

פרק 6

הערכה של תוכניות לימודים לפיתוח החשיבה

הערכה של תוכנית חינוכית פירושה מדידת תרומתה של התוכנית להישגי התלמידים. תכנון וביצוע של הערכת תוכניות לימודים לפיתוח החשיבה הוא עניין מסובך. חלק מהסיבוך אופייני לכל הערכה חינוכית באשר היא, אך יש גם קשיים המתעוררים מעצם טיבו של הדבר שאותו מנסים למדוד, כלומר החשיבה. נפתח את הפרק בהצגת הבעיה, לאחר מכן נתאר כמה דוגמאות ממבחנים 'קלאסיים' לחשיבה, ונסביר את הקשיים שהם מעוררים, ונסיים בהתוויה של כיווני הערכה אלטרנטיביים.

הגדרת מטרות

בעיה יסודית בכל דיון על הערכה היא השאלה מה בעצם אנו רוצים למדוד. שאלה זו מצדה צריכה להיגזר מן השאלה **מה רוצים להשיג** בתוכנית מסוימת המטפחת חשיבה. כאן ראוי להקפיד על התאמה מקסימלית בין צורת ההערכה (איך מודדים) לבין מטרות התוכנית (מה מודדים). כל חוסר התאמה בין שני מרכיבים אלה יביא להטיה של המסקנות שנגזר מן ההערכה.

כדי להציג את הבעיה הכרוכה בהערכת תוכניות לימודים לפיתוח החשיבה, נתאר מחקר מסכם שבו נידונו התוצאות של שורת מחקרים קודמים. מקמילן (McMillan, 1987) סקר עשרים ושבעה מחקרים שבדקו את השפעתן של שיטות הוראה ותוכניות לימודים על חשיבה ביקורתית של תלמידי קולג'. ברוב המחקרים שנסקרו לא נמצא הבדל מובהק בין השיפור בהישגים של קבוצת הניסוי שקיבלה

הכשרה בחשיבה ביקורתית לבין השיפור בהישגים של קבוצת ביקורת שלא קיבלה הכשרה כזאת.

מקמילן שואל מדוע לא נמצא הבדל ומסביר כי תיתכנה שתי תשובות אפשריות. אפשרות אחת היא שהתוכניות אכן לא תרמו לשיפור החשיבה. אפשרות שנייה היא שהתוצאות נובעות מבעיות במדידה שמקורן בכלי ההערכה שבו השתמשו. מבחן ווטסון-גלזר לחשיבה ביקורתית, שבו השתמשו בכשני-שליש מהמחקרים שנסקרו על ידי מקמילן, מדגים את הקשיים שמציבים מכשירי ההערכה הקיימים. המבחן פותח מלכתחילה כדי לבדוק את היכולת לחשוב באופן ביקורתי על תכנים שנתקלים בהם במסגרת חיי היום-יום (במסגרת העבודה או בעת קריאת עיתונים). לפיכך המבחן כולל בעיות, הצהרות, טיעונים ופירוש של נתונים שתוכנם לקוח מנושאים יומיומיים. מכאן יוצא שפריטי המבחן מודדים יכולות רחבות וכלליות מאוד.

מבנה המבחן מציב שתי בעיות: הבעיה הראשונה היא במעבר מתחום תוכן אחד לשני. אם נתרגם את דבריו של מקמילן למערכת המושגית שנידונה בפרקים הקודמים, הרי שמבחנים הבודקים יכולות חשיבה בתכנים כלליים אינם יכולים לשמש לבדיקת יכולות חשיבה שנלמדו בתכנים ספציפיים, משום שהם מניחים העברה. ייתכן שהיה שיפור ביכולת החשיבה בתכנים שהתוכנית לפיתוח החשיבה עסקה בהם, אולם הדבר לא ניכר בתוצאות ההערכה משום שפריטי המבחן עוסקים בתחום תוכן אחר, ואין העברה מלאה של מיומנויות חשיבה מתחום תוכן אחד לתחומי תוכן אחרים.

בעיה שנייה הכרוכה במבנה המבחן מתעוררת מהעדר חפיפה בין מטרות התוכנית לפיתוח החשיבה לבין מטרות המבחן. אם אין חפיפה מלאה בין שתי רשימות המטרות, סביר להניח שכלי המדידה לא יצליח לגלות שיפור, גם אם הפרויקט החינוכי אכן גרם לשינויים ביכולות חשיבה ספציפיות. דוגמא קיצונית לקושי כזה עשויה להתרחש כאשר מלמדים חשיבה יצירתית אך מעריכים את ההישגים באמצעות מבחן הבודק חשיבה ביקורתית. קשה לדמיין חוקרים שיתכננו מחקר הערכה כל כך בלתי-סביר. אולם הספרות מלאה

במקרים של מחקרי הערכה ברוח דומה. לדוגמא, תחת הכותרת 'חשיבה ביקורתית' עשויות להופיע במבחן מיומנויות חשיבה רבות. יש חוקרים שהשתמשו במבחן לחשיבה ביקורתית שבדק שורה של מיומנויות שלא הופיעו כלל בתוכנית הלימודים. מובן שקשה להניח ששיפור במיומנויות שנלמדו יגרור בעקבותיו גם שיפור בהישגי תלמידים במיומנויות אחרות שלא נלמדו.

אחת הדרכים להתמודד עם שני הקשיים האחרונים היא להשתמש במבחן מיוחד שיתאים לתחום התוכן ולמטרות הספציפיות של תוכנית לפיתוח חשיבה. כלומר, אם תוכנית לימודים מסוימת חוברה במטרה ללמד שורה של מיומנויות חשיבה בתחום תוכן מסוים (למשל בגיאוגרפיה או בהיסטוריה), אפשר להשתמש במבחן שיבדוק באופן ספציפי את המיומנויות האלה בתחום התוכן הנתון. כדי לבדוק העברה, ניתן להשתמש במבחן שיבדוק אותן מיומנויות חשיבה בתחום תוכן אחר. אולם המבחנים הדרושים לשם כך אינם מצויים תמיד בהישג ידם של החוקרים. נוריס ואניס (Norris & Ennis, 1989) מציינים כי ככל הידוע להם לא קיימים מבחנים לחשיבה ביקורתית בתחומי תוכן ספציפיים אשר נמכרים בשוק באופן מסחרי. הם מפרטים שורה של הצעות מעשיות שעשויות לעזור לאלה שיבחרו לחבר מבחן חדש שיתאים לצורכיהם יותר מהמבחנים הקיימים בשוק. יחד עם זאת, חשוב לזכור כי חיבור של מבחן תקף ומהימן הבדוק מיומנויות חשיבה דורש מומחיות לא מבוטלת. לכן נוריס ואניס ממליצים כי להוציא מקרים יוצאי דופן, רצוי להעדיף את המבחנים הקיימים משום שהם פותחו בתשומת לב ובהשקעה מרובה, ולכן, לדעתם, סביר שישרתו את המשתמש טוב יותר מאשר מבחנים שחוברו בחיפזון לצורך פרויקט זה או אחר. דבריהם מפנים אותנו למבחנים היקלאסיים לפיתוח חשיבה, שאכן פותחו תוך השקעת משאבים מרובים. שני מבחנים כאלה יידונו בסעיף הבא.

דוגמאות ממבחנים נפוצים לחשיבה ביקורתית

כדי להמחיש את הדברים שנאמרו בסעיף הקודם, נתאר להלן בקצרה שניים מן המבחנים הנפוצים לחשיבה ביקורתית.

מבחן ווטסון-גלזר לחשיבה ביקורתית (Watson & Glaser, 1964) שהוזכר בסעיף הקודם, חובר כבר בשנות השלושים ועדיין נמצא בשימוש נרחב. המבחן חובר מלכתחילה עבור תלמידי תיכון וקולג', אבל אפשר להשתמש בו גם בחטיבת הביניים. למבחן שני נוסחים מקבילים, בכל אחד מהם שמונים פריטים רבי-ברירה. שלושה מבין חמשת תת-המבחנים שמהם מורכב המבחן יתוארו להלן.

1. **תת-המבחן היסק** בודק באיזו מידה יכול הנבחן לשפוט האם מסקנה היא תקפה או לא. תת-המבחן מורכב מפסקאות המכילות מידע בנושאים שונים. בעקבות כל קטע מוצגים מספר פריטי מבחן. כל פריט הוא מסקנה שאפשר להסיק מן המידע הנתון בקטע. בכל פריט על הנבדקים לבחור בין חמש אפשרויות:

- * המסקנה אמיתית
- * סביר למדי שהמסקנה אמיתית
- * סביר למדי שהמסקנה שקרית
- * המסקנה שקרית
- * אין מספיק מידע על מנת להחליט

2. **תת-המבחן זיהוי הנחות.** הנחה חבויה היא פיסת מידע שמתבססים עליה בלי לומר זאת במפורש. לדוגמא, בשאלה "האם הפסקת להכות את בעלך?" חבויות שתי הנחות (לפחות): (א) שאת נשואה; (ב) שאת מכה את בעלך. תת-המבחן זיהוי הנחות בוחן את היכולת לזהות הנחות חבויות בטיעונים או היגדים. הפריטים של תת-המבחן כוללים מספר היגדים, שכל אחד מהם מתבסס על הנחות חבויות. לאחר כל הצהרה מוצעות מספר הנחות. המטלה היא להחליט בנוגע לכל אחת מההנחות המוצעות האם המצהיר מניח אותה בהכרח.

3. **תת-המבחן חשיבה דדוקטיבית** מכיל מספר פסקאות קצרות ומסקנות אפשריות. על הנבחנים להחליט לגבי כל מסקנה אם היא נובעת באופן דדוקטיבי מן המידע שבקטע.

(שני תת-המבחנים הנוספים, שבהם לא נדון כאן במפורש, הם **תת-המבחן פירוש ותת-המבחן הערכת טיעונים**.)

לוח 6.1 מציג שני פריטים לדוגמא מתוך מבחן ווטסון-גלזר. הדוגמא הראשונה לקוחה מתת-המבחן היסק ומורכבת מקטע מידע ומהיגד המובא בעקבותיו. הדוגמא השנייה לקוחה מתת-המבחן זיהוי הנחות.

לוח 6.1 - דוגמאות לפריטים ממבחן ווטסון-גלזר

דוגמא א':

מר בראון, המתגורר בעיר סאלס, הובא בפני בית-המשפט העירוני זו הפעם השישית במהלך החודש האחרון. הוא הואשם בכך שהשאיר את הפאב שלו פתוח לאחר השעה 1 לפנות בוקר. הוא שוב הודה באשמה ושוב נקנס ב-500 דולר (כלומר בעונש המירבי) כשם שנקנס בפעמים הקודמות.

בין הפריטים המובאים בעקבות הקטע הזה, מובא גם פריט מספר 6: בלילות מסוימים מר בראון רק הרוויח מהשאת הפאב שלו פתוח לאחר השעה 1 לפנות בוקר, למרות שבכך הסתכן בתשלום קנס של 500 דולר.

(התשובה הנכונה לפי המפתח היא: "סביר למדי שהמסקנה אמיתית.")

דוגמא ב':

"אני נוסע לדרום-אמריקה. אני רוצה להיות בטוח שלא אחלה בטיפוס, לכן אלך לרופא לפני היציאה לטיול ואקבל חיסון כנגד טיפוס."

הנחה מוצעת:

"טיפוס נפוץ יותר בדרום-אמריקה מאשר באזור שאני מתגורר בו."

(התשובה הנכונה לפי המפתח היא: "הנחה זו חבויה בהצהרה הנתונה." פריט זה לקוח מתת-מבחן 2)

* * *

כפי שחלק מן הקוראים ראה בוודאי בעצמו, התשובות הנכונות לפי המפתח אינן בהכרח התשובות האפשריות היחידות. נושא זה יידון

בפרוט בסעיף ב', העוסק בביקורת של המבחנים לחשיבה ביקורתית. מבחן נפוץ נוסף לחשיבה ביקורתית הוא **מבחן קורנל** (Ennis, Millman & Tomko, 1985). המבחן כולל שני נוסחים 'X' ו-'Z'. נוסח 'X' קל יותר ומיועד לתלמידים החל מגיל חטיבת הביניים. נוסח 'Z' נועד בעיקר למבוגרים. להלן נתרכז בתאור של נוסח 'X'. בנוסח זה ארבעה תתי-מבחנים: **היסק אינדוקטיבי, מהימנות של מקורות מידע ותצפיות, דדוקציה, וזיהוי הנחות חבויות**. המבוא למבחן מספר על קבוצת חוקרים מכדור הארץ שנחתה על כוכב דמיוני בשם ניקומה שהתגלה זה לא כבר. החוקרים הגיעו לניקומה כדי לחפש קבוצת חוקרים אחרת שנחתה על הכוכב שנתיים קודם ועקבותיה אבדו. הנבחנים מתבקשים לדמיין שהם נמנים על קבוצת החוקרים ונוטלים חלק בחיפוש אחר החוקרים האבודים. במבחן 71 פריטים רבי-ברירה. לוח 6.2 מציג מספר דוגמאות לפריטים מתוך המבחן.

לוח 6.2 - דוגמאות לפריטים ממבחן קורנל לחשיבה ביקורתית

(תרגום מתוך Ennis, Millman & Tomko, 1985)

א. דוגמאות 3-5 לקוחות מתת-המבחן העוסק בחשיבה אינדוקטיבית.

אתה וקצין התברואה מגיעים ראשונים לקבוצת בקתות עשויות מתכת. אתם קוראים בקול אבל לא מקבלים כל תשובה. "אולי הם כולם מתים," מציע קצין התברואה. אתה מנסה לגלות אם הוא צודק. להלן העובדות שאתה מגלה.

לגבי כל עובדה, סמן אחת מהאפשרויות הבאות על דף התשובות שלך:

א. עובדה זו **תומכת** בדעתו של קצין התברואה, שכל החוקרים מהמשלחת הראשונה מתים.

ב. עובדה זו נוגדת את דעתו של קצין התברואה, שכל החוקרים במשלחת הראשונה מתים.

ג. לא זה ולא זה: עובדה זו אינה עוזרת לנו להחליט.

3. אתה נכנס לבקתה הראשונה. הכול מכוסה שכבה עבה של אבק.

4. אתה נכנס לבקתה השלישית. אין אבק על תנור הבישול.

5. אתה מוצא פותחן קופסאות ליד תנור הבישול בבקתה השלישית.

לפי מפתח התשובות הנכונות, התשובה לשאלה 3 היא א' (כלומר, עובדה זו תומכת בדעתו של קצין התברואה), התשובה לשאלה 4 היא ב' (כלומר, עובדה זו נוגדת לדעתו של קצין התברואה) והתשובה לשאלה 5 היא ג' (כלומר, לא זה ולא זה: עובדה זו אינה עוזרת לנו להחליט).

ב. דוגמאות 26 ו-28 לקוחות מתת-המבחן העוסק בבדיקת המהימנות של מקורות מידע.

הערב יורד, ואתם מתכוננים ללינת לילה. עם שחר אתם שוב יוצאים לדרך. לאחר הליכה של שעה אתם מגיעים לכפר של בקתות מאבן. הכפר ריק. השמש מאירה באור בהיר וחזק. אתה מנהיג הקבוצה. חברי הקבוצה יוצאים לסייר בשטח ומדווחים לך. בכל פעם יובאו בפניך שני דיווחים. קרא את שניהם ולאחר מכן החלט לאיזה דיווח עליך להאמין יותר.

אם אתה חושב שעליך להאמין יותר לדיווח הראשון, סמן א' על דף התשובות.

אם אתה חושב שעליך להאמין יותר לדיווח השני, סמן ב'.

אם אתה חושב שעליך להאמין במידה שווה לשני הדיווחים סמן ג'.

לפניך דוגמא (ההיגד שאת אמינותו עליך לשפוט מסומן בקו):

26.

א. מכונאי הרכב חוקר את הנחל שזורם ליד הכפר ומדווח: "המים

אינם טובים לשתייה".

ב. קצין התברואה אומר: "עדיין אי-אפשר לקבוע אם המים

טובים לשתייה".

ג. יש להאמין במידה שווה לשני הדיווחים (א' ו-ב').

בתשובה לשאלה 26 עליך לסמן ב' בדף התשובות. קצין התברואה אמור להבין בטיב המים יותר מאשר מכונאי הרכב.

28. סמן את המשפט שנראה לך אמין ביותר

א. המכונאי אומר: "המים נראים צלולים".

ב. קצין התברואה (לאחר שביצע מספר בדיקות) אומר: "המים

בטוחים לשתייה".

ג. יש להאמין במידה שווה לשני הדיווחים (א' ו-ב').

* * *

לפי מפתח התשובות הנכונות, התשובה הנכונה לשאלה 28 היא א', כלומר יש להאמין יותר לדיווח של מכונאי הרכב.

קשיים בהערכה של יכולת חשיבה

לפני שנבחן את חסרונותיהם של מבחנים אלה ככלי מדידה, ראוי לשים לב לקושי הנוגע להיבט החינוכי שלהם. המבחנים שתוארו מורכבים משאלות רב-ברירות, כלומר שאלות סגורות שצריך לסמן בהן תשובה נכונה אחת בלבד מתוך מספר תשובות אפשריות (המונח שאלה רב-ברירותיות הוא תרגום של המונח האנגלי 'multiple choice'. כינוי נפוץ לשאלות אלה בעברית הוא 'שאלות אמריקאיות'). לכן קיימת סתירה מהותית בין נוקשותן של שאלות אלה לבין ערכים ומסרים חינוכיים המודגשים בתוכניות לפיתוח החשיבה: מסר חשוב בהוראת חשיבה הוא שיש חשיבות רבה לתהליך החשיבה עצמו, ולא דווקא לתוצאתו הסופית. אם רוצים שתלמידים ירגישו נוח להעלות בכיתה רעיונות נועזים ומקוריים כפתרון לבעיות, אי-אפשר לעבוד בשיטה שלפיה התלמיד חייב 'לקלוע' לתשובה האחת והיחידה שהמורה מצפה לה, אלא להיפך, יש להמעיט מערך התשובה הסופית ולתת לתלמיד חיזוקים על עצם ביצוע תהליך החשיבה. מכאן יוצא שהתוצר הסופי או התשובה 'הנכונה' הם לעתים קרובות בעלי חשיבות משנית (ראו דיון בפרק 4, עמ' 76). נוסף לכך, ראינו כי אחד האפיונים של חשיבה מסדר גבוה הוא שהיא מסתיימת לעתים קרובות בפתרונות מרובים, שלכל אחד מהם יש יתרונות וחסרונות, ולא בפתרון יחיד ובודד. אם אכן מתקיימת הוראה בגישה המדגישה את פיתוח החשיבה, יהיה זה מוזר להעריך אותה באמצעות מבחן שהמסר החינוכי שלו מנוגד לרוח ההוראה.

בעיה זו קשורה גם לחיסרון של המבחן כאמצעי מדידה, משום שנבחנים עשויים לסמן תשובה שונה מהתשובה המסומנת כנכונה על פי המפתח המלווה את המבחן, גם כאשר הם מפעילים תהליכי חשיבה 'נכונים'. לעתים נגרמת תופעה זו על ידי ידע מוקדם או מערכת של תיאוריות פרטיות שהתלמידים מביאים אתם למבחן. לפעמים, דווקא תהליכי חשיבה מתוחכמים ביותר עשויים להוביל את הנבחן לתשובה שונה מזו שעליה חשבו מחברי המבחן.

כדי להדגים נקודה זו, נחזור לפריט מס' 6 במבחן ווטסון-גלור המתאר את מר בראון והפאב שלו (ראו לוח 6.1 לעיל). נוסף לתשובה הרשומה במפתח התשובות כנכונה, אפשר לחשוב על הסברים אחרים לתופעה המתוארת. נוריס ואניס, (Norris & Ennis, 1989, עמ' 58-59), למשל, מציעים כי ייתכן שמר בראון שרוי באבל בגלל אסון שקרה לאדם קרוב. בעקבות האבל הוא החל לשתות לשוכרה ואיבד את השליטה על המתרחש בפאב שלו. או אולי מר בראון העביר את השליטה בפאב לבנו ונתן לו יד חופשית כדי לכפר על כך שהזניח את הבן בילדותו. במקרים אלה (או במקרים רבים אחרים שאפשר לדמיין), אומרים נוריס ואניס, לא יהיה זה נכון לומר כי למר בראון היה מה להרוויח מכך שהפאב נשאר פתוח לאחר השעה אחת לפנות בוקר. לעומתם, אפשר לטעון כי הכול תלוי בפירוש שנותנים למושג 'להרוויח'. באפשרות הראשונה שהעלו נוריס ואניס הרוויח מר בראון את השכחה שהעניקה לו השתייה, ובאפשרות השנייה את אהבתו של בנו. מכל מקום, נבחן עשוי להעלות בראשו אפשרויות כגון אלה ואז לבחור בתשובה "אין מספיק מידע כדי להחליט אם המסקנה אמיתית".

נבחן דוגמא נוספת, המופיעה בפריט מס' 28 במבחן קורנל. המפתח למבחן קובע כי המידע "המים נראים צלולים", המבוסס על תצפיתו של המכונאי, מהימן יותר מן המידע המבוסס על בדיקות שערך קצין התברואה ולפיהן "המים בטוחים לשתייה". אולם קביעה זו תלויה במידע קודם ובאמונות הנבחן לגבי טיבן של הבדיקות שערך קצין התברואה. אם הנבחן מאמין שבדיקות מים הנערכות בתנאי שדה ובזמן קצר אינן בעלות דרגת אמינות גבוהה, הוא יבחר בוודאי בתשובה המתאימה למפתח. אולם אם הנבחן יודע על קיומה של בדיקה אמינה ומהירה לטיב המים, סביר שיענה תשובה אחרת. ברוח

זו אפשר לחשוב על הסתייגויות דומות לגבי הפריטים האחרים שהובאו כדוגמאות.

ראינו שמבחן כללי לפעמים אינו מתאים להערכת לימוד ספציפי, הן מבחינת התוכן והן מבחינת יכולות חשיבה. יחד עם זאת, אם משתמשים במבחן שחובר במיוחד לצורך הערכת לימוד ספציפי, מתעוררת בעיה הפוכה הנובעת משינון החומר על ידי תלמידים. כדי להסביר את הקושי הזה, הבה נניח תוכנית המלמדת מספר מיומנויות חשיבה באמצעות שורה של תכנים, ומבחן המורכב משאלות העוסקות באותם תכנים עצמם ובאותן מיומנויות חשיבה. במקרה כזה, גם אם שאלות המבחן מנוסחות כשאלות הדורשות חשיבה ברמה גבוהה, אין הנבחן חייב להפעיל חשיבה ברמה גבוהה כדי לענות עליהן: כל שעל הנבחן לעשות הוא לשנן את הנאמר בכיתה ולהיזכר בעת המבחן בדיון הכיתתי שנערך בנושא. במלים אחרות, במקום שאלות הדורשות רמה גבוהה של חשיבה, לפנינו שאלות הדורשת שלילת מידע מהזיכרון.

במקרים אחרים, ייתכנו מצבים שבהם תוכן השאלה השתנה אמנם, אך דפוסה זהה או דומה לדפוסים של שאלות שנשאלו בכיתה. במצב עניינים כזה, ייתכן שיש בדפוס השאלה רמזים המכוונים את התלמיד לתשובה הנכונה. כך תלמיד יוכל לענות נכון על שאלה הדורשת חשיבה בעזרת דליית הפתרון מזיכרונו, מבלי שיצטרך לבצע בפועל את החשיבה הדרושה לפתרון הבעיה. לדוגמא, אם כאשר עוסקים בבידוד משתנים (ראו פרק 7), עושים זאת תמיד באמצעות דפוס שאלה כגון: "מהו תפקידה של מבחנה X?", התלמידים יזהו את הדפוס גם בבחינה ויענו: "מבחנה X היא מבחנת הבקרה". תשובה נכונה כזו אינה דורשת הבנה של מושג הבקרה או הפעלה של מיומנות החשיבה הדרושה כדי לבדוד משתנים. שתי ההסתייגויות האחרונות תופסות גם לגבי שאלות פתוחות, ונראה כי הדרך היחידה להשתחרר מהן היא שימוש בשיטות ההערכה האלטרנטיביות הנידונות לקראת סוף הפרק.

הצעות נוספות להערכת תוכניות

לאור הדברים שהובאו בסעיפים הקודמים, נראה שהערכה של תוכניות חינוכיות באמצעות שימוש במבחני חשיבה נפוצים (כגון ווטסון-גלזר או קורנל) הכוללים שאלות רבות-ברירה היא בעייתית. אם כך, באילו דרכים חלופיות אפשר להיעזר?

פול (Paul, 1992) התבקש על ידי מחלקת החינוך המרכזית של ארצות הברית לגבש נייר עמדה שיתאר מודל להערכת רמות חשיבה גבוהות בקנה מידה לאומי. פול מונה שלוש אסטרטגיות שלדעתו צריכות להיות כלולות בהערכה כזו:

א. שימוש בפריטים רבי-ברירה. לדעת פול יש מקום לשימוש בפריטים רבי-ברירה, אולם הוא טוען כי שימוש לגיטימי בפריטים כאלה הוא מוגבל, משום שהם נוטים לכיוון מיומנויות חשיבה שאינן מורכבות.

ב. שימוש בפריטים רבי-דירוג (multiple rating items). בפריטים רבי-דירוג נשאלות שאלות וניתנת רשימה ארוכה של תשובות אפשריות. התלמיד יכול לבחור **מספר בלתי-מוגבל של תשובות**. נוסף לכך, מאפשרת השאלה לתלמיד לדרג את התשובה: יש סולם (בדרך כלל מ-1 עד 5), והתלמיד צריך לקבוע עד כמה התשובה שבחר נכונה, סבירה או הגיונית בעיניו. לדוגמא:

* "דרג את הפריטים הבאים על סולם של 1-5 לפי מידת הבנתך אותם כמביעים את ההנחות החבויות של המחבר."

* "לפניך רשימה של ניסוחים המתארים את מטרות המחבר בקטע שקראת. דרג אותם לאור הקטע שקראת על סולם של 1-5 לפי רמת הסבירות שלהם."

פול טוען כי פריטים רבי-דירוג מאפשרים הערכה אמינה, עדינה ומפורטת יותר מאשר פריטים רבי-ברירה. כאשר דרך חישוב הציון במבחנים כאלה היא מתוחכמת מספיק, הרי שאסטרטגיות שעוזרות להצליח במבחנים המכילים פריטים רבי-ברירה (כגון ניחוש או שימוש באלימינציה) לא תהיינה יעילות. נוסף לכך, אפשר להשתמש בפריטים

כאלה גם כדי להעריך אספקטים ריגושיים המתייחסים לנטיות התנהגותיות ולעמדות הנוגעות לחשיבה.

ג. שימוש בשאלות פתוחות. פול טוען כי לא ניתן להעריך חשיבה במלוא היקפה מבלי לדרוש מהתלמיד, כחלק מהבחינה, ליצור תשובות מקוריות באמצעות התבטאות עצמית בכתב. לכן הערכה של תוכניות לפיתוח חשיבה חייבת לדעתו לכלול, נוסף לשאלות בחירה, גם שאלות פתוחות (שאלות נושא), שבהן מתבקש התלמיד ליצור את התשובה בעצמו. רק שאלות כאלה עשויות לשקף חשיבה בנושאים 'אמיתיים', שהבעיות המוצגות בהם דומות לבעיות המתעוררות במציאות. יכולות כגון שקילת אלטרנטיבות, מציאת איזון בין דרישות סותרות וקבלת החלטות מנומקות, אינן יכולות להיבדק במבחנים שבהם על התלמיד לבחור תשובה. פול מציע לבדוק יכולות כאלה באמצעות **שאלות נושא**. לוח 6.3 מציג פריט אחד מתוך בחינה לדוגמא ואת ההוראות הנלוות אליו:

לוח 6.3 - דוגמא לשאלת נושא מתוך מבחן לחשיבה ביקורתית

הוראות לבחינה:

מבחן זה נועד לבדוק את יכולתך לחשיבה ביקורתית, פתרון בעיות ותקשורת. הבוחן יעריך את תשובתך לפי ההיבטים הבאים:

1. האם הכותב מציג שאלה ברורה ובהירה, שאינה מכוונת מראש לאף אחת מהתשובות?
2. האם הכותב מצטט ראיות רלוונטיות, ניסויים ו/או מידע הכרחיים לדיון בנושא?
3. האם הכותב מבהיר מושגי מפתח כאשר הדבר דרוש?
4. האם הכותב מגלה רגישות להנחות מובלעות?
5. האם הכותב מפתח קו מחשבה מבוסס, המסביר היטב איך הגיע למסקנה שהגיע?
6. האם הכותב מגלה רגישות לנקודות ראיות אחרות או לקווי מחשבה חלופיים? האם הוא מתייחס ומגיב לזווית אפשרי עם נקודות מבט אחרות?

7. האם הכותב מגלה רגישות להשלכות ו/או לתוצאות של העמדה שנקט?

פריט מס' 3 - מוסר

הסוציולוג אירוויין גופמן הצביע על כך שכל הקבוצות החברתיות (כולל קבוצות שהמושג בין חבריהן הוא שכולם נמנים על בעלי מקצוע מסוים) מפתחות עמדות מגוונות כלפי חברי הקבוצה שלהם. התופעה מתקיימת אפילו כאשר מעשיהם של אחדים מחברי הקבוצה נתפסים כפגומים מבחינה מוסרית. דון בטיעונים בעד ונגד חשיפת מעשים בלתי-מוסריים של אנשים הקרובים אליך בקשר אישי או מקצועי. פתח עמדה בנושא, אשר תוכל לשמש כמדריך לכל אדם שייקלע למצב דומה.

* * *

שאלות נושא אכן בודקות היבטים של חשיבה שלא ניתן לבדוק באמצעות שאלות בחירה. יחד עם זאת, מדגימות שאלות אלה עד כמה בודקים מבחנים מן הסוג הזה יכולת **תקשורת** ולא רק יכולת חשיבה, שכן התשובה דורשת ללא ספק יכולת גבוהה של הבנת שפה והתבטאות מילולית. חשוב גם לזכור כי ליתרון של פריט מעין זה יש מחיר לא מבוטל מבחינת תהליך הבדיקה של המבחן. תהליך הבדיקה של שאלות רבות-ברירה הוא קל, מהיר וזול. לעתים קרובות נעזרים באמצעים טכנולוגיים מתקדמים (כגון קורא אופטי) המאפשרים לבדוק אלפי מבחנים בזמן קצר ביותר. מהימנות הבדיקה גבוהה משום שהיא אינה תלויה בשיקול דעתו של הבודק. כל בודק (אדם או מכונה), שיעבוד לפי מפתח הבדיקה המצורף למבחן (כלי לטעות), יעריך את המבחן בדיוק באותו אופן. אולם כאשר עוברים לשאלות נושא, יתרונות אלה נעלמים. שאלות פתוחות חייבות להיבדק על ידי מעריכים בשר ודם. לכל מעריך מערכת שיקולים סובייקטיבית משלו. כדי לצמצם את השפעת הסובייקטיביות של המעריך על הערכת המבחן, יש לצרף לכל שאלת נושא מפתח איכותי שיגדיר קריטריונים מדויקים וברורים לדירוג התשובות. כאשר המטרה היא להגיע להערכה אובייקטיבית, חייבים להגיע למפתח שיהיה ברור דיו כדי שמעריכים שונים הבודקים מבחנים זהים יעריכו אותם באותו אופן.

במלים אחרות, כשאנו עוסקים בהערכה של שאלות פתוחות, חשוב לשים את הדגש על **מהימנות גבוהה בין מעריכים שונים** (ראו דיון נוסף במושג זה בעמ' 121-122 להלן).

בניית מפתח כזה היא עבודה קשה, הדורשת זמן ומיומנות. גם תהליך הניתוח והבדיקה של כל שאלה נמשך זמן רב ודורש מעריך מומחה ומיומן. לכן, תהליך הבדיקה של שאלות פתוחות, כדוגמת שאלות הנושא שפול מציע, הוא תהליך יקר.

לסיכום, אפשר לראות כי להערכה שאינה מסתפקת עוד בשאלות סגורות יש יתרונות גדולים מצד אחד ומחיר כבד מצד שני. מצד אחד, יש בהערכה כזאת פוטנציאל למדידה תקפה יותר מאשר במבחנים רבי-בחירה; מצד אחר, שיטות הערכה פתוחות קשות יותר לביצוע ויקרות יותר. שני היבטים אלה של ההערכה יבלטו ביתר שאת בשיטות ההערכה האלטרנטיביות שיתוארו בסעיפים הבאים.

שיטות הערכה אלטרנטיביות

הקשיים בהערכה של מיומנויות חשיבה גבוהות שתוארו בסעיפים הקודמים הביאו אנשי חינוך רבים להכרה כי הכנסת חינוך לחשיבה ברמה גבוהה למערכת החינוך, מחייבת גם שינוי בדרכי הערכה (בנוסף לשיטות ההערכה שהוצעו על ידי פול). משאבים רבים מוקדשים כיום ברחבי העולם לחיפוש אחר גישות חלופיות להערכה. לשיטות ההערכה החדישות יש מספר שמות: הערכה חלופית (alternative assessment); הערכה אותנטית (authentic assessment); או הערכת ביצועים (performance assessment). יתרונו של השם 'הערכת ביצועים' הוא בכך שהוא משדר לא רק את עצם החידוש בגישה אלא גם את מהותה, כלומר הערכת אופן הביצוע של משימות שונות.

שיטות הערכה מסורתיות מתאפיינות בקיום מבחנים במועדים קצובים, בדרך כלל לאחר סיום תהליך הלמידה. מבחנים אלה דוגמים את רמת ההישגים של התלמיד באמצעות סדרה של שאלות שחוברה במיוחד לצורך ההערכה. שיטות ההערכה האלטרנטיביות עשויות אמנם לכלול גם מבחנים כאלה, אך הן מתבססות על יצירת דרכים

להעריך לא רק את התוצר הסופי של הלמידה (כפי שמקובל במבחנים מסכמים) אלא את **תהליך הלמידה** עצמו. הרעיון המרכזי של השיטה מתמקד ביצירת 'חלונות' שיאפשרו 'להציץ' על תהליך הלמידה בזמנים שונים ולהעריך את התקדמותו של התלמיד במשך הזמן. במקום להסתפק במבחנים שנבנו במיוחד לצורכי הערכה, משתמשים בשיטות החדשות **בתוצרים האוטנטיים** של תהליך הלמידה.

מגוון רחב מאוד של תוצרי תהליך הלמידה עשוי לשמש להערכה. בחירת התוצר המתאים להערכה צריכה להיגזר ממטרת תהליך ההוראה. בקורס שמטרתו לימוד כתיבה, התוצר יהיה קטעים שנכתבו על ידי התלמידים. בקורס שמטרתו רכישת ידע בשפה זרה, התוצר יהיה שיחה בשפה זו. גם בקורסים שעיקרם פיתוח החשיבה יהיו התוצרים פירות של תהליך העבודה של התלמידים: בקורס שמטרתו היא פיתוח מיומנויות חקר מדעיות יהיה התוצר ביצוע של חקירה על ידי התלמיד. בקורס שעיקרו פתרון בעיות יהיה התוצר קובץ של מהלכי פתרונות. בקורס שמטרתו הקניית מיומנויות ללימוד עצמאי יהיה התוצר עבודת סיכום או דיווח על נושא. בקורס שמטרתו הקניית מיומנויות חשיבה מוגדרות עשוי התוצר להיות מהלך פתרון של משימה הדורשת שימוש במיומנויות אלה. התוצרים עשויים להיאסף בדרכים שונות: הקלטה של דיון המתקיים בין תלמידים בזמנים שונים במהלך הלמידה, למשל, עשויה לשמש כחומר גלם לניתוח של מיומנויות החשיבה שבהם השתמשו, סוגי הנימוקים שהביאו כדי לתמוך בדעתם ושינויים בדפוסי השיחה ורמת ההקשבה ההדדית שלהם (Brown et al., 1993). תוצרים נבחרים של תהליך הלמידה כגון דפי עבודה, בעיות פתורות, דוחות מעבדה, חוות דעת מנומקות, מפות מושגים (ברנהולץ ואמיר, 1985) או ציורים (Rosato et al., in press) עשויים להוות חומר גלם לניתוח למטרות הערכה. מקורות נוספים להערכת ביצועים עשויים להתקבל מאיסוף דוגמאות מעבודות כתובות, הערכה בעקבות שיחה בעל פה או ריאיון, תצפית בתלמיד המבצע ניסוי מדעי או פותר בעיה, תערוכות שתלמידים מציגים בהן את עבודותיהם או הרצאות של תלמידים בפני הכיתה.

אחת השיטות הנפוצות לאיסוף מידע עבור הערכת הביצועים היא שיטת תיק העבודות (portfolio). לפי שיטה זו בוחרים התלמידים

חלק מן התוצרים של תהליך הלמידה שלהם ושומרים אותם בתיק מיוחד. המושג נגזר מתחום האמנות, כיוון שבאמנות מקובל זה שנים רבות שתלמידים אוצרים בתיק העבודות דוגמאות נבחרות של ציורים או של עבודות אחרות. בדומה לכך, תלמידים בשיעורים לפיתוח חשיבה ישמרו בתיק דוגמאות ממשימות שביצעו. כמעט כל מטלה לימודית בכל מקצוע עשויה להתאים לתיק העבודות. לדוגמא, בשיעורי מדע יכניסו התלמידים לתיק הסברים לתופעות, מסקנות מניסויים או תוכנית שהתוו לניסוי חדש. בשיעור מתמטיקה יאספו התלמידים בתיק את מהלך הפתרון של בעיות שהיו צריכים להתמודד עמן בעצמם. בשיעור היסטוריה יכניסו לתיק ניתוח של מקורות היסטוריים, השוואה בין שתי תקופות או ניתוח שערכו לגבי סיבות אפשריות למלחמה, והשלכותיה. בכל נושא (כולל בקורסים לפיתוח החשיבה בגישה הכללית) יבדוק המורה (או מעריך חיצוני), לפי קריטריונים מוגדרים, מדגם מן התוצרים שהוכנסו לתיק העבודות.

כאן מן הראוי להיכנס לסוגיית **הקריטריונים להערכה**, שהיא סוגיה מרכזית בכל דרכי ההערכה האלטרנטיבית. המשותף לכל דרכי הערכה אלה הוא שימוש בכלי ניתוח איכותיים, הכוללים פיתוח של אמות מידה מתאימות (קריטריונים) שבעזרתן מנתחים את הנתונים. למרות שהניתוח הוא איכותי (כלומר, יש להפעיל שיפוט באיזו מידה התוצר שאותו מנתחים עומד בקריטריון), התוצר הסופי של הניתוח עשוי (גם אם לא בהכרח) להתבטא בציון מספרי. בגלל אופיו האיכותי של הניתוח לפי קריטריונים, יש המכנים את ההערכה האלטרנטיבית גם בשם 'הערכה איכותית'. דיון מסוים בנושא המורכב של ניתוח איכותי לפי קריטריונים נערך קודם לכן כאשר תיארו את הקושי הטמון בהערכת שאלות נושא פתוחות (ראו עמ' 117-118). אולם הגיון הגדול של החומרים המשמשים בהערכה האלטרנטיבית גורם לכך שסוגיית הקריטריונים כאן תהיה מורכבת עוד יותר. הקריטריונים צריכים כמובן להיקבע לפי מטרות החשיבה שהמטלה דורשת, אותן מטרות שהנחו את מחברי הפעילות הלימודית.

חיבור הקריטריונים להערכה איכותית הוא תהליך ארוך הכולל מספר שלבים. בשלב הראשון מגבש צוות של אנשי מקצוע את הקריטריונים המתבקשים לאור מטרות הפעילות הלימודית. בשלב השני מנסים

חברי הצוות לנתח מדגם של עבודות תלמידים לפי הקריטריונים הראשוניים שגובשו. בשלב זה מתעוררים תמיד קשיים משום שמתברר כי רק חלק מהקריטריונים מתאימים לעבודות התלמידים שאותן רוצים לנתח: מצד אחד, נמצא כי קריטריונים מסוימים, שהוגדרו במהלך הניתוח התיאורטי שערכו חברי הצוות, אינם ניתנים ליישום על עבודות התלמידים. מקרים כאלה מתרחשים כאשר במדגם תשובות התלמידים לא מופיעות דוגמאות המתאימות לקריטריון שנקבע, או כאשר מסתבר כי קריטריון מסוים נוסח באופן עמום שאינו מאפשר למעריך להחליט בברור אם הוא מתאים לדוגמא ספציפית של עבודת תלמיד. מצד אחר, מתגלים מקרים של תשובות מפתיעות בכיוונים שחברי הצוות לא חשבו עליהם מראש ולכן לא הכינו עבורם קטגוריות מתאימות. בשלב השלישי של העבודה מתקנים חברי הצוות את הקריטריונים לאור הבעיות שעלו בעקבות הניתוח של מדגם התשובות. בשלב הרביעי ינסו להפעיל את הקריטריונים המתוקנים על מדגם תשובות נוסף. צפוי שהפעם תהיה התאמה גדולה יותר בין הקריטריונים לבין התשובות, אך בדרך כלל צריך לחזור על התהליך מספר פעמים נוספות עד שמגיעים לקריטריונים הסופיים. גורם נוסף שיש להתחשב בו הוא **מידת ההסכמה בין מעריכים שונים**. מספר מעריכים (שניים לפחות) מסווגים כל אחד בנפרד מדגם נוסף של תשובות תלמידים לפי רשימת הקריטריונים ואחר כך מחשבים את אחוז ההסכמה ביניהם. השאיפה היא להגיע להסכמה של כ-90%. בדרך כלל מתקבלים בהתחלה אחוזי הסכמה נמוכים יותר, המשקפים נקודות שבהן הקריטריונים אינם חד-משמעיים, ולכן ניתנים לפירושים שונים. הדרך לשפר את ההסכמה בין המערכים היא לקיים דיון בכל מקרה ומקרה של אי-הסכמה ולחדד את הקריטריון הרלוונטי (בין אם על ידי שיפור הניסוח ובין אם על ידי הוספת קריטריון חדש). גם תהליך זה דורש בדרך כלל מספר סיבובים לפני שמגיעים לתוצר משביע רצון. לפעמים מגיעים למסקנה שהערפול נובע מחוסר בהירות במטלה ועל כן חוזרים ומנסחים אותה מחדש. תהליך חיבור הקריטריונים האיכותיים הוא, אם כך, תהליך ארוך, הדורש השקעה רבה במשאבי אנוש מקצועיים. נוסף לכך, בהמשך יהיה צורך להשקיע בהכשרת מעריכים שיבדקו את עבודות התלמידים תוך שימוש בקריטריונים. מחקרים מעידים כי כאשר

מקפידים על כל השלבים שתוארו כאן, אפשר להגיע לאחוזי הסכמה גבוהים מאוד בין מעריכים (למעלה מ-90%) גם במחקרי הערכה רחבים המקיפים תלמידים ומעריכים רבים (למשל, בבחינות המקיפות שכבת גיל שלמה במדינות מסוימות בארצות הברית, או בבחינות בגרות מסוימות בישראל, כגון בחינות הבגרות במעבדה לביולוגיה). לפיכך אין זה נכון לזהות הערכה איכותית עם הערכה סובייקטיבית ובלתי-מהימנה.

להדגמת התוצר של התהליך שתואר בפסקה הקודמת נתאר שתי דוגמאות לקריטריונים איכותיים להערכה. הדוגמא הראשונה מתייחסת למחקר איכותי של חשיבה מדעית. במחקר זה מנתחים היסקים של תלמידים. לגבי כל היסק שואלים אם הוא תקף או לא. הקריטריון להיסק תקף הוא שעל ההיסק להיות מבוסס על השוואה בין שני ניסויים לפחות, ועל בידוד משתנים (בפרק 7 מתוארים קריטריונים נוספים לסיווג מפורט יותר של ההיסקים הבלתי-תקפים).

דוגמא שנייה לקריטריונים איכותיים מוצגת בלוח 6.4. דוגמא זו לקוחה מפרויקט המשלב חשיבה בהוראת מתמטיקה (מתוך Mitchell, 1992).

לוח 6.4 - דוגמא לכלי ניתוח איכותי לשאלה פתוחה במתמטיקה הדורשת פתרון בעיה (שהתשובה עליה אינה הפעלת אלגוריתם)

התלמיד מראה יכולת גבוהה

תשובה לדוגמא (הדירוג = 6 נקודות)

ניתנת תשובה שלמה הכוללת הסבר בהיר, קוהרנטי, חד-משמעי ואלגנטי; כולל תרשים בהיר ופשוט; כולל קומוניקציה יעילה עם קהל היעד; מגלה הבנה של הרעיונות והתהליכים המתמטיים הקשורים בבעיה הפתוחה; זיהוי כל האלמנטים החשובים של הבעיה; עשוי לכלול דוגמאות ודוגמאות נגד; מציג טיעונים חזקים התומכים בפתרון.

תשובה טובה (הדירוג = 5 נקודות)

ניתנת תשובה שלמה למדי הכוללת הסבר בעל דרגה סבירה של בהירות; עשויה לכלול תרשים מתאים; כולל קומוניקציה יעילה עם קהל היעד; מגלה הבנה של הרעיונות והתהליכים המתמטיים הקשורים בבעיה הפתוחה; זיהוי האלמנטים העיקריים של הבעיה; מציג טיעונים סבירים התומכים בפתרון.

התלמיד מגלה יכולת משביעת רצון

תשובה משביעת רצון עם פגמים קלים (הדירוג = 4 נקודות)

ניתנת תשובה משביעת רצון, אך ההסבר עשוי להיות מעורפל; הטיעונים עשויים להיות חלקיים; התרשימים עשויים להיות לא מתאימים או לא בהירים; התשובה מעידה על הבנת העקרונות המתמטיים הרלוונטיים ועל שימוש יעיל ברעיונות מתמטיים.

פגמים רציניים אבל תשובה כמעט משביעת רצון (הדירוג = 3 נקודות)

מופיעה התחלה של פתרון, אך הוא אינו מושלם, או שהוא מתעלם מחלקים חשובים של הבעיה; עשוי להיכשל בהבנה של הרעיונות והתהליכים המתמטיים; עשוי לעשות טעויות עקרוניות בחישוב; עשוי להימנע משימוש במושגים מתמטיים חשובים, או להשתמש בהם באופן שגוי; התשובה עשויה לשקף אסטרטגיה בלתי-מתאימה לפתרון הבעיה.

תשובה לא מספקת

מתחיל, אך לא מצליח להשלים את הפתרון (הדירוג = 2 נקודות)

ההסבר של התלמיד אינו מובן; התרשים עשוי להיות בלתי-מובן; התשובה אינה מראה הבנה של הבעיה; עשוי לעשות טעויות עקרוניות בחישוב.

העדר יכולת להתחיל לפתור את הבעיה (הדירוג = 1 נקודה)

מלים שאינן משקפות את הבעיה; תרשימים שאינם מייצגים נכונה את הבעיה; מעתיק חלק מן הבעיה מבלי שינסה לפתור אותה; אינו מצליח לרמוז איזה מידע מתאים לפתרון הבעיה.

אין ניסיון לפתור את הבעיה (הדירוג = 0)

* * *

לשיטות ההערכה האלטרנטיביות יש יתרון נוסף מלבד מה שנאמר עד כה: הערכה שוטפת של התקדמות התלמיד **במשך הלמידה**, ולא רק בסופה, מדגישה במיוחד את תרומתה הפוטנציאלית של ההערכה לעיצוב תהליכי הוראה ולמידה (נוסף לתפקיד ההערכה במתן דין וחשבון בדיעבד על איכות הלמידה או ההוראה). המורה מנתח את התוצרים המתקבלים מתהליך הלמידה לפי קריטריונים קבועים מראש, וכך מקבל מידע שוטף על מצבו של התלמיד. המידע הזה עומד גם לרשות התלמיד שיכול לגלות את מידת התקדמותו בשלבים מוקדמים של הלמידה, ולתכנן את המשך עבודתו בהתאם. כמו כן, יכולים התלמידים עצמם ללמוד את הקריטריונים להערכת עבודתם. הקריטריונים הללו גלויים לעין כל: בכיתות מסוימות אף תולים המורים את רשימת הקריטריונים להערכה על פלקט גדול בקדמת הכיתה. תלמידים רואים את הקריטריונים, דנים בהם, ושואלים את עצמם בהזדמנויות שונות באיזו מידה ממלאת עבודתם אחר הקריטריונים הללו. באופן כזה יש לקריטריונים להערכה תפקיד חשוב נוסף שכן הם **מכוונים את תהליך הלמידה המתרחש בכיתה**. היות שהקריטריונים קשורים קשר הדוק למטרות החשיבה של התוכנית הלימודית, תורם הדיון בקריטריונים להפנמת מטרות אלה. למעשה, הדיון בקריטריונים עשוי להוות את אחת הדרכים לדיון מטא-קוגניטיבי בעקרונות ובכללים המארגנים את החשיבה. בנוסף לכך, תלמיד הרוכש את היכולת להעריך בעצמו את עבודתו, מתקדם לקראת היותו לומד עצמאי, המקבל אחריות אישית על תהליך הלמידה שלו.

שיטות ההערכה האלטרנטיביות מתאימות להערכת הישגים בפרויקטים שמטרתם פיתוח מיומנויות חשיבה גבוהות (ואולי גם בתהליכי למידה באשר הם) הרבה יותר מאשר המבחנים שתוארו בסעיפים הקודמים. ניתוח של תיק עבודות פותר את בעיית התקיפות שתוארה קודם לגבי מבחנים הבודקים רמות חשיבה: כאן, אין צורך לשאול אם פריטי המבחן אכן משקפים נאמנה את המיומנויות או את הנושאים שנלמדו, שכן מעריכים ישירות מדגם מן התוצרים של תהליך

הלמידה עצמו. אולם, יחד עם זאת, חשוב לזכור כי התוקף הגבוה דורש מחיר: ניתוח תיק עבודות דורש משאבים רבים, במיוחד אם מסרבים להתפשר על דרגת המהימנות של הבדיקה. רות מיטשל (Mitchell, 1992) דנה בנקודה זו באומרה כי הקהילייה הפסיכומטרית היתה תחילה עוינת כלפי הערכת ביצועים, אולם עד מהרה החלה לראות בהערכת ביצועים אתגר חדש. במונחים המקצועיים של אנשי הפסיכומטריה, יש להערכת הביצועים תוקף גבוה (משום שהיא מודדת ישירות את מה שהתלמידים אמורים לדעת ולהיות מסוגלים לעשות), אך מהימנות נמוכה (משום שהציון אינו ניתן על ידי מכוונות אלא על ידי בני-אדם). במונחים פסיכומטריים, המהימנות הנמוכה גורמת לירידה באובייקטיביות של הערכת הביצועים. קיים מתח (trade off) בין 'אותנטיות', לבין מהימנות ואובייקטיביות. ככל שהתוקף עולה, כן יורדת המהימנות והאובייקטיביות.

אולם, ממשיכה מיטשל להסביר, הקהילייה הפסיכומטרית מתחילה כעת לרתום את הידע הסטטיסטי העצום שלה לפתרון בעיות הנוצרות עקב הצורך להשיג בו-זמנית רמה גבוהה של אותנטיות ושל מהימנות. מממצאיהם של גופים אחדים (כגון ETS - Educational Testing Service) עולה כי ניתן להגיע לדרגות גבוהות של מהימנות בשיטות של הערכת ביצועים, אולם אין ספק שנחוצה לשם כך השקעה של זמן ומומחיות (Mitchell, 1992).

לסיכום, נראה כי לשיטות ההערכה האלטרנטיביות יש יתרונות חשובים ביחס לשיטות ההערכה המסורתיות. יתרונות אלה תופסים לגבי חינוך באופן כללי, אך יש להם חשיבות מיוחדת בנוגע להערכה של מיומנויות חשיבה גבוהות, היות שבתחום זה נתקלת ההערכה הקונוונציונלית בקשיים מיוחדים. הערכה משפיעה על ההוראה. פער בין מטרות הוראה למטרות הערכה מביא לאי-השגת מטרות ההוראה, ולהיפך: הערכה מתאימה עשויה לזרז הכנסת שינוי במטרות ובדרכי ההוראה. היבטים רבים של דרכי הערכה אלטרנטיביות עולים בקנה אחד עם הוראה לקראת חשיבה. יחד עם זאת, העלות הגבוהה של פיתוח השיטות להערכה האלטרנטיבית ושל הפעלתן השוטפת מהווה מגבלה חמורה בדרך לאימוץ שיטות אלה על ידי מערכת החינוך. קשה

לנבא מה צופן העתיד בתחום זה. בארצות הברית, יותר ויותר מדינות (למשל ורמונט וקונטיקט) עוברות כיום להערכת ביצועים. באמצעות שיטה זו מעריכים הישגים במקצועות שונים, כאשר היכולת להשתמש במיומנויות חשיבה גבוהות מהווה אחת ממטרות ההוראה. ואולם במדינת קליפורניה, שהיתה מחלוצות הערכת הביצועים, הקפיאו לאחרונה את המשך הפיתוח של הנושא, בעיקר עקב העלות הגבוהה. ייתכן שקושי זה יוביל בעתיד לשילוב בין מגוון דרכי הערכה: הערכה קונוונציונלית, שהיא זולה יחסית, אך יש לה מגבלות חמורות, בצד הערכה אלטרנטיבית שתתרום לניווט ההוראה לכיוון של מיומנויות חשיבה גבוהות. כדי להראות ששילוב כזה אפשרי ומעשי, נסיים בדוגמא מבחינת בגרות בישראל המיישמת אותו הלכה למעשה.

בחינת הבגרות הישראלית בביוולוגיה בהיקף של חמש יחידות לימוד מורכבת מכמה חלקים (Tamir, 1985): א. בחינה בכתב, הכוללת שאלות רב-ברירות, שאלות פתוחות, וניתוח של מאמר מדעי (unseen); ב. בחינת מעבדה, המהווה אחת הדוגמאות הראשונות בעולם להערכת ביצועים בקנה מידה ארצי. בבחינת המעבדה מתמודדים התלמידים עם חקירה אמפירית שלא הכירו קודם. הבחינה בודקת גם מיומנויות ביצוע טכניות וגם מיומנויות חשיבה גבוהות האופייניות לתהליך החקר המדעי (הגדרת בעיה, ניסוח השערה, תכנון ניסוי, הסקת מסקנות); ג. עבודת ביוטופ, שהיא עבודת מחקר עצמאית שעל התלמיד לבצע בעצמו בנושא שהוא בוחר; ד. בחינה בהגדרת צמחים, שבה נבדקת יכולתו של התלמיד להגדיר של צמח לפי כללים קבועים. הציון הסופי בביוולוגיה הוא ציון משוקלל של כל חלקי הבחינה. המבנה המיוחד של בחינת הבגרות בביוולוגיה משפיע באופן ניכר על ההוראה. הוראה לפי שיטת 'החקר המדעי' חדרה להוראת ביוולוגיה בישראל יותר מאשר למקצועות המדעיים האחרים (פיסיקה וכימיה), ואף יותר מאשר להוראת הביוולוגיה בארצות הברית (משם הובאה השיטה לארץ בשנות השבעים). אחת הסיבות המרכזיות להצלחת השיטה נעוצה בכך, שנוסף לתרגום חומרי הלמידה המתאימים לעברית ולהכשרת מורים, הוכנסה הערכה של מיומנויות חקר מדעיות לבחינות הבגרות. עובדה זו המריצה את המורים ללמד בגישת החקר. הערכת מיומנויות החקר בבחינת הבגרות התאפשרה

בעיקר הודות לפיתוח בחינת המעבדה המיוחדת. אף שהבחינה פותחה הרבה לפני הגל הנוכחי של 'הערכות אלטרנטיביות', היא מהווה אבטיפוס להערכת ביצועים. הדגם הישראלי שימש לאחרונה כדגם לחיקוי במדינות אחדות, ששואפות לשלב הערכת ביצועים במערכת ההערכה הארצית שלהם.