בשל ציון חתך נמוך מדי או גבוה מדי.

## 5. שיטות לקביעה של ציוני חתך

יש שיטות רבות לקבוע ציוני חתך, אולם אנו נתרכז רק בעיקריות ובחשובות שבהן. לאורך השנים הציעו חוקרים דרכים שונות לסווג את השיטות כדי להקל את הבחירה ואת השימוש בהן. החלוקה הבסיסית ביותר היא החלוקה לשיטות מבוססות-נורמה (שיפוט יחסי) לעומת שיטות מבוססות-קריטריון (שיפוט מוחלט). **בשיטות מבוססות-נורמה** נקודות החתך נקבעות על-פי התפלגות של ביצועי מדגם רלוונטי במבחן שנקבע. עד אמצע המאה ה-20 התבססו השיטות לקביעה של ציוני חתך בעיקר על נורמות. לדוגמה, ציון החתך למבחן היה נקבע מראש בהתייחסות לציונים של כל הנבחנים באופן שאחוז מסוים מהנבחנים, למשל 70%, יצליחו במבחן, ואילו השאר ייכשלו. במילים אחרות, על מנת לקבל ציון עובר אין צורך לדעת את כל החומר במבחן, ומספיק לדעת יותר מ-30% מהנבחנים האחרים. שיטות מבוססות-נורמה מסוג אחר עשויות לקבוע שציון החתך הוא בנקודה שהיא סטיית תקן אחת מתחת לממוצע של קבוצת הנבחנים במחזור. מי שנופל מתחת לנקודה זו "לא עבר", ומי שנמצא מעליה "עבר". לשיטות מבוססות-נורמה יש שני חסרונות שבעטיים הן כבר כמעט לא-מקובלות. ראשון, קביעה של נקודות חתך בדרך זו כלל לא מתייחסת ליכולות או לידע של התלמיד בכל נקודת חתך. הפרשנות של הציון היא רק במונחים יחסיים לשאר ביצועי התלמידים ואיננה אומרת דבר על מה התלמיד יודע או איננו יודע. חיסרון שני מתייחס להשוואה שבין נבחנים ממחזורי בחינה שונים. מכיוון שמשנה לשנה רמת הידע של אוכלוסיית הנבחנים עשויה להשתנות ובהתאם לכך גם קושי הפריטים במבחן והרכבו, יש קושי בפרשנות הציון.

כיום השימוש ב**שיטות מבוססות-קריטריון** הוא הרזוח. העיקרון המנחה את השיטות המבוססות על קריטריון הוא זה: כדי לקבל ציון חתך מסוים על הנבחנים לעמוד בקריטריון מסוים של ידע ויכולת. לדוגמה, סטנדרט התוכן עשוי להגדיר שני תחומי ידע שהנבחנים צריכים לשלוט בהם ברמה מספקת, או הקריטריון יכול להיות פתרון נכון של 80% מהשאלות בכל אחד מהתחומים. שלא כבשיטות מבוססות-נורמה, השיפוט אינו מתייחס לכלל הנבחנים, ועל כן בשיטה מבוססת-קריטריון ייתכן שכל הנבחנים יצליחו במבחן, וייתכן גם שכל הנבחנים ייכשלו בו. בשיטה מבוססת-נורמה אחוז העוברים נשאר תמיד קבוע, אף אם ציון החתך נמוך מאוד, ועל כן עשוי שלא לאפיין את הידע ואת היכולת הרצויים. אף ששתי הגישות לקביעת סטנדרטים נראות שונות בתכלית, יש הטוענים כי כל שיפוט משלב היבטים מוחלטים ויחסיים. על בסיס טיעון זה הוצעו

שיטות שהן בין שיפוט מוחלט לשיפוט יחסי ונקראות שיטות פשרה. מרבית השיטות שנציג בפרק זה הן שיטות מבוססות-קריטריון, ורק האחרונה שבהן היא שיטת פשרה.

חלוקה אחרת של השיטות לקביעה של ציוני חתך שהוצעה בספרות מתמקדת באובייקט ההערכה: שיפוט מבוסס-מבחן, ובו השופטים מעריכים את היכולות הנבדקות במבחן, או **שיפוט מבוסס-נבחן**, ובו הם מעריכים את יכולות הנבחנים. ואולם במרבית השיטות לא ניתן להפריד באופן מוחלט בין שני התהליכים מכיוון ששיפוט מבוסס-מבחן מתייחס גם לאוכלוסיית הנבחנים, ועל כן הם גם מבוסס-נבחן. סיווג מעין זה נוח בהתייחסות לתמונה הכללית, אך עשוי להסתבך, כאשר מתחשבים בפרטים הספציפיים של כל שיטה ושיטה. בהמשך הפרק מתוארות שמונה שיטות עיקריות לקביעה של ציוני חתך. מידע נוסף על השיטות השונות אפשר למצוא במקורות האלה: Cizek (2001), Cized & Bunch (2007), Hambleton & Pitoniak (2006), Zieky, Perie & Livingston (2008).

## 5.1 שיטות מבוססות-מבחן

### 5.1.1 שיטת נדלסקי

שיטת נדלסקי (Nedelsky Method, Nedelsky, 1954) מתאימה למבחנים המורכבים מפריטי בָּרָה בלבד. המשימה המוטלת על השופטים היא להעריך את ההסתברות שנבחנים יצליחו לזהות את המסיחים בכל אחד מהפריטים ולהימנע מבחירתם. נקודת ההתייחסות של השופטים היא תלמיד שיכולתו מינימלית או מצויה על קו הגבול בין נכשל לעובר (Minimally Competent, Just Barely Passing). נדלסקי טבע את המונח **F-D Student** לתיאור תלמיד היפותטי כזה: F מציין ציון נכשל, ו-D הוא הציון הנמוך ביותר מעל ציון נכשל. תלמיד נכשל לא יוכל לפסול שום מסיח, ואילו תלמיד F-D יוכל לפסול את מקצת המסיחים. על השופטים להתעמק בכל פריט ופריט במבחן ולקבוע אילו מסיחים יוכל תלמיד F-D לפסול. לאחר מכן מחשבים את "דירוג נדלסקי", שהוא ההופכי של מספר אפשרויות התשובה שנותרו לאחר פסילת המסיחים. למעשה, דירוג זה הוא ההסתברות שתלמיד שיכולתו מינימלית יבחר בתשובה הנכונה בפריט. לדוגמה, אם השופטים מחליטים שבפריט בעל ארבע תשובות אפשריות יש מסיח אחד בלבד שתלמיד F-D יכול לפסול - נותרו שלוש תשובות אפשריות שרק אחת מהן היא נכונה, ולכן ההסתברות שהתלמיד המתואר יענה נכון הוא  $1/3=0.33$ , וזהו דירוג נדלסקי לפריט זה.

על מנת לקבוע ציון חתך למבחן יש לסכם את דירוגי נדלסקי של כל הפריטים. לדוגמה, אם במבחן של 50 פריטים, שבכל אחד מהם יש ארבע תשובות אפשריות, היו 30 פריטים בדירוג של 0.33 (כלומר, שהשופטים העריכו שנבחן F-D יפסול תשובה אחת) ועוד 20 פריטים בדירוג של 0.25 (כלומר, שהשופטים העריכו שנבחן F-D לא יפסול אף תשובה אפשרית), ציון החתך במבחן הוא  $15 \approx 9.99 + 5 = 30 \times 0.33 + 20 \times 0.25$ . באופן זה אפשר לחשב ציון חתך במבחן מבלי להשתמש בנתוני ביצוע אמתיים. במילים אחרות, שיטת נדלסקי מאפשרת לקבוע סטנדרטים לפני שנעשה שימוש כלשהו במבחן. במובן זה זוהי שיטה מבוססת-קריטריון ומבוססת-מבחן.



אף ששיטת נדלסקי היא שיטה אינטואיטיבית ופשוטה לחישוב, יש לא מעט הסתייגויות מהשימוש בה. ראשית, היא מוגבלת לפריטי בִּרְרָה. שנית, לרוב נקבעים בה ציוני חתך נמוכים מדי, שכן נדיר שייקבעו באמצעותה דירוגי נדלסקי שהם בין 0.5 ל-1 (Berk, 1984). דבר זה נובע מנטייתם של רוב השופטים להניח שתלמיד F-D לא יוכל לפסול את כל התשובות המוטעות, ולכן יישאר לכל הפחות מסיח אחד מלבד התשובה הנכונה; כלומר, במרבית המקרים יניחו השופטים שעל תלמיד F-D לבחור לפחות בין שתי אפשרויות תשובה, ועל כן דירוג נדלסקי לא יהיה גבוה מ-0.5. שלישית, ההנחה שנבחני F-D יבחרו אקראית מתוך התשובות שלא נפסלו - לא קיבלה תמיכה מחקרית (Melican, Mills & Plake, 1987). שיטת נדלסקי קיימת כבר שנים רבות, והיא השיטה המבוססת-קריטריון הראשונה שהוצעה, אך כיום משתמשים בה בעיקר במבחני הסמכה מקצועית ופחות בתחום החינוך הבית ספרי.

### 5.1.2 שיטת אבל

גם שיטת אבל (Ebel Method, Ebel, 1972) היא מבוססת-קריטריון ומבוססת-מבחן. השופטים בשיטת אבל מתבקשים להעריך שני היבטים של פריטי המבחן: (א) הקושי של כל פריט: קל, בינוני או קשה; (ב) הרלוונטיות של כל פריט לעולם התוכן הנמדד: מהותי, חשוב, מתקבל על הדעת, מוטל בספק. באופן זה השופטים מחלקים את כל הפריטים במבחן ל- $3 \times 4 = 12$  קטגוריות נפרדות; לדוגמה, פריטים קלים ומהותיים, פריטים בינוניים וחשובים וכולי. שלושת הטורים הראשונים בטבלה 1 מדגימים חלוקה של הפריטים לקטגוריות הללו במבחן של 100 פריטים. לאחר חלוקה זו השופטים מתבקשים להעריך - עבור כל קומבינציה של קושי ורלוונטיות - את פרופורציית הפריטים שתלמידים גבוליים יצליחו לענות עליהם נכון (הטור הרביעי בטבלה). למשל, מבין עשרת הפריטים שנשפטו כקשים וכמהותיים, הערכת השופטים היא שתלמיד גבולי יענה על 80% מהם או 0.8 מהם. לצורך קביעה של ציון החתך מכפילים בכל שורה את מספר הפריטים (טור 3) בפרופורציה (טור 4) ומקבלים את טור 5. סכמה של טור זה קובעת את ציון החתך, ובמקרה שמתואר בטבלה 1 ציון זה הוא 63.

טבלה 1. דוגמה לקביעה של ציוני חתך למבחן בשיטת אבל

| קושי   | רלוונטיות     | מספר פריטים | פרופורציית העונים נכון | מכפלה |
|--------|---------------|-------------|------------------------|-------|
| קל     | מהותי         | 5           | 1.0                    | 5     |
|        | חשוב          | 10          | 0.9                    | 9     |
|        | מתקבל על הדעת | 5           | 0.8                    | 4     |
|        | מוטל בספק     | 0           | --                     | --    |
| בינוני | מהותי         | 10          | 0.8                    | 8     |
|        | חשוב          | 25          | 0.7                    | 17.5  |
|        | מתקבל על הדעת | 10          | 0.6                    | 6     |
|        | מוטל בספק     | 0           | --                     | --    |
| קשה    | מהותי         | 10          | 0.5                    | 5     |
|        | חשוב          | 15          | 0.4                    | 6     |
|        | מתקבל על הדעת | 5           | 0.3                    | 1.5   |
|        | מוטל בספק     | 5           | 0.2                    | 1     |

אף ששיטת אבל מיושמת בעיקר במבחן של שאלות בִּכְרָה, המקבלות ציון דיכוטומי 0 או 1, על ידי שינוי פשוט אפשר להשתמש בה גם בשאלות המקבלות ציון פוליטומי, לדוגמה: 0, 1, 2 או 3. במקרה זה השופטים מעריכים את החלק היחסי שקבוצת תלמידים גבוליים יקבלו מתוך סך הניקוד האפשרי על שאלות אלו. חלופה אחרת בשיטת אבל היא שימוש בנתונים אמיתיים על מנת לקבל אומדני קושי לפריטים ובדרך זו לעקוף את הצורך בשיפוט סובייקטיבי של קושי. שיטת אבל נחשבת לשיטה פשוטה ואינטואיטיבית ליישום וזוכה לשימוש נרחב עד היום, בעיקר בבחינות הסמכה ברפואה. עם זאת, מבקרי השיטה טוענים שההפרדה בין שני ממדי ההערכה - קושי ורלוונטיות - בעייתית משום שבמצבים מסוימים יש התאמה חזקה בין ממדים אלו (Berk, 1984). ביקורת אחרת מתייחסת להיבט התוקף (Cizek, 1996): אם הרלוונטיות של מקצת השאלות מוגדרת כמוטלת בספק, מדוע שאלות אלו נכללות במבחן, והאם מוצדק לקבוע את ציון החתך בהסתמך על הצלחת הנבחנים בשאלות אלו?

### 5.1.3 שיטת אנגוף

שיטת אנגוף (Angoff Method, Angoff, 1977) על מגוון הרחבותיה יושמה ונחקרה באופן המקיף ביותר מכל השיטות לקביעה של ציוני חתך, במיוחד בתחום של מבחני רישוי והסמכה. גם שיטה זו, כמו קודמותיה, מבוססת-קריטריון ומבוססת-מבחן.

הפשטות והפרקטיות של שיטת אנגוף מקנות לה יתרון על פני השיטות האחרות, ויעילותה זכתה לתמיכה ממחקרים רבים המשווים בין השיטות. עם זאת, לעתים מיטשטשים ההבדלים בין השיטה המקורית ובין המודיפיקציות השונות שהוצעו לה. למעשה, במרבית המקרים שבהם חוקרים מצהירים שהשתמשו בשיטת אנגוף<sup>2</sup> ניתן להניח שהשתמשו באחת המודיפיקציות של השיטה (Cizek & Bunch, 2007). לשם הסרת העמימות חשוב להבין את מהותה של כל מודיפיקציה בשיטה (אבל לא בהכרח את השם שהוצע לה).

השיטה הראשונה שהציע אנגוף היא שיטה פשוטה ביותר: עבור כל פריט השופטים מתבקשים לציין אם תלמיד שיכולתו מינימלית יענה על הפריט נכון או לא נכון. אם נעניק את הערך 1 לכל פריט שהתלמיד יענה עליו נכון, ו-0 לכל שאר הפריטים - סכום הערכים האלה יהיה ציון החתך במבחן. כיום נהוג לקרוא לשיטה זו **שיטת ה"כן/לא"** (Yes/No Method, Impara & Plake, 1997) או **שיטת "הגבול הישיר"** (Direct Borderline, Downing et al., 2003).

כהערת שוליים לשיטה הראשונה שהציע הוסיף אנגוף חלופה שהייתה, למעשה, לשיטה המכונה כיום **"שיטת אנגוף המקורית"** (Unmodified Angoff's Method): במקום להעריך אם התלמיד שיכולתו מינימלית יענה נכון או לא נכון על הפריט, על השופטים להעריך את ההסתברות שהתלמיד יענה נכון על הפריט. למעשה, השופטים חושבים על קבוצת תלמידים הנמצאים על קו

<sup>2</sup> על פי אנגוף, מקור השיטה הוא דווקא פסיכומטריקאי ידוע אחר: לדיארד טאקר.

הגבול בין "נכשל" ל"עובר" ומעריכים את האחוז מהם שיענו נכון על הפריט. ערכים אלו מכונים "דירוגי אנגוף". סכום דירוגי אנגוף של כל הפריטים יהיה ציון החתך במבחן.

כיום נהוג להשתמש במודיפיקציות שונות של שיטת אנגוף (Modified Angoff's Methods), ובהן - בדומה לשיטות אחרות - כמה סבבים של שיפוט. בין כל צמד סבבים השופטים מקבלים משוב על שיפוטיהם, דנים בשיפוטיהם ומסתייעים בנתונים על ביצועי אמת של נבחנים. פעולות אלו מסייעות להקטין את שונות ההערכות בין השופטים ועקב כך מאפשרות להתכנס לדירוגים מדויקים, יציבים וריאליים יותר. על מנת לקבוע את ציון החתך במבחן, בדומה לשיטות אחרות, נהוג למצע את דירוגי השופטים של כל הפריטים בסבב האחרון.

כאמור, לאורך השנים הוצעו שינויים רבים לשיטה המקורית. אחת הווריאציות נקראת **שיטת אנגוף המורחבת** (Extended Angoff's Method; Hamblen & Plake, 1995), והיא מיועדת לשימוש בפריטים פתוחים, פריטים מסוג הפקת תגובה. בדומה לשיטה המקורית, בשיטה זו עבור כל פריט על השופטים להעריך את הניקוד מתוך סולם הציונים שצפויים לקבל נבחנים שיכולתם מינימלית. לדוגמה, אם בפריט מסוים אפשר לקבל ניקוד של 0, 1, 2 או 3, שופט אחד עשוי להעריך שהציון הצפוי לנבחנים שיכולתם מינימלית הוא 1, ואילו שופט אחר עשוי להעריך שהציון הצפוי הוא 2. קביעת ציון החתך תיעשה על ידי מיצוע דירוגי השופטים של כל הפריטים בסבב השיפוט האחרון והכפלה במספר הפריטים. פעולה זו מניחה שהניקוד המרבי זהה בכל הפריטים. אם הניקוד המרבי שונה מפריט לפריט, יש למצע את דירוגי השופטים עבור כל פריט ולסכם את ממוצע הדירוגים של כל הפריטים.

על אף השימוש הנרחב בשיטת אנגוף הושמעו ביקורות נגדה בעקבות דוח של האקדמיה הלאומית לחינוך בארצות ברית (National Academy of Education). הדוח בחן את הקביעה של ציוני החתך למבחן ה-National Assessment of Education Progress (NAEP) וביקר את השימוש בשיטת אנגוף ובשיטות דומות בטענה שמושגים, כגון "תלמיד שיכולתו מינימלית", הם מושגים לא-נתפסים, לא-שמישים ומקשים על השופטים ליצור תפיסה כללית של יכולת (Shepard et al., 1993). עם זאת, ביקורת זו לא קיבלה אישוש מחקרי חזק, ושיטת אנגוף נותרה השיטה הפופולרית ביותר לקביעה של ציוני החתך.

#### 5.1.4 שיטת הסימנייה

השיפוט בשיטת הסימנייה (Bookmark Method; Lewis, Mitzek & Green, 1996) נעשים באמצעות הנחת סימנייה במקום המתאים בתוך סט של פריטים המסודרים על-פי אומדני קושי, שהתקבלו בהעברת המבחן בקבוצה רלוונטית. השיטה זכתה לשימוש נרחב משום שאפשר ליישמה בהערכות המורכבות מסוגי פריטים שונים, וכן משום שהמשימה המוטלת על השופטים פשוטה ואינטואיטיבית. לשם השוואה, נניח שהשופטים משתמשים במבחן של 50 פריטים. בשיטת אנגוף עליהם לקבוע עבור כל פריט את ההסתברות לתשובה נכונה של נבחן שיכולתו מינימלית, כלומר, לעשות 50 שיפוט. בשיטת הסימנייה עליהם לעשות רק שיפוט אחד, הבא

לידי ביטוי בהנחת הסימנייה במקום בחוברת המציין את נקודת החתך. נוסף על כך, שיטת הסימנייה מתבססת על תורת התגובה לפריט (IRT), שהיא המסגרת התאורטית לפיתוחם של מרבית המבחנים הסטנדרטיים רחבי ההיקף ועתיירי הסיכון (Cizek & Bunch, 2007). מכיוון שב-IRT האומדנים של מאפייני היכולת של הנבחנים ואומדני הקושי של הפריטים נמצאים על אותו סולם, השימוש במתודולוגיה זו מקל את הפירוש של הסטנדרטים הנקבעים ואת קישורם לפרוצדורות אחרות מבוססות-IRT, כגון סילום וכיוול.

שיטת הסימנייה מבוססת על שימוש בחוברת של פריטים המסודרים בסדר קושי עולה, **חוברת הפריטים הסדורים** (OIB-The Ordered Item Booklet). החוברת יכולה להכיל פריטים מסוג אחד או מכמה סוגים כמו פריטי בִּרְרָה ופריטים של הפקת תגובה. עם זאת, אין הכרח שהפריטים יבואו כולם מנוסח אחד של המבחן. העיקרון המנחה בבחירת הפריטים לחוברת הוא הבטחת כיסוי נאות של התוכן ושל רמות הקושי של הפריטים במבחן. כל פריט בררר יופיע בחוברת רק פעם אחת, במקום המתאים לקושי של הפריט על-פי IRT, כלומר, על-פי ערכו של הפרמטר  $b$ . כל פריט מסוג הפקת תגובה יופיע בחוברת כמה פעמים, על-פי מספר הנקודות המרבי שאפשר לקבל בפריט ובהתאם לקושי של השלב המתאים לציון. לדוגמה, אם בפריט מסוים הציון המרבי הוא חמש נקודות, יופיע הפריט חמש פעמים בחוברת - אף שיש לו שישה ציונים אפשריים: 0, 1, 2, 3, 4 ו-5 - משום שאין צורך לקבוע נקודת חתך על בסיס הציון 0. בפריטים מסוג הפקת תגובה יצורפו גם המחווין ודוגמאות לתשובות ברמות ציון שונות.

בקביעת הסטנדרטים בשיטת הסימנייה על השופטים לבחון כל פריט ופריט ולהחליט אם סביר שנבחן שיכולתו מינימלית או גבולית יענה נכון על השאלה או יקבל ציון מסוים על השאלה. ערך הסבירות תלוי במודל ה-IRT שעל פיו דורגו הפריטים, ונוטים לקבוע אותו בין 0.5 ל-0.8. לדוגמה, יש המעדיפים להשתמש בערך של  $RP=0.5$ , אם המבחן נותח על-פי מודל Rasch או 1PL (Wang, 2003), ובערך  $RP=0.67$  (או  $2/3$ ), אם המבחן נותח על-פי מודל 2PL.<sup>3</sup> במילים אחרות, אם השופט מעריך שלנבחן יש סיכוי של לפחות 0.67 לענות נכון על שאלה  $X$ , הוא יכול לעבור הלאה ולבחון את השאלה הבאה. באופן מעשי השופט ממשיך לעבור מפריט לפריט לפי סדרם, כל עוד הוא מעריך שההסתברות שהנבחן יענה על הפריט היא יותר מ-0.67. כאשר השופט נתקל בפריט שלהערכתו סיכויי הנבחן לענות עליו נכון קטנים מ-0.67, עליו לסמנו בסימנייה. מקום זה מסמן שעד כה היו הפריטים ברמת יכולתו של הנבחן הגבולי, ואילו מכאן והלאה הפריטים הם מעל יכולתו של הנבחן. באופן דומה אפשר לקבוע יותר מנקודת חתך אחת, כמו במצב של כמה רמות יכולת. לדוגמה, לאחר קביעת החתך לרמה הבסיסית ימשיכו השופטים לעבור הלאה מפריט לפריט, ונקודת ההשוואה שלהם תתבסס על נבחנים ברמת היכולת השנייה, וכן הלאה.

מנהל התהליך יכול לקבוע שעל השופטים להניח את הסימנייה בפריט האחרון שסיכויי הנבחן להצליח בו היו 0.67 או יותר. ההחלטה אם הסימנייה תונח על הפריט האחרון שסביר שהנבחן

<sup>3</sup> ההבדל העיקרי בין מודל 1PL למודל 2PL הוא במספר הפרמטרים המאפיינים כל פריט. במודל הראשון יש רק פרמטר של קושי, ובמודל השני יש גם פרמטר של הבחנה.

יוכל לפתור נכון או על הפריט הראשון שנמצא ברמת קושי גבוהה יותר - היא החלטה שרירותית. הדבר החשוב הוא שכל השופטים יהיו מודעים להוראות וימלאו אותן בהתאם.

ציון החתך במבחן מחושב ביחידות לוג שהן היחידות של היכולת ( $\theta$ ) במודלים של IRT. פרטי הנוסחה לחישוב ציון החתך הם מעבר לנדרש בפרק זה, אבל באופן כללי נאמר שציון החתך הוא הסכום של שני מרכיבים: (א) קושי הפריט שעליו הונחה הסימנייה, ו-(ב) קבוע המחושב כפונקציה של ה-RP הנבחר.

אף ששיטת הסימנייה קיימת למעלה מעשור וזוכה לשימוש נרחב, המחקר עליה עדיין מועט יחסית (Karantonis & Sireci, 2006). הטענות העיקריות שהועלו נגד השיטה הן שסדר הפריטים בחוברת מושפע יתר על המידה ממודל המדידה הנבחר, לדוגמה 1PL או 2PL.

## **5.2 שיטות מבוססות-נבחן**

שתי השיטות המוצגות בחלק זה דורשות משופטים להתמקד בקבוצת תלמידים אמיתית ובהישגיהם במבחן.

### **5.2.1 שיטת הקבוצות המנוגדות**

שיטת הקבוצות המנוגדות (Contrasting Groups Method; Berk, 1976) מבוססת על תהליך המסייע לקבוע ציוני חתך המבחינים בין קבוצות שונות בתכלית: א. קבוצת השולטים בחומר (True masters) - מי שלמדו לימודים אפקטיביים בנושא המבחן.

ב. קבוצת הלא-שולטים בחומר (True non-masters) - מי שלא למדו לימודים אפקטיביים בנושא המבחן.

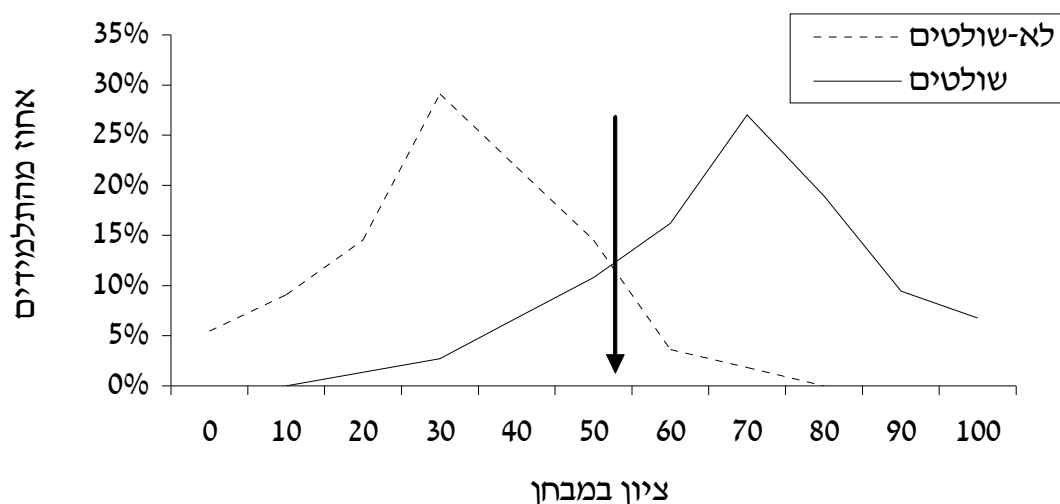
בשיטת הקבוצות המנוגדות יש להשתמש בנתונים אמיתיים של תלמידים שנבחנו במבחן לפני שנקבעו ציוני החתך. הצוות הקובע את ציוני החתך מורכב משופטים בעלי היכרות אישית עם התלמידים שנבחנו ועם תחומי הידע המוערכים במבחן. השופטים מתבקשים - מבלי לדעת מהם הישגי התלמידים במבחן - לקבוע עבור כל תלמיד, אם הוא משתייך לקבוצת השולטים או לקבוצת הלא-שולטים בחומר. עבור כל אחת משתי הקבוצות הללו בודקים את התפלגות ציוניה במבחן. תרשים 2 מראה דוגמה היפותטית להתפלגות הציונים בשתי קבוצות. ציון החתך במבחן נקבע לפי הנקודה על ציר הציונים בתרשים שמבחינה באופן המיטבי בין שתי הקבוצות.

במקרה זה ציון החתך נקבע לנקודת המפגש בין שתי ההתפלגויות, ציון של 52 מתוך 100 נקודות. נבחנים שקיבלו ציונים מתחת ל-52 - סביר יותר להניח שבאו מקבוצת הלא-שולטים, והנבחנים שקיבלו ציונים מעל נקודה זו סביר יותר להניח שהם מקבוצת השולטים.

במצבים אמיתיים נדיר לקבל התפלגויות חלקות ורציפות. לרוב יראו ההתפלגויות "רעש" במידה לא-מבוטלת (הקווים בתרשים יהיו משוננים, לא חלקים) בשל השימוש במדגמים קטנים או בגלל אי-הופעת ציונים כלשהם בנתונים. דבר זה מקשה על איתור נקודת החיתוך האופטימלית.

במצבים שבהם ההתפלגויות אינן רציפות, אפשר להשתמש בשיטה להחלקת ההתפלגות, לדוגמה ממוצע נע.<sup>4</sup> כאשר אי-אפשר לאתר את נקודת החיתוך באופן חזותי, אפשר לקבוע אותה באופן סטטיסטי. לדוגמה, נקודת החיתוך יכולה להיות נקודת האמצע בין ממוצעי (או חציוני) שתי הקבוצות. השימוש במודל סטטיסטי עדיף על ניתוח חזותי, במיוחד כאשר בנתונים יש הרבה "רעש". ישנן שיטות סטטיסטיות נוספות למצוא את נקודת החתך במצבים אלה שלא כאן המקום לפרטן. המעוניינים יוכלו לקבל תמצית קצרה של שיטות אלה בהרחבה שלהלן.

תרשים 2. התפלגויות היפותטיות להדגמה של שיטת הקבוצות המנוגדות



#### ----- הרחבה: שיטות סטטיסטיות לקביעת נקודת החתך בין שתי התפלגויות ----

הגרסיה היא שיטה סטטיסטית לבחינת הקשר בין שני משתנים או יותר. כאשר ישנם רק שני משתנים, אחד מהם מוגדר כמשתנה התלוי או המנובא, והמשתנה השני הוא הלא-תלוי או החזאי. התוצר של ניתוח הרגרסיה הוא נוסחה של קו ישר מסוג  $Y=aX+b$ , כאשר  $Y$  הוא המשתנה התלוי ו- $X$  הוא המשתנה הלא-תלוי.  $b$  הוא נקודת חיתוך הקו עם ציר  $Y$ , ו- $a$  הוא השיפוע של הקו. הקו שנקבל מניתוח הרגרסיה הוא זה שמרחק כל הנתונים ממנו הוא הקטן ביותר; כלומר, אם נשתמש בנוסחת הקו הזה כדי לנבא את ערכי  $Y$  מערכי  $X$ , אזי הטעות הממוצעת שלנו תהיה הקטנה ביותר.

כאשר המשתנה התלוי הוא דיכוטומי (בעל שני ערכים, כגון עובר/לא עובר), יש להשתמש בגרסיה לוגיסטית. בניתוח זה ציוני המבחן משמשים בניבוי ההסתברות להשתייכות לקבוצת השולטים לעומת קבוצת הלא-שולטים. תוצאת הרגרסיה היא ערכי המקדמים  $a$  ו- $b$  הניתנים בנוסחה לניבוי ערכי  $Y$  מערכי  $X$ :  $Y=aX+b$ , כאשר  $X$  הוא הציון במבחן, ו- $Y$  הוא לוג הסיכוי,

<sup>4</sup> בממוצע נע כל ערך בהתפלגות מוחלף בממוצע של  $K$  הערכים הבאים בהתפלגות לרבות הערך עצמו. לדוגמה, אם  $K=5$ , אזי הנקודה הראשונה היא ממוצע חמשת הערכים הראשונים (הערך הראשון עד החמישי), הנקודה השנייה היא ממוצע של הערכים 2 עד 6 וכן הלאה.

היחס בין ההסתברות להשתייכות לקבוצת השולטים ובין ההסתברות להשתייכות לקבוצת הלא-שולטים. מכיוון שהנקודה שאנו מעוניינים לאתר היא נקודת האמצע (קו התפר) בין שתי הקבוצות, אנו מעוניינים בהסתברות של 50% להשתייכות לכל אחת מהקבוצות. אם כן, לוג הסיכוי הוא:

$$Y = \log[p/(1-p)] = \log[0.5/(1-0.5)] = \log(1) = 0$$

במילים אחרות, נקודת החתך תימצא על ידי השוואה של משוואת הרגרסיה לאפס ופתרון עבור ציון המבחן,  $X$ .

אף ששיטת הקבוצות המנוגדות נראית קלה ליישום, שופטים עשויים להתקשות בסיווג תלמידים לשולטים או לא-שולטים. השיטה המפורטת להלן מנסה להתגבר על בעיה זו.

### 5.2.2 שיטת קבוצת הגבול

שיטת קבוצת הגבול (Borderline Group Method; Zieky & Livingston, 1977) מתמקדת בקבוצת התלמידים השולטים בחומר, אבל לא ברמה של "שולטים אמיתיים". בראשית התהליך מתקיים דיון בין השופטים, ובו נבנה במשותף הפרופיל המאפיין תלמידים בקבוצת הגבוליים. לאחר מכן השופטים מתבקשים לאתר את התלמידים הגבוליים בקבוצת הנבחנים - שוב, מבלי להתייחס לציוניהם במבחן. לחלופין, השופטים מתבקשים לסווג את כל התלמידים לשלוש קבוצות: שולטים, גבוליים ולא-שולטים. התוצר של תהליך זה הוא התפלגות הציונים בקבוצת הגבול. ציון החתך ייקבע על-פי אחד מאומדני הנטייה המרכזית, לרוב החציון.

שיטת קבוצת הגבול דורשת שיפוט מדוקדק יותר משיטת הקבוצות המנוגדות משום שהיא דורשת לסווג את התלמידים הנמצאים באזור עמום של ההתפלגות, להבדיל מן התלמידים הנמצאים בקצוות ההתפלגות. באופן דומה אפשר להרחיב את השיטות האלה על ידי חלוקה לרמות בתוך כל קבוצה. לדוגמה, בשיטת השיפוט האנליטי (Analytical Judgment Method; Plake & Hambleton, 2001) ראשית מחלקים את התלמידים לשלוש קבוצות יכולות: בסיסית, מיומנת ומתקדמת, ואז מחלקים כל קבוצה לשלוש רמות: נמוכה, בינונית וגבוהה. באופן טבעי תהיה חפיפה בין קבוצות סמוכות, כמו בסיסית-גבוהה ומיומנת-נמוכה. לאחר מכן קובעים את ציון החתך על-פי שיטת הקבוצות המנוגדות, כלומר - באופן חזותי - נקודת האמצע בין ממוצעי הקבוצות הסמוכות, או ברגרסיה לוגיסטית (ראו הרחבה), או שמאחדים את שתי הקבוצות הסמוכות ומשתמשים בשיטת קבוצת הגבול לקביעה של ציון החתך, לדוגמה החציון של הקבוצה המאוחדת.

### 5.2.3 חסרונות של שיטות מבוססות-נבחן

השיטות האלה לוקות בחיסרון הנובע מהתבססותן על מדגם אמיתי. במקרים של מדגם קטן, ובמיוחד אם בקבוצת הגבול יש מעט תלמידים, האומדנים שבאמצעותם נקבעים ציוני החתך אינם יציבים מספיק. גם שיטות ההחלקה שמשתמשים בהן יכולות ליצור הטיות באומדני נקודות החתך. בעיה אחרת היא ההסתמכות הבלעדית של השיטות על יכולתם של השופטים לסווג

תלמידים על-פי רמת השליטה שלהם בחומר. ידוע כי לשיפוטם כאלו יש, לרוב, מהימנות נמוכה משום שאף לשופטים מיומנים קשה להעריך היבט אחד בלבד של התלמיד ולהתעלם ממכלול היכולות האחרות שלו. בעיה זו חמורה במיוחד בשיטה זו, שכן השופטים חייבים להכיר את התלמידים באופן אישי, ועל כן שיפוטיהם מושפעים בהכרח מגורמים אחרים מלבד היכולת והידע של כל תלמיד.

### 5.3 שיטות הוליסטיות – שיטת אוסף עבודות

העיקרון המנחה שיטות הוליסטיות לקביעת סטנדרטים (Body of Work Method; Kingston, 2001) הוא הערכה של מכלול הביצועים של התלמיד באמצעות ציון אחד המשקף את אוסף עבודותיו. השיטה יועדה בעיקרה להערכות מורכבות יותר ממבחני בִּכְרָה, כגון: הערכה של מטלת כתיבה, הערכת ביצוע, הערכה של תיק עבודות (portfolio) ועוד. בשיטה זו השופטים מתבקשים לסווג את עבודות (או את הערכות) התלמידים לרמות ביצוע בכמה סבבים, ובכל סבב המיקוד על טווחי הציונים מתכנס מסביב לקו התפר בין רמה אחת לאחרת. בזמן הקצר שעבר מאז פרסום השיטה היא זכתה לשימוש נרחב במבחנים עתירי סיכון מסוגים שונים.

שיטת אוסף העבודות מורכבת מארבעה שלבים. **בשלב האימון** (training) המטרה העיקרית היא לבחור את הרכב העבודות שיישפטו, להסכים על מאפיינים של רמות הביצוע שעל פיהן יש לשפוט את העבודות ולהגביר את ההיכרות עם הפרטים של שיטת העבודה. ראשית, הצוות המנהל בוחר מעט עבודות שמוסכם עליו כי הן מדגימות בבירור את רמות הביצוע השונות. הקונצנזוס מכוון להקל על השופטים את ההיכרות עם רמות הביצוע. לאחר מכן השופטים מתבקשים לסווג כל דוגמה לאחת מכמה רמות ביצוע. בשלב הבא הם מתבקשים להשוות בין הערכותיהם ולדון בדמיון ובשוני ביניהן במטרה להעמיק את היכרותם עם רמות הביצוע ולחזק את ההסכמה ביניהם.

בשלב השני, **שלב מציאת הטווח** (range finding), השופטים בוחנים קבוצות של עבודות המאורגנות בתיקיות לפי גובה הציונים: בכל תיקייה כמה עבודות מחלק כלשהו של טווח הציונים, התיקיות מסודרות בסדר עולה (או יורד) של חלקי טווח הציונים, והעבודות בכל אחת מהתיקיות מסודרות בסדר עולה (או יורד) של ציונים. לדוגמה, בתיקיה מס' 1 יש שלוש עבודות שציוניהן 25, 28 ו-30, ואילו בתיקיה מס' 2 יש עבודות שציוניהן 34, 38 ו-41. השופטים יודעים שהעבודות והתיקיות מסודרות לפי מדרג הציונים, אבל אינם יודעים מה הציון המדויק שקיבלה כל עבודה. הם מתבקשים לדרג כל עבודה על-פי רמות הביצוע שהתאמנו בהן בשלב הראשון. אף שאת הדירוג עושה כל שופט בנפרד, השופטים יכולים להתייעץ זה עם זה ולשנות את דירוגיהם לפי הצורך. המטרה העיקרית בשלב זה היא לקבל הערכה גסה של ציוני החתך לקראת סבב השיפוט הבא. לשם כך בתום הדירוג הצוות המנהל בוחר את דירוגי השופטים ומצמצם את מאגר העבודות על ידי הוצאתן של העבודות שאינן תורמות לקבלת ההחלטה. לדוגמה, אם כל השופטים דירגו את העבודות בתיקיה מס' 1 כ"מתחת לרמה הנדרשת", פירוש הדבר כי אין ספק שהשופטים מסכימים על הקריטריונים להערכת עבודות ברמה זו, ואין צורך לשוב ולדרגן בסבב הבא. יתרה מזו, סביר להניח שציון החתך לא יהיה בטווח הציונים שבתיקיה מס' 1. במקום עבודות אלו אפשר להוסיף עבודות אחרות, הקרובות יותר לקו התפר בין רמת ביצוע אחת



לאחרת, וכך אפשר למקד טוב יותר את התהליך של קביעת ציון החתך. בסינון העבודות לקראת הסבב הבא המטרה העיקרית היא לקבל מדגם עבודות ששונות הדירוגים בו גדולה, כלומר, שאי-ההסכמה על הדירוג המתאים לעבודות גדולה יותר. מספר העבודות שיוותרו לסבב השני צריך להיות גדול מספיק כדי לספק בסיס רחב דיו לקביעת כל ציוני החתך, אך עם זאת לא גדול מדי על מנת שלא להקשות מדי על השופטים.

בשלב הבא, **שלב האיתור הממוקד** (pinpointing), נעשה סבב שיפוט שני. סבב זה מתבסס על מדגם העבודות שנוצר מהתהליך של מציאת הטווחים בשלב הקודם; כלומר, השופטים עושים את אותה פעולה כמו בשלב הקודם, אלא שאוסף העבודות שהם מתבססים עליו נמצא סביב אזורי התפר בין רמות ביצוע סמוכות.

בשלב האחרון בתהליך, שלב הקביעה של ציוני החתך, אפשר להשתמש במגוון שיטות אנליטיות להחלטה על הערכים המתאימים. המטרה המרכזית היא למצוא את הציון הגולמי שעבורו ההסתברות שהעבודה תדורג ברמה כלשהי גדולה מההסתברות שהיא תדורג ברמה אחת מתחתיה או שווה לה, כלומר, הסתברות של 50%. השיטה המקובלת למציאה של ציוני החתך היא רגרסיה לוגיסטית (ראו הרחבה למעלה), שבה הציון הגולמי הוא חזאי לרמת הביצוע שהעריכו השופטים בסבב האחרון.

#### 5.4 שיטות פשרה – שיטת הופסטי

שיטת הופסטי (Hofstee, 1983) היא אחת משיטות הפשרה לקביעת סטנדרטים. על-פי העיקרון המנחה שיטות אלו התהליך של קביעת הסטנדרטים טומן בחובו שיקולים ואינטרסים שונים ממגוון גורמים: חינוכיים, פוליטיים, כלכליים, פרקטיים ועוד, ועל כן קביעה של ציון החתך היא בבחינת פשרה בין גורמים אלו. יתרה מזו, השיטה שהופסטי פיתח מבוססת על התפיסה שקביעת הסטנדרט אינה מוחלטת ואינה מבוססת-קריטריון במלוא מובן המילה. למעשה, שיפוט אנושי תמיד משתמש במידה כלשהי במידע נורמטיבי. לדוגמה, כאשר שופטים מתבקשים לדמין תלמיד שיכולתו מינימלית, סביר להניח שיעלו בדמיונם מקרים ספציפיים של תלמיד שגילה יכולת מינימלית. ואולם השופטים אינם יכולים לנתק את התלמיד הספציפי משאר התלמידים שהם מכירים: בהכרח - שיפוט יכולתו של התלמיד מושפע גם מהשוואתו לתלמידים אחרים, ועל כן יש קשר בין הקריטריון ובין הנורמה, קשר שלא ניתן להתעלם ממנו.

שיטת הופסטי היא שיטה פשוטה, ולא דרוש לה זמן רב. לאחר שלב האימון וההיכרות השופטים מתבקשים לענות על ארבע שאלות בלבד. בהנחה שזו הבחינה הראשונה של כל הנבחנים, שתי שאלות מתמקדות ברמת הידע שנבחנים צריכים לגלות, ושתי שאלות מתמקדות באחוז הנכשלים שניתן לשאת. להלן נוסח מתורגם של השאלות והסימון המתאים לערך המתקבל מכל שאלה.

1. להערכתך, מהו הערך הגבוה ביותר - במונחים של אחוז התשובות הנכונות במבחן - שיכול להיחשב לקביל עבור ציון החתך, אפילו אם כל הנבחנים יקבלו ציון זה או גבוה ממנו?

— זו הערכה של הידע המרבי הנדרש מהנבחנים ( $k_{max}$ ).

2. להערכתך, מהו הערך הנמוך ביותר - במונחים של אחוז התשובות הנכונות במבחן - שיכול להיחשב לקביל עבור ציון החתך, אפילו אם שום נבחן לא יקבל ציון זה?

— זו הערכה של הידע המינימלי הנדרש מהנבחנים ( $k_{min}$ ).

3. להערכתך, מהו הערך הגבוה ביותר שיכול להיחשב לקביל עבור אחוז הנכשלים במבחן?

— זו הערכה של אחוז הנכשלים המרבי שניתן לשאת ( $f_{max}$ ).

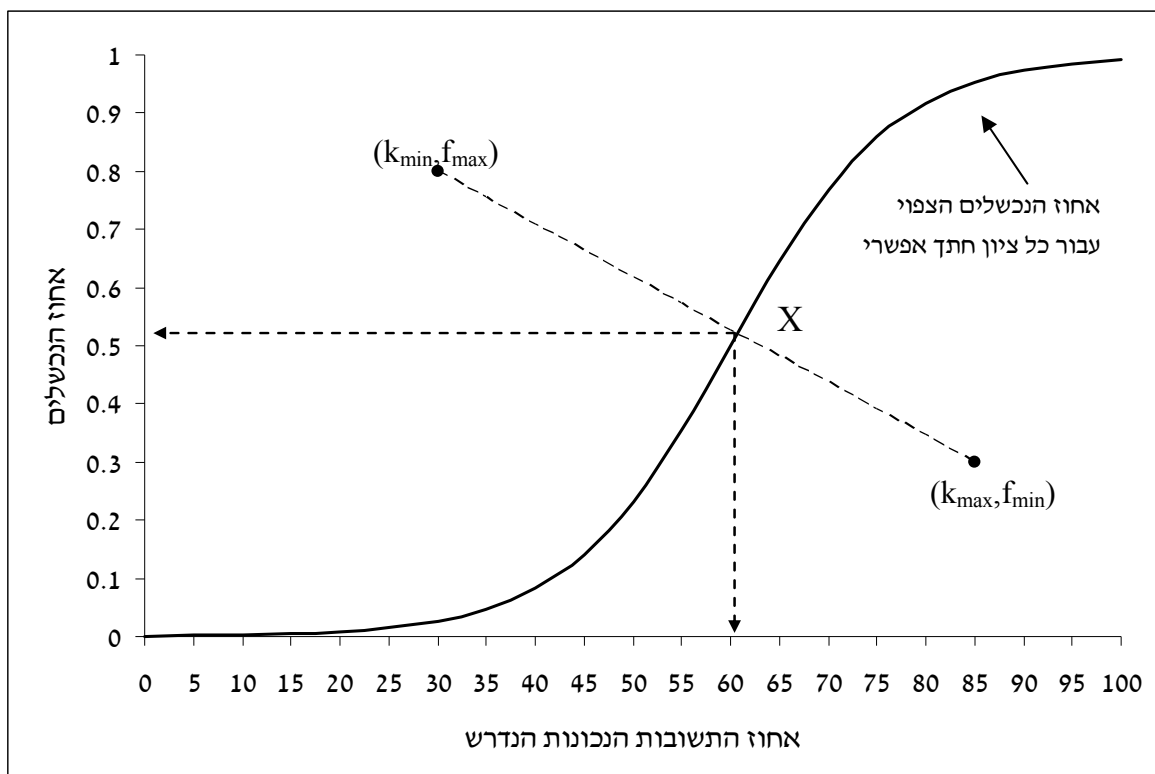
4. להערכתך, מהו הערך הנמוך ביותר שיכול להיחשב לקביל עבור אחוז הנכשלים במבחן?

— זו הערכה של אחוז הנכשלים המינימלי שניתן לשאת ( $f_{min}$ ).

הערכים שהתקבלו מכל השופטים נאספים והממוצע שלהם משמשים לקביעה של ציון

החתך כך: צמדי הערכים ( $k_{min}, f_{max}$ ) ו- ( $k_{max}, f_{min}$ ) מסומנים כנקודות על גרף שבו אחוז התשובות הנכונות הנדרש מופיע בציר האופקי, ואחוז הנכשלים מופיע בציר האנכי, כפי שנראה בתרשים 3. בין שתי נקודות אלו מצויר קו ישר, – ועל-פי הופסטי עליו נמצאת נקודת הפשרה של ציון החתך. בתרשים 3 מופיעה גם עקומה המבוססת על התפלגות הציונים האמיתית. העקומה מייצגת את אחוז הנכשלים הצפוי עבור כל ציון חתך אפשרי. נקודת המפגש בין הקו לעקומה מייצגת את ציון החתך הסופי על-פי שיטת הופסטי. בתרשים 3 נקודה זו מסומנת באות X, והיא מייצגת ציון חתך של 60.7%, ואחוז הנכשלים הצפוי הוא 52%.

תרשים 3. הדגמה של קביעת נקודת החתך בשיטת הופסטי



לרוב, שיטת הופסטי מוצעת כגיבוי לשיטה אחרת של קביעת סטנדרטים. מכיוון שלעבודת השופטים לא דרוש זמן רב, אפשר לאסוף את התשובות לארבע השאלות בתום סבב השיפוט בשיטה אחרת של קביעת סטנדרטים. אף שסיזק ובונץ (Cizek & Bunch, 2007) מתריעים מפני שימוש בכמה שיטות בעת ובעונה אחת, נראה כי המשאבים המועטים הדרושים לאיסוף הנתונים בשיטת הופסטי מאפשרים שימוש בשיטה זו יחד עם שיטות אחרות. אחד החסרונות הבולטים של שיטת הופסטי הוא אי-התאמתה לקביעה של כמה נקודות חתך במתכונתה העכשווית. על אף פשטותה השימוש בה אינו נרחב, ולא נעשה עליה די מחקר בהשוואה לשיטות אחרות.

## 6. בקרת איכות ותיקוף הסטנדרטים

כל תהליך של קביעת סטנדרטים של ביצוע מחייב תהליך מלווה של בקרת איכות ותיקוף. ברמה הבסיסית ביותר יש לוודא שהתהליך כולו, מבחירת צוות השופטים ועד לקביעת ציוני החתך הסופיים, נעשה בדרך תקינה. בספרות מופיעות כמה מערכות של הנחיות לניהול תקין של תהליך קביעת הסטנדרטים, לדוגמה: Cizek, 1996; Hamblen, 1998. הדגש בהנחיות אלו הוא בתכנון מוקדם ובביצוע מדויק של התהליך על-פי השיטה שנבחרה, בתיעוד מדויק של התהליך ובהערכתו ובתיקוף התוצר הסופי - ציוני החתך. קיין (Kane, 1994) מציין שיש לבדוק שתי הנחות הנוגעות לציוני החתך: (1) שציון החתך תקף; כלומר, שבסבירות גבוהה, נבחן שציונו גבוה מציון החתך אכן השיג את הסטנדרט (דהיינו, שולט בחומר כפי שמצוין בתיאור רמת הביצוע), ונבחן שציונו נמוך מציון החתך לא השיג את הסטנדרט; (2) שהסטנדרט לביצוע הוא סביר בהתאם למטרות של קבלת ההחלטות על בסיס המבחן. לשם תיקוף ההנחות האלה יש לאסוף עדויות רלוונטיות לאורך כל התהליך.

המבלטון ופיטוניאק (Hamblen & Pitoniak, 2006) פורטים את תהליך התיקוף לבדיקות משלושה סוגים: בדיקות נוהליות, בדיקות פנימיות ובדיקות חיצוניות.

בדיקות נוהליות מתייחסות להרכב של צוות השופטים ולשיטה הנבחרת. בבדיקות אלו יש לבדוק: (א) באיזו מידה הוצג התהליך באופן ברור ומפורש לפני ביצועו; (ב) באיזו מידה היה התהליך קל לביצוע, אמין ומובן להדיוטות; (ג) באיזו מידה נעשה התהליך באופן שיטתי וממצה; (ד) באיזו מידה חש צוות השופטים בנוח עם מרכיבי התהליך כולו; (ה) באיזו מידה תועד התהליך ונבחן לשם בקרת איכות. עדויות לניהול תקין של התהליך תורמות מאוד לביסוס הסטנדרטים הנבחרים ולצידוקם (Carson, 2001).

בדיקות פנימיות מתייחסות לעקיבות הפנימית בביצוע התהליך. בבדיקות אלו יש לבדוק: (א) באיזו מידה בחזרה על התהליך עשויים להתקבל אותם ציוני החתך שהתקבלו; (ב) באיזו מידה הערכות השופטים, למשל הערכת קושי הפריטים, תואמות עדויות אמפיריות הנוגעות למבחן; (ג) באיזו מידה שיפוט השופטים משתנים מסבב לסבב; (ד) באיזו מידה השופטים מסכימים על הערכות קושי של הפריטים, על הערכות של ביצועי הנבחנים ועל ציוני החתך.

בדיקות חיצוניות מתייחסות לקריטריונים חיצוניים שאפשר להשוות אליהם את הסטנדרטים. בבדיקות אלו יש לבדוק: (א) באיזו מידה ציוני החתך שנקבעו בשיטה הנבחרת דומים לציוני החתך שהיו נקבעים בשיטות אחרות; (ב) אם יש קשר בין החלטות שיתקבלו על סמך ציוני החתך האלה ובין קריטריונים אחרים, כגון מבחן דומה; (ג) באיזו מידה ציוני החתך שהתקבלו סבירים ובני-ביצוע.

## 7. סיכום

פרק זה סקר את הנושא של הגדרת סטנדרטים של תוכן והתמקד בשיטות להצבת סטנדרטים לביצוע, דהיינו, לקביעה של נקודות חתך. סיכום השיטות השונות מותג בטבלה 2. אף שהשימוש בסטנדרטים במערכות חינוך עשוי לשפר את ההוראה, את ההערכה ואת הלמידה - ישנם לא מעט גורמים המקשים על ההצלחה בקביעה של נקודות חתך למבחן. השיטות שהוצגו בפרק זה נועדו, בין השאר, לייעל את התהליך של קביעת נקודות החתך ולהתגבר על מקצת המגבלות של התהליך. על ידי בחירה של צוות שופטים מיומן, שימוש אחיד בשיטה שנקבעה, הערכה בכמה סבבי שיפוט, מתן משוב לשופטים ותיקוף התהליך - אפשר להעלות את הסיכוי שציון החתך במבחן אכן ישקף את רמת הידע, את היכולת ואת הביצוע המוגדרים על-פי סטנדרט התוכן המתאים. עם זאת, ריבוי השיטות לקביעה של ציוני חתך עשוי להקשות על המשתמשים המקווים לבחור ב"שיטה הטובה ביותר" לצורכיהם. השיטות המיושמות כיום בחינוך הן בעיקר שיטת אנגוף ושיטת הסימנייה, אבל יש גם מקרים שבהם משתמשים בשיטות אחרות - לפי הצורך הספציפי בכל מקרה. ההחלטה באיזו שיטה יש להשתמש בקביעה של ציוני החתך נעשית לרוב בדרג גבוה ומונעת משיקולים מהותיים וטכניים כאחד.

חשוב להבין שאין שיטה אחת עדיפה באופן מוחלט על פני כל שאר השיטות (Zikey, 2001). לשיטות שונות יש יתרונות במצבים שונים. מחקרים שונים מנסים להשוות בין שיטות מקובלות לשיטות חדשות במטרה לנסח הנחיות שיסייעו לבחור את השיטה המתאימה על-פי צורכי המבחן, לדוגמה: Buckendahl, Smith, Impara & Plake, 2002. עם זאת, ריבוי המצבים שבהם משתמשים בסטנדרטים, התכנים השונים של המבחנים ואופיים וכן השונות בין השופטים – כל אלה מקשים על קביעת מסקנות כלליות הנוגעות לעדיפויות של שיטה זו או אחרת. על אף קשיים אלו מחקרים מדעיים המשווים בין שיטות שונות לקביעה של ציוני חתך הכרחיים לשיפור התהליך של קביעת הסטנדרטים.

על אף היתרונות הרבים של סטנדרטים במערכות חינוך ביקורת לא מועטה נמתחת על אופן הקביעה של הסטנדרטים - בין שהם סטנדרטים של תוכן ובין שהם סטנדרטים של ביצוע, על השימוש בהם ועל ההשפעה שלהם על התכנים הנלמדים ועל אופי הלימוד. על-פי חשש אחד הסטנדרטים יסייעו לשמר פערים חברתיים קיימים ויעוררו תחושות תסכול וחוסר ביטחון בקרב תלמידים מאוכלוסיות חלשות. מרזאנו וקנדל (Marzano & Kendall, 1996) הצביעו על טענות אחרות: (א) לבניית הסטנדרטים וליישומם דרושים משאבים רבים הניתנים לניצול טוב יותר בשיפור החינוך; (ב) השימוש בסטנדרטים מגביר את הלחץ על המורים ועל תלמידים מאוכלוסיות חלשות; (ג) ישנן דוגמאות היסטוריות לכישלונות של יישום סטנדרטים בחינוך; (ד) הסטנדרטים

מקיפים חומר נרחב מדי, ולימודו אינו בר-מימוש ; (ה) לעתים התוכן המוגדר בסטנדרטים שנוי במחלוקת. חוקרים אחרים הביעו ביקורת גם על התהליך של קביעת ציוני החתך. הטענות העיקריות שלהם גורסות שקביעת הציון שרירותית ואף גחמנית, לא-מבוססת ומייצגת תובנות של קובעי הסטנדרטים ולא דווקא תכונות/יכולות אמיתיות של תלמידים (Shepard, 1984). על אף ביקורות אלו השימוש בסטנדרטים במערכות חינוך נפוץ מאוד משום שהוא נובע מהצורך הבלתי נמנע לפרש ציונים שלא באופן נורמטיבי ולקבל החלטות בנוגע לתלמידים, למורים ולמוסדות.

נספח א. מאפיינים עיקריים של שבע שיטות לקביעת סטנדרטים לביצוע

| שיטה                        | מאמר עיקרי                             | תיאור קצר                                                                                                                                                                                                                                      | השופטים<br>מתמקדים ב-                                     | שימוש עיקרי                      | התייחסות<br>לכל<br>הפריטים* | התייחסות לנבחן/<br>לקבוצה שיכולותיה | קביעת כמה<br>נקודות<br>***חתך | שימוש<br>בנתונים<br>קיימים**** |
|-----------------------------|----------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------|----------------------------------|-----------------------------|-------------------------------------|-------------------------------|--------------------------------|
| נדלסקי                      | Nedelsky, 1954                         | זיהוי המסחים שנבחן שיכולתו מינימלית יכול לפסול. הסתברות לתשובה נכונה בפריט היא ההופכי של מספר התשובות שנותרו. ציון החתך הוא סכום ההסתברויות של כל הפריטים.                                                                                     | מסחים בתוך פריטים                                         | מבחני רישוי והסמכה               | כן                          | כן                                  | לא                            | לא                             |
| אבל                         | Ebel, 1972                             | מיון פריטים לקבוצות על-פי רמת הקושי והרלוונטיות. הערכה של אחוז הפריטים מכל קבוצה שתלמידים שיכולתם מינימלית יצליחו לפתור נכון. ציון החתך הוא סכום מכללת מספר הפריטים והאחוז בכל קבוצה.                                                          | סיכויי תשובה נכונה בפריטים המקובצים על-פי קושי ורלוונטיות | הערכה ברפואה, הערכה בחינוך       | כן                          | כן                                  | אפשרי - בשינוי קל             | לא בהכרח                       |
| אנגוף                       | Angoff, 1971                           | הערכת ההסתברות - באופן רציף או דיכוטומי - שנבחן גבולי יענה נכון על כל אחד מהפריטים במבחן. ציון החתך הוא סכום ההערכות של כל הפריטים.                                                                                                            | הסיכוי לבחור בתשובה נכונה בפריטים                         | מבחני רישוי והסמכה, הערכה בחינוך | כן                          | כן                                  | כן                            | לא בהכרח                       |
| סימנייה                     | Lewis, Mitzel & Green, 1996            | בחברת המסודרת על-פי קושי הפריטים - זיהוי המקום שבו ההסתברות שנבחן גבולי יענה נכון על הפריט היא מתחת לערך מסוים, לדוגמה 0.5. ציון החתך מבוסס על הקושי של הפריט על-פי IRT.                                                                       | סיכויי תשובה נכונה בפריטים                                | הערכה בחינוך                     | כן                          | כן                                  | כן                            | כן                             |
| קבוצות מנוגדות /קבוצות גבול | Berk, 1976<br>Zieky & Livingston, 1977 | קבוצות מנוגדות - חלוקת מדגם של נבחנים אמיתיים לשולטים/לא-שולטים על סמך היכרות מוקדמת וקביעת ציון החתך על-פי נקודת החתך בין התפלגויות שתי הקבוצות. קבוצת גבול - זיהוי קבוצת הנבחנים הגבוליים וקביעת ציון החתך על-פי הנטייה המרכזית של התפלגותם. | מאפייני נבחנים בקבוצות שונות                              | הערכה בחינוך                     | לא                          | כן                                  | כן                            | כן                             |
| אוסף עבודות                 | Kingston, Kahl, Sweeney & Bay, 2001    | הערכה הוליסטית של כמה רמות ביצוע של תלמידים על בסיס אוסף עבודותיהם. ציוני החתך נקבעים על ידי גרסיה לוגיסטית של רמות הביצוע על ציוני העבודות.                                                                                                   | אוסף עבודות של תלמידים                                    | הערכה בחינוך                     | כן                          | כן                                  | כן                            | כן                             |
| הופסטי                      | Hofstee, 1983                          | הערכת הערך הגבוה/הנמוך ביותר של ציון החתך והערך הנמוך/הגבוה ביותר של אחוז הנכשלים. ציון החתך נמצא בנקודת המפגש בין הקו המוגדר על ידי נקודות אלו ובין עקומת הקשר בין ציון החתך לאחוז הנכשלים.                                                   | רמת ידע של נבחנים ואחוז הנכשלים                           | מבחני רישוי והסמכה               | לא                          | לא                                  | לא                            | כן                             |

הערות: \* האם השיטה מחייבת שהשופטים יתייחסו לנבחן או לקבוצת נבחנים שיכולותיהם מינימליות? \*\* האם השיטה מחייבת שהשופטים יתייחסו לנבחן או לקבוצת נבחנים שיכולותיהם מינימליות? \*\*\* חתך: האם השיטה מאפשרת לקבוע כמה נקודות חתך? \*\*\*\* האם השיטה מחייבת שימוש בנתונים קיימים על ביצוע במבחן?

## 8. שאלות תרגול

### שאלה 1

מהו סטנדרט של ביצוע? כיצד הוא נבדל מסטנדרט של תוכן?

### שאלה 2

לפניכם שני משפטים. איזה משפט מתאים יותר להיחשב סטנדרט של ביצוע? מדוע?

א. תלמידים ברמה "גבוהה" יכולים להשלים באופן מלא פעלים מאותו שורש ובאותו בניין על-פי ההקשר של המשפט.

ב. תלמידים ברמה "גבוהה" חייבים לקבל לפחות 85 נקודות במבחן של ידע לשוני.

### שאלה 3

מרבית השיטות לקביעה של ציוני חתך משתמשות במושג "נבחנים גבוליים". מהם נבחנים גבוליים, ומדוע מתמקדים דווקא בקבוצה זו?

### שאלה 4

תארו בקצרה את שיטת אנגוף המקורית.

### שאלה 5

במבחן מסכם באנגלית ישנן שתי קבוצות פריטים :

א. 40 פריטים עם חמש אפשרויות תשובה

ב. 21 פריטים עם שלוש אפשרויות תשובה.

קבוצת שופטים השתמשה בשיטת נדלסקי לקביעה של ציון החתך במבחן. לאחר סיום ההליך חילקו השופטים את הפריטים לתת-קבוצות האלה :

1. ב-10 פריטים מקבוצה א נבחנים גבוליים יוכלו לפסול רק אפשרות תשובה אחת.

2. ב-15 פריטים מקבוצה א נבחנים גבוליים יוכלו לפסול שתי אפשרויות תשובה.

3. ב-15 פריטים מקבוצה א נבחנים גבוליים לא יוכלו לפסול אף אפשרות תשובה.

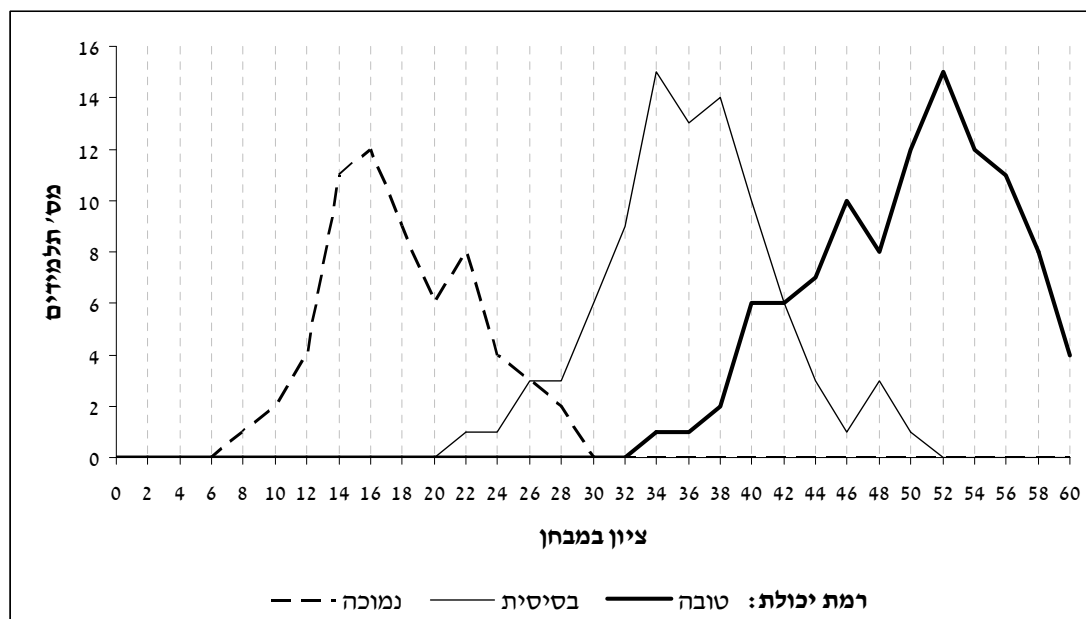
4. ב-12 פריטים מקבוצה ב נבחנים גבוליים יוכלו לפסול רק אפשרות תשובה אחת.

5. ב-9 פריטים מקבוצה ב נבחנים גבוליים לא יוכלו לפסול אף אפשרות תשובה.

מהו ציון החתך במבחן? הראו את דרך הפתרון.

### שאלה 6

לפניכם תרשים של שכיחות הציונים במבחן כלשהו עבור נבחנים משלוש קבוצות יכולת.



על-פי שיטת הקבוצות המנוגדות, היכן יש לקבוע את ציוני החתך בין רמה נמוכה לבסיסית ובין רמה בסיסית לטובה?



- American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 508-600). Washington, DC: American Council on Education.
- Berk, R. A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education*, 45, 4-9.
- Berk R. A. (1984). *A guide to criterion-referenced test construction*. Baltimore, MD: The Johns Hopkins University Press.
- Buckendahl, C. W., Smith, R. W., Impara J. C., and Plake, B. S. (2002). A comparison of Angoff and Bookmark standard setting methods. *Journal of Educational Measurement*, 39, 253-263.
- Carson, J. D. (2001). Legal issues in standard setting for licensure and certification. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 427-444). Mahwah, NJ: Erlbaum.
- Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement*, 30, 93-106.
- Cizek, G. J. (1996). Standard setting guidelines. *Educational Measurement: Issues and Practice*, 15, 12-21.
- Cizek, G. J. (2001). *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Erlbaum.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.
- Downing, S. M., Lieska, N. G., & Raible, M. D. (2003). Establishing passing standards of classroom achievement tests in medical education: A comparative study of four methods. *Academic Medicine*, 78, S85-S87.
- Ebel, R. L. (1972). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall.

- Hambleton, R. K. (1998). Setting performance standards on achievement tests. In L. H. Hansche (Ed.), *Handbook for the development of performance standards: Meeting the requirements of Title I*. Washington, DC: U.S. Department of Education. Netherlands: IEA.
- Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement*, 24, 355-366.
- Hambleton, R. K., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433-470). Westport, CT: Praeger.
- Hambleton, R. M., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education*, 8, 41-56.
- Hofstee, W. K. B. (1983). The case for compromise in educational selection and grading. In S. B. Anderson & J. S. Helmick (Eds.), *On educational testing* (pp. 109-127). San Francisco, CA Jossey-Bass.
- Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement*, 34, 353-366.
- Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64, 425-461.
- Karantonis, A., & Sireci, S.G. (2006). The bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice*, 25, 4-12.
- Kingston, N. M., Kahl, S. R., Sweeney, K., & Bay, L. (2001). Setting performance standards using the body of work method. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 219-248). Mahwah, NJ: Erlbaum.
- Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard setting procedures utilizing behavioral anchoring*. Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.
- Marzano, R. J., & Kendall, J. S. (1996). *The fall and rise of standards-based education. A National Association of State Boards of Education issues in brief*. Washington, DC: National Association of State Boards of Education.

- Melican, G. J., Mills, C. N., & Plake, B. S. (1989). Accuracy of item performance predictions based on the Nedelsky standard setting method. *Educational and Psychological Measurement*, 49, 467-478.
- Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, 14, 3-19.
- Plake, B. S., & Hambleton, R. K. (2001). The analytic judgment method for setting standards on complex performance assessments. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 283-312). Mahwah, NJ: Erlbaum.
- Ravitch, D. (1995). *National standards in American education: A citizen's guide*. Washington, DC: Brookings.
- Shepard, L. A. (1984). Setting performance standards. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 169-198). Baltimore, MD: John Hopkins University Press.
- Shepard, L. A., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement*. Stanford: National Academy of Education.
- Wang, N. (2003). Use of the Rasch IRT model in standard setting: An item mapping method. *Journal of Educational Measurement*, 40, 231-253.
- Zieky, M. J. (2001). So much has changed: how the setting of cutscores has evolved since the 1980s. In G. J. Cizek, (Ed.), *Setting Performance Standards: Concepts, methods, and perspectives* (pp. 19-52). Mahwah, NJ: Erlbaum.
- Zieky, M. J., & Livingston, S. A. (1977). *Manual for setting standards on the Basic Skills Assessment Tests*. Princeton, NJ: Educational Testing Service.
- Zieky, M. J., Perie, M. & Livingston, S. A. (2008). *Cutscores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

## תשובות לשאלות תרגול

### תשובה 1

סטנדרט של ביצוע הוא הציון שתלמיד נדרש לקבל במבחן כדי להיחשב כבעל רמת ביצוע מסוימת (לדוגמה, עובר/לא עובר או נמוך/בינוני/גבוה). על כן, המונח סטנדרט של ביצוע מקביל למונח ציון חתך. למבחן מסוים אפשר לקבוע ציון חתך אחד או יותר, וכל ציון חתך מציין את נקודת המעבר בין רמת ביצוע אחת לאחרת. ציון החתך משקף את הידע ואת הכישורים הנדרשים ברמת ביצוע מסוימת במבחן.

סטנדרט של תוכן כולל את הידע ואת הכישורים הנלמדים בתוכנית הלימודים. הסטנדרט הוא הגדרה מפורטת של התכנים שמורים צריכים ללמד, האופן שבו תכנים אלו צריכים להילמד והידע והכישורים שתלמידים צריכים לפתח בעקבות תהליך זה של הלמידה. סטנדרטים של תוכן משמשים לבניית תוכניות לימוד, לקביעת מטרות לימודיות וציוני דרך, לבניית מבחנים ולסיוע בפירוש תוצאותיהם.

בעוד סטנדרט של תוכן הוא תיאור של הידע והכישורים הרצויים בתוכנית הלימודים, סטנדרט הביצוע הוא ציון המשקף רמת ביצוע מסוימת במבחן המיועד לתוכנית הלימודים הזו.

### תשובה 2

משפט ב מתאים יותר להיות סטנדרט של ביצוע משום שהוא כולל את ציון החתך במבחן.

### תשובה .

נבחנים גבוליים הם נבחנים הנמצאים על קו התפר בין ביצוע אופייני לבעלי יכולת ברמה מסוימת ובין ביצוע שהוא נמוך או גבוה מרמה זו. לדוגמה, אם במבחן יש שתי רמות ביצוע, עובר ולא עובר, הנבחנים הגבוליים הם מי שהביצוע שלהם קרוב לגבול בין שתי הרמות. לנבחנים כאלה יש מאפיינים של שתי רמות הביצוע; כלומר, הם טובים יותר מהנבחן הטיפוסי בקבוצת ה-"לא עוברים" וטובים פחות מהנבחן הטיפוסי בקבוצת ה-"עוברים". נבחנים כאלה עוזרים להבליט את הכישורים ואת הידע שנדרשים כדי להיחשב ברמת ה-"עובר" בדיוק בגלל שהם קרובים לרמה זו, אך לא משיגים אותה. אילו היו השופטים מתמקדים בנבחנים חלשים מדי, כלומר, במי שנמצאים הרבה מתחת לקו הגבול בין שתי הרמות, הם לא היו נחשפים לביצוע ברמת ה-"עובר". אילו היו השופטים מתמקדים בנבחנים חזקים מדי, כלומר, במי שנמצאים הרבה מעל לקו הגבול בין שתי הרמות, הם לא היו נחשפים לביצוע ברמת ה-"לא עובר". על ידי מיקוד השופטים בקבוצה הגבולית אפשר להגדיר טוב יותר את המאפיינים של נקודת המעבר בין רמות ביצוע שונות.

### תשובה 4

בשיטת אנגוף המקורית צוות השופטים מתמקד בקבוצת תלמידים הנמצאים על קו הגבול בין "נכשל" ל"עובר". השופטים עוברים על כל פריט במבחן ומעריכים את האחוז מתוך קבוצת הגבוליים שיענו נכון על הפריט. סכום האחוזים מעבר לכל הפריטים הוא ציון החתך במבחן.

#### תשובה 5

ראשית, יש לחשב את דירוג נדלסקי לכל תת-קבוצת פריטים. להלן החישובים בהתאמה :

$$1. \quad 10 \times \frac{1}{4} = 2.5$$

$$2. \quad 15 \times \frac{1}{3} = 5.0$$

$$3. \quad 15 \times \frac{1}{5} = 3.0$$

$$4. \quad 12 \times \frac{1}{2} = 6.0$$

$$5. \quad 9 \times \frac{1}{3} = 3.0$$

סכום דירוגי נדלסקי מעבר לתת-הקבוצות הוא  $2.5 + 5 + 3 + 6 + 3 = 19.5$ . על כן ציון החתך במבחן הוא 19.5.

#### תשובה 6

ציון החתך בין רמה נמוכה לבסיסית הוא 26, וציון החתך בין רמה בסיסית לטובה הוא 42.