

**מבחנים ממוחשבים -  
סיכויים, סיכונים וסייגים**

יואב כהן

**ת מ צ י ת**

**1. מ ב ו א**

**2. סיווג מבחנים ממוחשבים**

- 2.1 מבחנים קשיחים
- 2.2 מבחנים גמישים בעלי הסתעפות קבועה
- 2.3 מבחנים גמישים
- 2.4 תורת התגובה לפריט-IRT

**3. יתרונות וחסרונות**

- 3.1 בתחום המנהלי והטכני
- 3.2 בתחום הפסיכומטרי
  - 3.2.1 צורת העקומה האופינית של פריט
  - 3.2.2 מימדיות המבחנים
  - 3.2.3 אומדן פרמטרים
  - 3.2.4 יעילות פסיכומטרית
  - 3.2.5 תקפות
  - 3.2.6 תהליך העברת המבחן
  - 3.2.7 השפעת שיטת המבחן ואופנותו על תוצאותיו
  - 3.2.8 סוגי מבחנים חדשים והעשרת מדדי תגובה
- 3.3 תגובת הנבחנים והציבור
  - 3.3.1 הנבחנים
  - 3.3.2 הציבור

**4. המעבר למבחנים ממוחשבים**

**5. ס י כ ו ס**

**6. מקורות**

## ת מ צ י ת

מבחנים ממוחשבים, ובמיוחד אדפטיביים, נדונים במאמר זה כפי שהם משתקפים בספרות המקצועית העדכנית.

שני יתרונות עיקריים למבחנים ממוחשבים: היתרון הראשון הוא האפשרות להתאימם לכל נבחן ונבחן על-ידי בחירה נבונה של פריטים המותאמים לכשריו של הנבחן. היתרון השני הוא באפשרות ליישם סוגי מבחנים ומדדי תגובה חדשים שלא ניתן ליישם במבחני נייר ועפרון. בנוסף לכך, יש למבחנים ממוחשבים יתרונות ניכרים בתחום המנהלי והתפעולי; הם מאפשרים בקרה טובה יותר על מהלך הבחינה, ניתן למנוע בעזרתם טעויות אדמיניסטרטיביות וניתן לייצל את הבחינה ובכך לחסוך זמן.

לעומת זאת, מיחשובם של מבחנים אינו פשוט. פרט לעובדה שמיחשובם מחייב השקעת ידע ומשאבים כספיים, קיימות בעיות תיאורטיות שעדיין לא נפתרו. בעיות אלו מתייחסות בעיקר לתורת התגובה לפריט (IRT), תיאוריה העומדת בבסיסם של מבחנים אדפטיביים.

החשש שמא בחינה ממוחשבת תעורר חרדה, וכך תפגע בנבחנים, אינו מוצדק כנראה; חשש זה מקורו כנראה בעמדותיהם של הבוחנים יותר מאשר באלו של הנבחנים עצמם.

בעבודה זהירה של מחקר ויישום ניתן יהיה להגיע למבחנים ממוחשבים מדויקים ויעילים, העשויים למדוד נאמנה את כל מרחב הכשרים והידע של הנבחנים.

## 1. מבוא

בתחילת המאה פירסמו Binet ו-Simon את מבחן האינטליגנציה הנקרא על שמם. מבחן זה הוא מבחן יחידני, המועבר על ידי בוחן מאומן. גם במבחן המקורי וגם בפיתוחים המאוחרים יותר (Stanford-Binet), מתאים הבוחן את המבחן ליכולתו של הנבחן; הוא בוחר את פריט המבחן הראשון (השאלה הראשונה) על-פי גילו של הנבחן ומפסיק את הבחינה כאשר הנבחן מגיע לרמת ביצוע שנקבעת מראש. בשמונים השנים שחלפו מאז פירסום המבחן, לבשו מבחני האינטליגנציה ומבחנים דומים להם צורות שונות ובהדרגה ירדה חשיבותו של המבחן היחידני המותאם לרמת הנבחן, והתחזקה המגמה לבחון קבוצות גדולות של אנשים בתנאים תקינים וקבועים. מקורה של המגמה הוא בשאיפה להוריד את עלות הבחינה בלשכות הגיוס של צבא ארה"ב, אשר נדרשו למיין כמליון וחצי מגויסים לקראת סוף מלחמת העולם הראשונה (ראה סקירות היסטוריות אצל Gould, 1981; Carroll, 1982).

שתי התפתחויות שחלו בשלושים השנים האחרונות, מאפשרות לשוב כיום אל שיטת הבחינה היחידנית, המותאמת ליכולתו של הנבחן. התפתחות אחת היא טכנולוגית: העליה העצומה בכוח החישוב, שחלה בד בבד עם הירידה המשמעותית במחירי המחשבים והציוד ההיקפי. התפתחות זו מאפשרת להחליף בתנאים מסויימים את הבוחן האנושי בתוכנית מחשב. התפתחות שניה היא תיאורטית: הופעתן של תורות מבחנים מסוג חדש - תורות התכונה החבויה (Latent Trait Theories), או, בשם אחר, תורות התגובה לפריט (Item Response Theories) או בקיצור (IRT). על תורות אלה ניתן לבסס מבחנים מותאמים לנבחן, שכן הן מאפשרות להביא למכנה משותף (סולם ציונים אחיד) מבחנים המודדים את אותו הכושר או אותה היכולת, אך מורכבים מקבוצות פריטים שיש ביניהן חפיפה מועטת ואף כאלה שאין ביניהן חפיפה כלל. יישומן של תיאוריות אלה אפשרי רק בעזרתם של מחשבים מהירים, הן בשלב פיתוח המבחן, הן בשלב קביעת הציון של הנבחן.

אין ספק שבעשורים הבאים נהיה עדים למספר הולך וגדל של מבחנים פסיכולוגיים ממותשבים, ובמילותיה של אנטטזי (1986):

"Because this development is now at an early stage and is rapidly gaining momentum, it is likely to constitute a major feature of psychological assessment in the twenty-first century".

האיגוד המקצועי והשוק הפרטי, שניהם כבר הצטרפו למגמה זו; האיגוד האמריקאי לפסיכולוגיה כבר פירסמה הנחיות לשימוש במבחנים ממותשבים (APA, 1986a), וחברה פרטית הוציאה לאור קטלוג מוצרי-תוכנה עבור מבחנים ממותשבים (Krug, 1984).

במאמר זה נשווה בין מבחנים קונבנציונליים לבין מבחנים ממותשבים. ננסה לעמוד על היתרונות ועל המגבלות של מבחנים ממותשבים לעומת הקונבנציונליים, משלוש נקודות מבט: נקודת המבט המינהלית והטכנית של העברת מבחנים, נקודת המבט הפסיכומטרית,

ונקודת המבט של הנבחן והציבור. מאחר ובתחום המושגים חלה התפתחות מהירה ומתמדת, ומאחר ותחום המבחנים הממוחשבים מצוי בתחילת דרכו, אין ספק שחלק מן הדברים הנכתבים כאן יתיישנו במהרה. הנכתב כאן הוא, איפוא, בבחינת צילום-מצב עכשווי, ורבות מן המגבלות של המבחן הממוחשב ייעלמו בעתיד הנראה לעין.

התחום של מיחשוב כלים למדידה פסיכולוגית הוא תחום רחב. במאמר זה לא נדון במיחשוב של מבחני אישיות וראיונות, ואף לא בפירוש של תוצאות מבחנים (על נושאים אלה ראה סקירות: Fowler, 1985; Space, 1981; Skinner & Pakula, 1986). כמו-כן, לא נעסוק במבחנים שמטרתם היא תירגול ואימון, כי אם רק באלה שמטרתם הסופית היא קביעת רמתו של הנבחן לצורך דירוג, מיון והשמה.

משקל מיוחד ניתן במאמר זה לדיון במבחנים גמישים, אלה המותאמים ליכולתו של כל נבחן ונבחן. זאת משום שעיקר התועלת הפסיכומטרית הגלומה במיחשוב מבחנים, היא היכולת להתאימם אינדיבידואלית לנבחנים. עם זאת, צעד זה אינו פשוט, והוא אחד האתגרים הגדולים ביותר העומדים כיום בפני החוקרים בתחום תורת המבחנים.

בטרם נדון ביתרונות ובחסרונות של מבחנים ממוחשבים, נייחד פרק לתיאור צורות בחינה שונות, בו נרחיב את תיאור המבחנים הממוחשבים המבוססים על IRT. זאת משום שכמה מיתרונותיהם, וגם כמה מחסרונותיהם (נכון לעכשיו) של מבחנים מסוג זה, קשורים קשר הדוק לתיאוריה העומדת בבסיסם. בעקבות פרק זה יוצגו היתרונות והחסרונות של מבחנים ממוחשבים משלוש נקודות המבט שהוזכרו לעיל; לבסוף, בטרם סיכום, יובא פרק קצר על הצעדים אותם יש לנקוט במעבר ממבחן קונבנציונלי למבחן ממוחשב.

## 2. סיווג מבחנים ממוחשבים.

אחד המימדים שלפיהם ניתן לסווג את המבחנים הממוחשבים הוא מידת הגמישות (אדפטיביות) שלהם; באיזו מידה הם ניתנים להתאמה לרמת הידע של הנבחן או לכושרו. הרצף נע בין מבחנים קשיחים לחלוטין - בהם ברור וידוע מראש מה תהיה קבוצת הפריטים שתוצג בפני הנבחן, ועד מבחנים גמישים ביותר - בהם בחירת הפריט הבא שיוצג בפני הנבחן תלויה ברמת היכולת שהפגין הנבחן עד כה והמשתקפת בסדרת הפריטים שהוצגו ובתשובותיו לכל פריט ופריט. המבחנים הגמישים ביותר מגיעים למיצוי מירבי של הפוטנציאל הטמון במחשב.

### 2.1 מבחנים קשיחים.

במבחנים אלה המחשב משמש לצורך הצגת ההוראות ופריטי התירגול, הצגת הפריטים התפעוליים, קליטת תגובות הנבחנים, וצייונו (scoring) של המבחן מיד עם סיומו. קבוצת הפריטים וסידרם קבועים מראש, ומבחינה זו אין המבחנים הללו שונים במידה

רבה ממבחנים קבוצתיים קונבנציונליים. השימוש במחשב מאפשר הקפדה על זמני הבחינה, שהינה קריטית במבחני מהירות (speeded tests), ואף מאפשר לשלוט על זמן הצגתו של כל פריט ופריט. מאחר ויש שליטה מלאה על הצגת הפריטים, ניתן לקבוע אם יותר לנבחן לחזור לפריטים שעליהם כבר השיב. בהינתן אפשרות כזו, יהיו מבחנים אלה דומים ביותר למבחנים קונבנציונליים, בהם הנבחן יכול לדלג על פריט על-מנת לחזור אליו מאוחר יותר, או אף לחזור לפריטים שעליהם כבר השיב, לצורך עיון מחודש ושינוי התשובה בעת הצורך.

מיחשוב מבחנים קשיחים הינו תהליך פשוט יחסית, ועדות לכך היא השפע היחסי של תוכנות-מדף עבור מבחנים כאלה (ראה למשל מודעות פרסומת ב-APA, 1986). למיטב ידיעתנו, המבחנים הממוחשבים היחידים למדידת כשרים המיושמים בארץ בהיקף תפעולי מלא, הם מבחנים קשיחים (בן-דור, 1985).

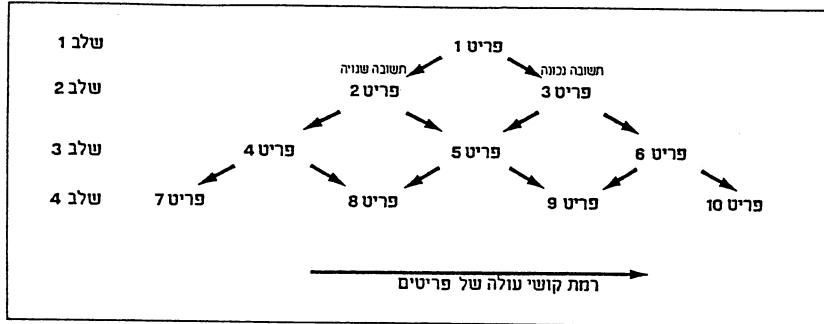
## 2.2 מבחנים גמישים בעלי הסתעפות קבועה.

במבחנים אלה בא לידי ביטוי כוחו של המחשב בהתאמת הפריטים לרמת הידע או הכושר של הנבחן. בכל שלב ושלב של המבחן, לאחר הצגת סדרת פריטים, נקבעת רמת הקושי של סדרת הפריטים שתוצג לנבחן לאור תשובותיו לסדרת הפריטים הקודמת. במידה וביצועו של הנבחן הוא מעל רמה מסויימת, הנקבעת מראש, תוצג בפניו בשלב הבא סדרת פריטים קשה יותר.

רבות הן הוואריאציות של מבחנים מסוג זה והן נבדלות זו מזו במספר השלבים, במספר ברירות ההחלטה שעומדות בסוף כל שלב, ובמספר הפריטים המוצגים בכל שלב. המבחנים הפשוטים ביותר בנויים משני שלבים ומחייבים החלטה אחת בלבד, אשר יש לקבלה בתום השלב הראשון (Cleary, Linn & Rock, 1968; Lord, 1971a).

כל נבחן עובר את השלב הראשון, ובהתאם לציונו הוא מופנה לאחת משתי סדרות הפריטים להמשך המבחן. סדרות הפריטים המופיעות בשלב השני אינן חייבות להוציא זו את זו; תיתכן חפיפה ביניהן. במבחנים המורכבים ביותר מוצג בכל שלב פריט אחד בלבד, וההחלטה מתבצעת לאחר כל פריט. מקרה מיוחד של מבחן רב-שלבי, שבו פריט אחד בכל שלב, הוא המבחן הפירמידלי המתואר בצירוף 1 (Weiss, 1982). לכל הנבחנים במבחן זה מוצג פריט 1. לנבחנים שהשיבו עליו נכונה מוצג פריט 3 שהוא קשה יותר, ולאילו שגגו בתשובתם מוצג פריט 2 שהינו קל יותר. בחירת הפריט הבא שיוצג לנבחן תלויה, איפוא, בתשובתו לפריט הקודם, ואינה יכולה להתבצע לפני שהנבחן משיב בפועל. המבחן מסתיים לאחר שלכל הנבחנים הוצג מספר פריטים שנקבע מראש. בדוגמה המוצגת בצירוף 1 מסתיים המבחן לאחר הצגת ארבעה פריטים. במבחנים גמישים בעלי הסתעפות קבועה, מתעוררת לראשונה בעית כיוול (calibration) בין סולמות ציונים שונים. למרות שלכל הנבחנים מוצג מספר זהה של פריטים, לא ניתן לקבוע את הציון על סמך מספר התשובות הנכונות, מאחר ולנבחנים שונים מוצגות סדרות פריטים שונות, הנבדלות זו מזו ברמת הקושי של הפריטים. במבחן המתואר בצירוף 1, למשל, מוצגת לכל נבחן את משמונה סדרות פריטים שונות.

**ציור 1 תיאור סכימטי של מבחן פירמדלי.**



בארץ נערכים ניסויים על-ידי מחלקת מדעי ההתנהגות של צה"ל, למטרת בדיקת אפשרויות היישום של סוללות מבחנים גמישות בעלות הסתעפות קבועה (Dover, 1986). כמו כן, המרכז לטכנולוגיה חינוכית פיתח ומפעיל מערכות תרגול ואבחון באמצעות מחשב (מערכות תוא"ם). אחד ממרכיבי המערכות הללו מיועד למדוד את רמת השליטה של התלמידים בחשבון, באנגלית, ובהבנת הנקרא, והוא למעשה מבחן גמיש בעל הסתעפות קבועה (אוסין ונשר, 1979; Osin, 1984).

**2.3 מבחנים גמישים**

בשם זה נכנה את המבחנים בהם מבנה ההסתעפויות אינו קבוע מראש, אלא נקבע תוך כדי העברת המבחן. את מבנה המבחנים הללו לא ניתן לתאר על-ידי גראף בצורת עץ או בצורת פירמידה, בהם מופיע כל פריט פעם אחת בלבד. במבחנים כאלה עדיף לנסח ולתכנת אלגוריתם של בחירת פריטים במקום לקבוע את סדרם מראש.

נתאר בקצרה שלושה סוגי מבחנים המופיעים תחת הכותרת "מבחנים גמישים": "מבחן הרמה הגמישה" (Flexilevel test), "מבחן גמיש-שכבתי" (Stradaptive test) ומבחנים המבוססים על IRT. מבחן הרמה הגמישה הוצע על-ידי Lord (1971b, 1980) והוא ניתן ליישום גם ללא מיחשוב. במבחן זה נתון מאגר של  $2n+1$  פריטים וכל נבחן עונה על  $n+1$  פריטים מתוכם. הפריטים סדורים לפי רמת הקושי, והפריט הראשון המוצג לכל הנבחנים הוא הפריט האמצעי. פריט זה מחלק את מאגר הפריטים כך שמחציתם "קשים" ממנו ומחציתם "קלים" ממנו. בחירת הפריט הבא בכל צעד תלויה בתשובת הנבחן לפריט הקודם. אם תשובתו היתה נכונה, יוצג בפניו הפריט הקל ביותר מבין כל הפריטים "הקשים" שעדיין לא הוצגו לו; אם שגה בתשובתו, יוצג בפניו הפריט הקשה ביותר מבין כל הפריטים "הקלים" שעדיין לא הוצגו לו. המבחן מסתיים לאחר שהוצגו בפני הנבחן

n+1 פריטים. ציונו של הנבחן הוא מספר התשובות הנכונות שנתן (בתוספת מחצית הנקודה לנבחן שתשובתו לפריט האחרון הייתה שגויה). Lord הראה שדירוג הנבחנים על-פי ציוניהם במבחן זה משקף נאמנה את יכולתם. (ראה מחקרים על מבחן הרמה הגמישה אצל אליוף, 1987; DeAyala & Koch, 1986).

הכללתו של מבחן הרמה הגמישה יכולה להיעשות על-ידי כך, שבמקום כל פריט עומדת קבוצת פריטים והמעבר מקבוצה לקבוצה נקבע על-פי מספר התשובות הנכונות בכל קבוצה, או על-פי אומדן אחר של היכולת; אם מספר הקבוצות הינו קטן, ניתן לראות מבחן זה כמבחן גמיש בעל הסתעפות קבועה. מבחן כגון זה הוצע על-ידי Weiss (1973) וכונה על-ידו "מבחן גמיש שכבתי" (Stratified adaptive או בקיצור Stradaptive). מבחן זה, כמו מבחן הרמה הגמישה נחקר מעט (למשל Hicks, 1986; Waters, 1977) ומירב המאמץ המחקרי מכוון כיום למבחנים גמישים המבוססים על IRT והמתוארים להלן.

המבנה הכללי של מבחן גמיש המבוסס על IRT הוא כדלהלן: נניח שהנבחן ענה כבר על מספר פריטים. על-סמך פריטים אלה, ניתן לחשב אומדן לרמת הידע שלו או ליכולתו. בהינתן רמת ידע או יכולת זו, נבחר הפריט הבא אשר יוצג לנבחן, וזהו בדרך כלל הפריט המבחין ביותר עבור נבחנים שהם ברמת יכולת דומה. הפריט מוצג, הנבחן משיב, רמת יכולתו נאמדת שוב, וחוזר חלילה. בתחילת המבחן - אם אין כל מידע מוקדם על יכולתו של הנבחן - מוצג בדרך כלל הפריט המבחין ביותר עבור ממוצע האוכלוסייה שלה נועד המבחן. המבחן מסתיים לאחר שהוצג לנבחן מספר פריטים קבוע, או לאחר שהושגה רמת דיוק המדידה אשר הוגדרה מראש. אומדן רמת היכולת של הנבחן, בחירת הפריט המתאים ביותר ואומדן לטעות המדידה בכל שלב ושלב של המבחן, יכולים להישען רק על תורה פסיכומטרית חזקה ביותר. תורה זו (IRT), פותחה במשך 40 השנים האחרונות ותחום היישום שלה הולך ומתרחב. מאחר וכמה מן המגבלות של מבחנים גמישים, מקורן בהנחות החזקות של התורה, יובא בסעיף הבא תיאור קצר של עיקריה.

יישומם של קבחים גמישים הולך ומתרחב, כאשר עיקר הפעילות מתבצעת בצבאות של מדינות שונות (ארה"ב: McBride & Martin, 1983; McBride & Sympson, 1985; גרמניה המערבית: Wildgrube, 1983) אך גם בחברות פרטיות (Vale, 1985).

#### 2.4 תורת התגובה לפריט - (Item Response Theory)

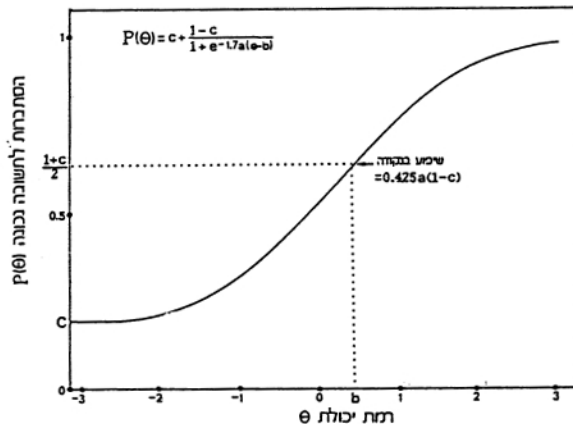
למעשה אין זו תורה אחת, כי אם משפחה של מודלים המכונה בספרות בשני שמות: "מודל התכונה החבויה" (Latent Trait Model) על-ידי Birnbaum (1986), או "תורת התגובה לפריט" על-ידי Lord (1980). אחד הווריאנטים של התורה, אשר נחקר רבות, ידוע בשם "המודל של Rasch" (Rasch, 1980). הרעיון הבסיסי של התורות הללו הוא, שההסתברות של נבחן לענות נכונה על פריט מסוים היא פונקציה מונוטונית עולה של יכולתו. יכולת זו היא "אימיתית" ועל כן היא חבויה (Latent) ואינה בת תצפית. התורות השונות מניחות כי לפונקציה צורה מוגדרת ובכולן היא דמוית S. צורתה המדויקת של



הפונקציה ומספר הפרמטרים שלה משתנים מתורה לתורה. קיים דמיון רב בין פונקציה זו לבין פונקצית הסף בפסיכופיזיקה (אשר כונתה אף בשם "פונקציה פסיכומטרית" על-ידי Urban ב-1908. ראה Boring, 1942, p. 39. בשתיהן, מוגדרת המדידה במונחים הסתברותיים, אלא שבעוד האחת מייחסת את הסתברות התגובה לגדלים פיזיקליים שניתן למדודם בנקל, הרי השנייה מייחסת את הסתברות התגובה לעוצמתה של תכונה חבויה, שאינה ניתנת למדידה ישירה. עקומת הפונקציה המתארת את הקשר בין רמת היכולת של הנבחן לבין ההסתברות שישב תשובה נכונה על פריט מסוים, נקראת "העקומה האופיינית לפריט" (Item Characteristic Curve).

דוגמה לעקומה כזו, הנגזרת מן המודל הנחקר ביותר מבין תורות התגובה לפריט, המודל הלוגיסטי התלת-פרמטרי, מוצגת בציור 2. על הציר האופקי בציור מיוצגת רמת היכולת של הנבחן ( $\theta$ ) ועל הציר האנכי מיוצגת ההסתברות לענות נכונה על הפריט. צורתה הכללית של העקומה נקבעת על-ידי שלושה פרמטרים: הראשון הוא הפרמטר הקובע את האסימפטוטה הנמוכה של העקומה ( $c$ ) והוא מכונה פרמטר הניחוש או פרמטר הניחוש-כביכול ("pseudo guessing"). ככל שפרמטר זה גבוה יותר, כן גדלה ההסתברות לכך שנבחנים בעלי רמת יכולת מיזערית יענו נכונה על הפריט, בשל ניחוש מוצלח או בשל הטיית-תגובה (response bias). הפרמטר השני ( $b$ ), מתייחס לקושי הפריט והוא הערך של  $\theta$  שעבורו ההסתברות לענות נכונה על פריט שווה למחצית הערך  $1-c$ . ככל שערכו של  $b$  גבוה יותר (אם שאר הפרמטרים מוחזקים קבועים), יתקשה נבחן בעל רמת יכולת נתונה לענות עליו. הפרמטר השלישי ( $a$ ) מתייחס לשיפוע העקומה בנקודה  $b$  ומייצג את רמת האבחנה של הפריט. ככל שהשיפוע תלול יותר, כך תשתפר ההבחנה בין נבחנים שיכולתם היא בטווח שבו העקומה תלולה.

ציור 2 העקומה האופיינית של פריט. מודל לוגיסטי תלת-פרמטרי.



כאמור, תורת התגובה לפריט היא בעצם משפחה שלמודלים, ובחקר הנוכחי ראוי לציין שני ווריאנטים נוספים. הראשון הוא המודל הלוגיסטי החד-פרמטרי, המכונה גם המודל של Rasch. במודל זה מניחים שפרמטר הניחוש (c) שווה לאפס עבור כל הפריטים (כלומר: ההסתברות שנבחן שרמת יכולתו נמוכה ישיב נכונה על הפריטים קרובה לאפס), וכן, שפרמטר האבחנה (a) הינו זהה עבור כל הפריטים. במודל השני - המודל הלוגיסטי הדו-פרמטרי - מניחים שפרמטר הניחוש שווה לאפס, והפריטים נבדלים זה מזה ברמת הקושי (b) וברמת האבחנה (a). מודל זה נחקר פחות מן האחרים; מירב המאמץ מושקע במחקר המודל הלוגיסטי התלת-פרמטרי והחד-פרמטרי.

לפי תורת התגובה לפריט (תת"ל), מבחן מיוצג ע"י אוסף עקומות-התגובה האופייניות של הפריטים המרכיבים אותו. מתוך העקומות הללו ניתן לחשב מדדים כגון מידת ההבחנה שמספק המבחן בכל תחום של רמות יכולת. ציונו של נבחן על-פי תת"ל נקבע בשיטת אומדן הנראות המירבית; בהינתן תשובותיו של נבחן לקבוצות פריטים, ניתן לשאול את השאלה: אילו היה הנבחן ברמת  $\theta$  מסוימת, מה היתה ההסתברות לכך שיענה על הפריטים כפי שענה למעשה. כציונו של הנבחן נבחרת אותה רמה בה הסתברות זו היא מירבית. בשיטת אומדן כזו, מניחים שההסתברות של נבחן לענות נכונה על פריט א', אינה תלויה בהסתברות שישב נכונה גם על פריט ב'. הנחה זו אינה דורשת שלא יהיה מתאם בין הפריטים; היא רק דורשת שעבור רמת  $\theta$  קבועה לא יהיה מתאם כזה, ועל-כן היא מכונה בשם הנחת "אי-התלות המקומית" (local independence). קיום ההנחה גורר, בין היתר, את הדרישה שהפריטים במבחן יהיו חד-מידיים. במילים אחרות, מבחן המבוסס על תת"ל יכול וצריך למדוד תכונה אחת בלבד. לצורך מדידת תכונות נוספות נחוצים מבחנים כמספר התכונות.

בשל חוזקן של הנחות תת"ל - צורת העקומה האופיינית לפריט, אי-התלות המקומית וחד-מימיות המבחן - מוטל על מי שמפתח מבחנים המבוססים על התורה להוכיח שההנחות מתקיימות. לחילופין, אם ההנחות אינן מתקיימות, עליו להראות שאי-קיומן אינו מחליש את המסקנות שהוא מסיק מתוצאות המבחנים. בפרק הבא ידונו ביתר הרחבה המגבלות הפסיכומטריות הנובעות משימוש במבחנים המבוססים על תת"ל, והמחקר שבוצע בדבר ישימותה של התיאוריה.

### 3. יתרונות וחסרונות

#### 3.1 בתחום המינהלי והטכני

המגבלה הגדולה ביותר של מיחשוב היא החשקה הכלכלית הנדרשת על-מנת לפתח מערכת העונה על הדרישות. מעבר לכך, יש להכשיר כח אדם המסוגל לתפעל ולתחזק את המערכת, ויש להביא בחשבון את העובדה שרמת הניוד של המערכת יורדת. מבחינות אלו, אין הבדל בין מיחשוב מבחנים לבין מיחשוב של פונקציות אחרות בארגון. כמו כן, בדומה למערכות בנקאיות וצבאיות, מערכת מבחנים ממוחשבת חייבת להיות מוגנת בפני תקלות מיכניות וערוכה להתגברות עליהן. מבחן שהחל ונקטע במהלכו בשל

תקלה, לא ניתן להתחילו מחדש כאילו לא אירע דבר. לא ניתן לבקש מנבחן לחזור ולהשיב על פריטים שעליהם כבר ענה, כאילו לא ראה אותם בעבר. הדבר נכון במיוחד במבחנים בהם ניתן לנבחן משוב מיידי לאחר כל פריט. קיימים פתרונות טכנולוגיים לצמצום שיעור התקלות, למשל, באמצעות שימוש במערכת כפולה, בה לכל רכיב מכני או אלקטרוני יש גיבוי, ובמקרה של תקלה, מוחלף הרכיב באופן אוטומטי על-ידי כפילו. אולם ברור שפתרונות מסוג זה, מייקרים את המערכת באופן ניכר. את ההשקעה הכלכלית והמאמץ האנושי הכרוכים במיחשוב יש לשקול, איפוא, כנגד התועלת הצפויה מן המיחשוב. התועלת מתמצה בשלושה תחומים: שיפור הבקרה והניהול, צמצום שיעורי הטעויות, והחיסכון בזמן.

שיפור הבקרה והניהול מקורו ביכולת של מערכת ממוחשבת לרשום ולתעד כל פעולה המתבצעת בה. מובן שלצורך כך יש להקים מערכת תוכנה וניהול נתונים מתאימה, אולם מרגע שזו מוקמת ופועלת כהלכה, יש בידי המשתמש יכולת רישום נתונים ושליפתם באופן שהינו מוגבל כמעט אך ורק על-ידי קיבולת אמצעי האיחסון העומדים לרשות המערכת.

שיעור הטעויות במערכת ממוחשבת הינו זעום בהשוואה למערכות מבחנים קונבנציונליות. בין קבלת תשובותיו של הנבחן לבין חישוב ציונו, אין תיווך של ניקוב או ציינון ידני של המבחן, ששיעור הטעות בהם גבוה. מובן שניתן להוריד את שיעורי הטעות במערכות קונבנציונליות באמצעות בדיקה חוזרת של כל ציון, בין אם בצורה ידנית, בין אם על-ידי קלידה חוזרת. אך גם במערכות קונבנציונליות המשתמשות בסורקים אופטיים לקליטת גליונות תשובה (אנקווה, 1985) יש עדיין טעויות בשיעור של 0.1%, ושיעור זה נמוך בהרבה משיעור הטעות שבבדיקה ידנית כפולה, המגיע ל- 0.5%-1.0% (בן-דור, 1985). טעויות מסוג אחר, הנחסכות על-ידי מיחשוב המבחנים, הן טעויות אנוש בזמן העברת המבחנים. גם בוחן מאומן ביותר יכול להיכשל במתן הוראות למבחן, או לטעות במדידת זמני הבחינה. טעויות כאלה מפרות את התיקניות של תנאי הבחינה ועל כן עלולות לפגום בתקפות המבחן.

מיחשוב המבחנים חוסך הן מזמנו של הנבחן, הן מזמנו של הבוחן. מקור אחד לחיסכון בזמן הוא היכולת לספק לנבחן את תוצאות הבחינה מיד עם סיומה. נחסכים השלבים המייגעים של חישוב קבוצתי של הציונים עם תום הבחינה והפצתם לנבחנים או לצרכנים האחרים לאחר מכן. מקור שני לחסכון בזמן מושג באמצעות מבחנים גמישים לסוגיהם. מאחר ומבחנים אלה מותאמים לרמתו של הנבחן, אין צורך להציג לנבחן שרמתו נמוכה את הפריטים הקשים יותר, שכן הסיכוי שיענה עליהם נכונה, ולא-על ידי ניחוש, הינו זעום. אם רמתו של הנבחן גבוהה, אין צורך להציג בפניו את הפריטים הקלים ביותר, שכן מלכתחילה קיים סיכוי גבוה שיענה עליהם נכונה.

נערכו מספר מחקרים לבדיקת מספר הפריטים שניתן לחסוך באמצעות מבחן גמיש. התוצאות הדרמטיות ביותר הושגו על-ידי Uzey (1977), אשר טוען כי ניתן להגיע במבחן גמיש לאותה רמת מהימנות של מבחן קונבנציונלי עם מחצית ממספר הפריטים,

ובמקרים מסויימים אף עם חמישית ממספר הפריטים. הביקורת על ממצאיו מתמקדת בגודל המדגם שעליו התבסס ובטכניקות הסטטיסטיות שעליהן ביסס את אומדני המהימנות. במחקר של Weiss (1982) נבדקה שאלת קיצור המבחן הגמיש על סוללה בת 12 מבחנים. נמצא כי במוצע, במבחן גמיש יש צורך ב-40%-65% ממספר הפריטים הנחוצים במבחן קונבנציונלי, על-מנת להגיע לאותה רמת דיוק במדידה.

McBride & Martin (1983) בדקו מבחנים גמישים בתנאים תפעוליים על המתגייסים לחיל הנחתים של הצי האמריקאי. הם בחרו במבחן יכולת מילולית - אחד מתוך סוללת ASVAB של עשרה מבחנים הבודקים כשרים קוגניטיביים וחמשמת את הצבא האמריקאי למיון מקצועי ראשוני של המתגייסים לשורותיו. לצורך אומדן המהימנות הורכבו שני מבחנים קונבנציונליים מקבילים ושני מבחנים גמישים מקבילים. כל המבחנים הועברו על-ידי מחשב, וכל מתגייס השיב על כל הארבעה. במערך הניסוי היה איזון של סדר המבחנים, ופריטי המבחנים המקבילים מכל סוג עורבבו זה בזה. מסקנות החוקרים הן, כי עבור היישום הספציפי אותו בדקו, ניתן להגיע לרמות דיוק גבוהות של המדידה עם פחות ממצחית מספר הפריטים של מבחן קונבנציונלי. ממצא זה חזר על עצמו גם לגבי מבחנים אחרים של סוללת ה-ASVAB (Moreno, Wetzel, McBride & Weiss, 1984).

הממצאים לעיל מראים שבמעבר ממבחן קונבנציונלי למבחן גמיש ניתן להפחית ממספר הפריטים, ללא פגיעה בדיוקה של המדידה. עדיין יש לבדוק באיזו מידה קיצור המבחן מתבטא גם בקיצור הזמן הדרוש להעברתו, אך סביר להניח, שגם אם זמן התגובה לפריט במבחן גמיש מתארך מסיבה כלשהי, הזמן הנדרש להעברת המבחן הגמיש כולו עדיין קצר מן הזמן הדרוש להעברת מבחן קונבנציונלי. עדות אחת לכך ניתן למצוא במחקרם של English ועמיתיו (English, Reckase & Patience, 1977), שבדקו את איכותו של מבחן הישגים גמיש. מסקנתם היא, שניתן לחסוך כ-75% מן הפריטים ובכך להגיע לחיסכון של 50% בזמן הבחינה, ללא פגיעה בטיב המדידה. יש לזכור כי הזמן המוקצב למבחני עוצמה (power tests) נקבע בדרך כלל על-פי מהירות הביצוע של הנבחנים האיטיים ביותר. מעצם טבעו של המבחן הגמיש היחידני, אין הגבלה זו חלה עליו; הנבחנים האיטיים ביותר יכולים לנצל זמן רב מכפי שהיה ניתן להם במבחן קשיח, ואילו הנבחנים הזריזים אינם חייבים להמתין לטיוס המבחן על ידי נבחנים אחרים.

אחת הבעיות הטכניות המתעוררות בעיקר בהקשר של מבחן גמיש המבוסס על תת"ל, היא חוסר האיזון בשכיחות הצגתם של הפריטים לנבחנים. בשל הכלל הנהוג במבחנים מסוג זה - שהפריט הבא שיוצג לנבחן הוא הפריט המבחין ביותר עבור רמת יכולתו כפי שנאמדה עד כה - יש הטייה הגורמת לכך שהפריטים המבחנים ביותר הנמצאים במאגר הפריטים ייבחרו ויוצגו לנבחנים שוב ושוב (Tung, 1986). כך למשל, בצורתו הטהורה של המבחן, קיים פריט אחד במאגר הפריטים המוצג לכל הנבחנים, וזאת משום שהוא הפריט המבחין ביותר עבור רמת יכולת ממוצעת. אם המבחן מועבר לאוכלוסייה גדולה בזמנים שונים, קיים חשש כי תוכנו של פריט כזה יהיה לנחלת הכלל בתוך זמן קצר, ושוב לא יהיה טעם בשימוש בו. על מנת להתגבר על בעיה זו, אפשר

לחכנים מרכיב של בחירה מקרית בין פריטים דומים זה לזה בכל צעד במבחן שבו יש לבחור פריט חדש להצגה. מנגנון כזה יבטיח אמנם התגברות על החטייה בבחירת הפריטים, אולם מאידך, יביא למדידה תת-מיטבית ובעקבותיה - להארכת המבחן.

ניתן למנות שני יתרונות מנהליים נוספים שיש למבחנים ממוחשבים, אך אלה משניים בחשיבותם. היתרון הראשון הוא באופן בדיקתם-מראש של פריטים נסיוניים. במבחנים קונבנציונליים נהוג להוסיף פריטים נסיוניים לפריטים התפעוליים, על-מנת להפיק מידע סטטיסטי עליהם עוד בטרם נעשה בהם שימוש תפעולי. הבחן מעוניין, כמובן, שהנבחנים יענו על פריטים אלו כפי שהיו עונים על פריטים תפעוליים, אך לא תמיד קל להשיג את שיתוף הפעולה של הנבחנים, בעיקר במידה והנבחנים יכולים לזהות בנקל אילו מן הפריטים הם נסיוניים ואילו אינם, משום שבדרך-כלל מוצגים הפריטים הנסיוניים כתוספת למבחן התפעולי ואינם מהווים חלק טבעי שלו. במבחנים ממוחשבים קל יותר להציג את הפריטים הנסיוניים ברצף אחד עם התפעוליים, קל יותר להפריד אותם מן התפעוליים בזמן הצינון, וקל יותר לשלוט בגודל מדגם הנבחנים ולהתאימו לגודל הרצוי לצורך אומדן טוב של הפרמטרים.

היתרון המינהלי השני קשור לבטיחותם של הפריטים, אך אין הסכמה מלאה לכך שאכן זהו יתרון. ניתן לטעון, כי מאחר והנבחן אינו מקבל לידי את פריטי המבחן כשהם מודפסים, יהיה קשה לו, להעביר לאחר את תוכנה של הבחינה. כמו כן, במבחן ממוחשב נחשכים שלבי הדפסת המבחן, כריכתו, הפצתו ואיסונו, שלבים שעלולים להיות קריטיים לגבי בטיחותו. מאידך, ניתן לטעון, כי לאור התחכום הרב של עברייני המחשבים, הסכנה הבטיחותית קיימת ועומדת גם לגבי מאגרי מבחנים ממוחשבים, והיא אולי אף תמורה יותר שכן פריצה אחת למאגר, די בה כדי לסכן את כולו. יש פנים לכאן ולכאן, ובעצם, קשה לחשוות בצורה כוללת את בטיחותו של מבחן ממוחשב לזו של מבחן קונבנציונלי, מבלי לפרט מהם אמצעי הבטחון הספציפיים הננקטים הלכה למעשה בכל אחת מן השיטות ובכל מקרה ומקרה.

### 3.2 בתחום הפסיכומטרי

#### 3.2.1 צורת העקומה האופיינית של פריט

אחת הסיבות לכך ש-F. Lord נמנע במשך שנים רבות ממחקר בתחום התת"ל, היא ההנחה החזקה של התיאוריה בדבר צורתה של העקומה האופיינית לפריט (Lord, 1980, p. 17). אולם מתקר רחב-היקף (Lord, 1970) הראה שהפונקציה הלוגיסטית בעלת שלושה פרמטרים, הולמת יפה עקומות אופייניות אמפיריות שנאמדו באופן בלתי-תלוי ע"י פונקציות בעלות כ-180 דרגות חופש. מתוך 150 פריטי SAT-V שנבדקו, רק בשישה התגלתה סטייה ממונוטונית בעקומה האופיינית של הפריט, וסטייה זו היתה מזערית.

במחקר מאוחר יותר, שבוצע על כל סוגי הפריטים המופיעים בבחינות ה-GRE (Kingston & Dorans, 1982, 1985), נבדקה התאמת הפונקציה הלוגיסטית התלת-פרמטרית לעקומות אופייניות אמפיריות, ונמצא שמידת ההתאמה תלויה בעיקר בסוג הפריט. ההתאמה הטובה ביותר נמצאה עבור כל הפריטים המילוליים, ושניים מסוגי הפריטים

האנליטיים (דיאגרמות לוגיות וחשיבה אנליטית). התאמה טובה פחות נמצאה עבור פריטי מתימטיקה מקובלים ועבור פריטים שבהם מתבקש הנבחן לפרש נתונים המוצגים לו. פריטים שבהם נמצאה בבירור אי-התאמה למודל הם פריטי "ניתוח הסברים" ופריטי השוואות-כמותיות. אי-ההתאמה שמוצאים לעיתים בין הנתונים לבין הפונקציה האופיינית לפריט אינה מכשול בלתי-עביר. קיימים בספרות נסיונות חדשים לבסס תת"ל על צורות פונקציונליות אחרות פרט לפונקציה הלוגיסטית או הנורמלית-מצטברת (Winsberg, Thissen & Wainer, 1984; Lord, 1984), שיאפשרו כנראה טיפול נאות בפריטי מבחן מורכבים.

אחד המודלים שזכו להתעניינות מרובה הוא המודל של Rasch, וזאת בשל תכונותיו הסטטיסטיות והפשטות היחסית שבה ניתן ליישמו. פשטותו של המודל היא מקור כוחו וחולשתו כאחד. פשטותו של המודל נובעת מכך שהפונקציה האופיינית לפריט היא חד-פרמטרית. לפי מודל זה, פריטים נבדלים זה מזה רק ברמת הקושי ולא בדרגת האבחנה, שהיא קבועה, ואף לא בפרמטר הניחוש - אשר מניחים שהוא שווה לאפס. למרות שקיימים סוגי מבחנים שהמודל מתאים להם, הרי ספק אם הוא מתאים למבחנים רבי-ברירה (multiple-choice), שבהם, ביחוד אם מספר המסחים הוא קטן, יש הסתברות גבוהה יחסית שהנבחן ינחש את התשובה הנכונה. בבדיקה שערך Divgi (1986), נמצא שלמעלה ממחציתם של 1942 הפריטים שבדק לא התאימו למודל. הפריטים שבדק היו כולם מן התחום המילולי ונועדו לבדוק אוצר מילים והבנת קטעים בכיתות ד'-ו' של ביה"ס היסודי. בבדיקה שנערכה על בחינה שנועדה לבדוק כשרים בסיסיים (מבחן CTBS של McGraw-Hill), המורכבת משמונה מבחנים מילוליים ומתימטיים, הראתה Yen (1981) כי הפריטים אינם מתאימים לתורת התגובה לפריט, שהיא חד-פרמטרית. לגבי כל אחד משמונה המבחנים התאמת הנתונים למודל חד-פרמטרי היתה גרועה מהתאמתם למודל דו- או תלת-פרמטרי. מאחר והמבחנים הם כולם מטיפוס רב-ברירה, התאמתו של המודל התלת-פרמטרי (שבו נכלל, כזכור, גם פרמטר הניחוש) טובה יותר מהתאמתו של המודל הדו-פרמטרי. עם זאת, הראו Balla & McDonald (1985) כי ניתן להסביר בצורה עקבית את המקרים בהם פריטים אינם מתאימים למודל החד-פרמטרי. ניתוח שיטתי של הפריטים מאפשר לבנות מבחנים שיתאימו למודל החד-פרמטרי.

למרות כל הנאמר עד כה, הביקורת החזקה על המודל החד-פרמטרי, הנובעת הן משיקולים אפריוריים (Traub, 1983), הן מבדיקות אמפיריות, כפי שתוארו לעיל, אינה מונעת את יישומו במקרים קיצוניים. Lord (1983) הראה שעבור מדגמים קטנים יש סיבות טובות להעדיף את המודל החד-פרמטרי (אך ראה גם de Gruijter, 1986; van de Vijver, 1986).

בעשור האחרון הוחל בפיתוח אינדקסים לבדיקת מידת התאמתם של נתוני מבחנים למודלים של תורת התגובה לפריט (ר' למשל: McKinley & Mills, 1985; Yen, 1981). הבעיה העיקרית בתחום זה היא מציאת אינדקסים אשר מחד יצביעו על סטיות מן המודל, ומאידך לא יושפעו במידה רבה ע"י סטיות מזעריות ממנו. סיכומים עדכניים של תחום זה

(Rogers & Hattie, 1987; Hambleton & Murray, 1983) מצביעים על כך שעדיין לא נמצא אינדקס אופטימלי יחיד, ובבדיקות התאמה מסוג זה יהיה צורך להשתמש בשורה של אינדקסים.

### 3.2.2 מידיות המבחנים

כפי שהוזכר לעיל, על-פי התיאוריה העומדת בבסיסם של המבחנים הגמישים, המבחן חייב להיות חד-מימדי על-מנת שניבויי התיאוריה יתקיימו. מבלי להיכנס לשאלה של הגדרת מידיות (ראה דיון אצל McDonald, 1981) ניתן לקבוע בוודאות שרוב המבחנים הקונבנציונליים המצויים בשימוש אינם חד-מימדיים. הדבר נכון גם במידה והמידיות נקבעת בשיטות מטריות (בודסקו, 1985; Holland & Rosenbaum, 1985; Rosenbaum, 1984) וגם בשיטות נון-מטריות (בלר, 1982). מקרה פרטי של רב-מידיות הוא קיומם של אשכולות פריטים (item clusters), בהם הפריטים מתואמים זה עם זה יותר מאשר עם פריטים אחרים. המתאם נובע מסדר הופעתם של הפריטים במבחן, או מן הצורה בה הם מנוסחים. Kingston & Dorans (1984) הראו, למשל, כי שינוי סדר הצגתם של פריטים במבחן, מביא לשינויים דרסטיים באומדן פרמטרי הקושי שלהם. הדבר נכון לגבי פריטים שההוראות לפתרונם מסובכות ודורשות אימון. הועלתה אף הטענה (Traub, 1983) כי מבחני הישג, שמטרתם לבדוק ידע, הם מטבעם רב-מימדיים, שכן אימון וחינוך גורמים לדיפרנציאציה של כשרים (מאידד, Phillips & Mehrens, 1987, הראו שהבדלים בתוכניות הלימודים משפיעים בצורה מיזערית על המבנה הגורמי של מבחני הישג). כמו כן, נשמעת לעיתים הטענה שמבחני הבנת הנקרא הם בהכרח רב-מימדיים, שכן הם בנויים על רכישת מיומנויות ממקורות שונים (Cannale, 1986). קיימת אף טענה (Linn, 1986) שמבחנים דיאגנוסטיים מנוגדים לחלוטין לרעיון החד-מידיות, מאחר וכל מטרתם היא לגלות את הגורמים (המידים) הנוספים המתערבים במהלך הלמידה ומפריעים לה.

כיצד ניתן אפוא ליישם את המבחן הגמיש, אם הנחה מרכזית של תת"ל אינה עומדת במבחן המציאות? שתי תשובות לשאלה. הראשונה היא, שניתן להכליל את תת"ל גם למקרה הרב-מימדי (למשל, Sympson, 1978; Embretson, 1984; Whitely, 1980; Fischer, 1983; Goldstein, 1980; Jannarone, 1986; Samejima, 1974) אך מודלים כאלה עדיין לא הגיעו לשלב בו ניתן ליישם. התשובה השנייה היא, שניתן לנסח תנאים בהם אי-קיומה של הנחת חד-המידיות אינו מחליש את תוקפן של ההחלטות או הניבויים הנעשים על-פי תוצאות המבחן. לשם כך יש לדאוג שמאגר הפריטים ממנו נבנה המבחן יהיה חד-מימדי במידת האפשר (על דרכים לקביעת מידיות של מאגר פריטים ר' Hattie, 1985 וכן פרק 8 ב-Hambleton & Rovinelli, 1986; Hulin, Drasgow & Parsons, 1983; Drasgow & Lissak, 1983). במונחי ניתוח גורמים, יש לשאוף לכך שהגורם הדומיננטי יסביר את מירב השונות, ויש לסלק מן המאגר פריטים שטעינותם על הגורם הדומיננטי היא נמוכה. לחילופין, מאחר ופריטים רב-מימדיים נוטים להופיע כבעלי יכולת אבחנה נמוכה, סילוקם של אלה יביא במקרים רבים להקטנת משקלם של הגורמים הלא-דומיננטיים. Drasgow & Parsons (1983), בדקו את המידה בה הפרת הנחת חד-המידיות משפיעה על הדיוק באומדן היכולת של הנבחנים. מתקדם התבטס על

סימולציה ולא על נבחנים בפועל. מתברר, כי עד גבול מסוים ניתן לאמוד את היכולת בעזרת מאגר פריטים רב-מימדי. מובן שהפרת הנחת החד-מימדיות מפחיתה מדיוק האומדן, אך עבור דרגות נמוכות של רב-מימדיות, ההפחתה אינה משמעותית (אך ראה Ansley & Forsyth, 1985 על סייגים למסקנה זאת). עדות עקיפה נוספת ניתנת במחקרם של Dorans & Kingston (1985), אשר בניגוד למחקר Drasgow & Parsons, המבוסס על נתוני אמת. הם בדקו את השפעת דו-מימדיותו של המבחן המילולי בבחינת ה-GRE על היכולת לכייל אותו בשיטות המבוססות על תת"ל. מסקנתם היא, שלמימדיות הפריטים יש אומנם השפעה על תוצאות הכיול, וזאת בעיקר משום שבהינתן מאגר פריטים שאינו חד-מימדי, יש תת-הערכה באומדן פרמטר האבחנה של הפריטים, אך ההשפעה היא מיזערית וניתן לומר שכיול המבוסס על תת"ל הוא עמיד (robust) בפני הפרת הנחת החד-מימדיות.

למטרות יישום, ניתן לפתור את בעיית החד-מימדיות באמצעות חלוקה של מאגרי הפריטים הקיימים למאגרים חד-מימדיים ושימוש במספר מבחנים גמישים בצוותא על-מנת להגיע למדידת אותם כשרים הנמדדים על-ידי מבחן קונבנציונלי יחיד. הנבחן עצמו אינו חייב להיות מודע להפרדה בין מאגרי הפריטים שכן הפריטים מן המאגרים השונים יכולים להיות מוצגים לסירווגין. חוקרים העוסקים ביישום סבורים שמגבלה זו יכולה אף להצמיח טובה למבחני ההישג שכן דרישת החד-מימדיות מחייבת הנהרה של תחומי התוכן הנמדדים ומשרה סדר על המבחנים ועל מטרות ההוראה (Cannale, 1986). (Tung, 1986).

### 3.2.3 אומדן פרמטרים

מרבית סוגי המבחנים הגמישים מבוססים בצורה זו או אחרת על אומדנים של פרמטר או מספר פרמטרים לכל פויט. במבחנים המבוססים על תורות מורכבות, כמו למשל תורת התגובה לפריט בצורתה התלת-פרמטרית, בעיית אומדן הפרמטרים היא בעיה שבפתרונה מושקעים מאמצים רבים. כדי להמחיש את גודל הבעיה נדון במבחן רב-ברירה בן 100 פריטים המועבר לקבוצה של 1000 נבחנים. כל נבחן נותן 100 תגובות, ויש איפוא 100,000 נתונים בינאריים (תשובה נכונה או שגויה לכל פריט), שעליהם יש לבסס את אומדן הפרמטרים. אם ידועות רמות היכולת האמיתיות של הנבחנים כי אז יש לאמוד 3 פרמטרים לכל פריט ובסה"כ 300 פרמטרים. על פי רוב המצב אינו כזה, ויש לאמוד גם את 1000 רמות היכולת של הנבחנים - ובסך-הכל 1300 פרמטרים. למען הדיוק נציין כאן, שמאחר ובדרך כלל סולם היכולת הוא סולם-רווחים, ישנן שתי דרגות חופש שהן נקודת האפס של הסולם וגודל היחידה; ועל כן, בדוגמא הנדונה יש צורך לאמוד 1298 פרמטרים.

שיטות האומדן מבוססות בדרך כלל על קירובים בשלבים. אומדן בשיטת הניראות המירבית למשל, עשוי להיראות כך: בתחילת התהליך קובעים ערכים התחלתיים לכל הפרמטרים. אוסף הערכים שניתנו ל-1300 הפרמטרים מהווה מודל להתנהגות 1000 נבחנים בהינתן



להם 100 פריטים. ניתן לחשב לגבי מודל זה מהי ההסתברות (ניראות) לקבלת 100,000 התגובות שאמנם ניתנו בפועל (הנתונים). בשלב הבא של התהליך, ניתן לבדוק לגבי כל פרמטר איזה ערך חדש ניתן לתת לו, כך שהתאמת הנתונים למודל תהיה מירבית (ולכן שיטה זו נקראת שיטת הניראות המירבית). הערכים החדשים של כל הפרמטרים מחליפים את ערכיהם הישנים והתהליך חוזר חלילה. השיטה מביאה לתוצאות נאות אך היא דורשת זמן חישוב ארוך, ואינה מבטיחה פתרון אופטימלי בכל מקרה. יש עדויות לכך, ששימוש בנתונים נוספים לשם קביעת הערכים ההתחלתיים של הפרמטרים - כמו שיוך הנבחנים לקבוצות אוכלוסיה שונות - עשוי להביא להעלאת דיוק האומדנים (Mislevy, 1987).

השיטה שתוארה לעיל עומדת בבסיסה של תוכנית המחשב הידועה ביותר לצורך אומדן פרמטרים: LOGIST (Wingersky, Barton & Lord, 1982; Wingersky, 1983). קיימות שיטות אומדן נוספות (ר' סקירות אצל Hambleton & Swaminathan, 1985; Baker, 1987; Swaminathan, 1983; Tsutakawa, 1985) וכן שיטות מתקדמות יותר (אצל: Tsutakawa, 1985; Bock & Aitkin, 1981; Mislevy, 1986; Tsutakawa & Hsin, 1986). אך בכלן קיימת בעיה בסיסית של חוסר עיקביות של האומדנים; כאשר רמות היכולת האמיתיות ידועות, אזי אומדני הפרמטרים של הפריטים יתקרבו לערכיהם האמיתיים ככל שנגדיל את מספר הנבחנים; בדומה, כאשר הפרמטרים האמיתיים ידועים, אזי אומדני היכולת יתקרבו לערכיהם האמיתיים ככל שמספר הפריטים יגדל. בשני מקרים אלה נאמר שהאומדנים הם עקביים (consistent), אך כאשר מנסים לאמוד בו-זמנית את רמות היכולת ואת פרמטרי הפריטים גם יחד, האומדנים אינם עקביים. Ree (1981), למשל, הראה כי באומדני פרמטר האבחנה ישנה הטתיה (bias) שהולכת וגדלה ככל שגודל המדגם גדל; עם זאת, הראה כי הגדלת המדגם שעליו מתבצע אומדן הפרמטרים מביאה להגדלת הדיוק ולהקטנת ההטתיה באומדן של פרמטרי הקושי והניחוש. Lord (1980) אימץ את ההשערה, כי האומדנים הם עקביים רק כאשר מגדילים את מספר הנבחנים ואת מספר הפריטים גם יחד; Swaminathan & Gifford (1980) הראו שהשערה זו עומדת כנראה במבחן המציאות.

לבעיית עקביות האומדנים משמעות תיאורטית בעיקר, ואין לה השלכות מעשיות או יישומיות חמורות. מבחינה מעשית חשוב יותר לדעת את טעויות הדגימה של אומדני הפרמטרים. טעויות הדגימה של אומדני הפרמטרים נחקרו, וכן גם טעויות הדגימה המשותפות לפרמטרים השונים (Lord & Wingersky, 1985). ידיעה זו מאפשרת לעמוד על ההשפעה שיש להכללתם של פריטים שאומדניהם אינם יציבים, על יציבות האומדנים של פריטים אחרים הנכללים במבחן. במקביל, נעשתה גם עבודה על יציבות האומדן של עקומות התגובה לפריטים (Thissen & Wainer, 1985), המאפשרת לקבל מושג על מידת יציבות עקומת התגובה האופיינית לפריט בכל נקודה ונקודה.

#### 3.2.4 יעילות פסיכומטרית

עיקר התועלת הצפויה ממבחנים גמישים היא בהגדלת יעילותם הפסיכומטרית בהשוואה למבחנים קונבנציונליים. כבר הוזכר לעיל (סעיף 3.1), כי במבחנים גמישים ניתן להגיע לאותה רמת דיוק במדידה עם מספר פריטים קטן יותר. השאלה שתידון כאן היא,

האם ניתן לנצל עובדה זו להגדלת מהימנותם של המבחנים. על האדם המיישם את המבחן הממוחשב יהיה להחליט על יחס החליפין הנכון בין משך הבחינה לבין רמת הדיוק של המדידה המתקבלת ממנה.

הגדלת דיוק המדידה המושגת במבחן גמיש, תועלתה רבה במיוחד אם המבחן משמש למדידה על-פני סולם ציונים רציף (ולא לקביעת עבר/נכשל בלבד) ואם האוכלוסיה הנבחנת בו היא הטרוגנית מבחינת כישוריה. כל מבחני המשכל המקובלים, וכן רבים ממבחני ההישג הסטנדרטים שייכים לקטיגוריה זו והמשתמש בהם מניח - בצדק או שלא בצדק - כי דיוק המדידה ברמות המשכל הנמוכות דומה לדיוק המדידה ברמות הגבוהות, ובשני הטווחים הדיוק אינו נופל מזה שברמות הבינוניות. הנחה זו אינה נכונה בעליל כאשר מדובר במבחני משכל קבוצתיים. במבחנים רבי-ברירה, בהם מספר המסיחים אינו עולה על ארבעה או חמישה, ואשר בהם ניתנת לנבחן האפשרות לנחש, ניתן להראות שבתחום ציונים מסוים טעות המדידה הולכת וגדלה ככל שציונו של הנבחן נמוך יותר. באיזור הציונים הקרובים לרמת הניחוש (מספר הפריטים מחולק במספר ברירות התשובה), שונות המדידה היא ברובה שונות טעותית. סביר, אם כן, שמבחנים גמישים יפתרו את הבעיה על-ידי הצגת אותם פריטים המותאמים לרמתו של כל נבחן ונבחן.

השערה זו הועמדה במבחן חלקי ע"י Lord (1977, 1980), שהרכיב מבחן כושר מילולי המבוסס על מאגר של 363 פריטים המכסה את טווח הכושר החל בילדים בכיתה ד' ועד רמתם של מועמדים לתואר שני. טעות המדידה של המבחן עבור 13 רמות של כושר מילולי נאמדה ע"י הדמייה, כאשר בכל רמת כושר נערכה ההדמייה על 200 נבחנים מדומים, ולכל נבחן כזה נבדק מבחן בן 25 פריטים. נמצא כי טעות המדידה היא כמעט זהה בכל 13 רמות היכולת. כמו כן, נמצא כי רמתו של הפריט הראשון המוצג לנבחן אינה משפיעה רבות על דיוק המדידה, אם כי, כצפוי, רצוי להתחיל בפריטים קלים יחסית. מבחן גמיש זה הושווה למבחן קונבנציונלי באותו אורך, ונמצא כי בכל רמות היכולת, טעות המדידה של המבחן הקונבנציונלי גדולה לפחות פי שניים מטעות המדידה של המבחן הגמיש.

את יעילותו הפסיכומטרית של המבחן הגמיש הדגים Bejar (1978) בתנאים קרובים יותר למציאות. הוא הראה שניתן להשיג במבחן גמיש גם מדידה מדויקת יותר וגם קיצור במספר הפריטים. לצורך זה הישווה מבחן הישגים קונבנציונלי בביולוגיה שאורכו 25 פריטים, עם מבחן גמיש שאורכו הממוצע היה גם הוא 25 פריטים. טעות המדידה של המבחן הגמיש היתה לאין ערוך קטנה יותר, וכצפוי, יתרון זה היה ניכר יותר ברמות ההישגים הנמוכות. Bejar הראה שגם אם יקוצר המבחן הגמיש לאורך של 17 פריטים, עדיין יהיה מדויק יותר מן המבחן הקונבנציונלי.

בישראל נערך מחקר שבו נאמדה מהימנות המבחן החוזר של מבחנים גמישים לעומת מבחנים קונבנציונליים (אללוף, 1987). שיטת המחקר היתה גם היא בהדמייה, אלא שנתוני המבחנים היו אמיתיים; נשלפו נתוניהם של 454 נבחנים שנבחנו בשתי בחינות

פסיכומטריות של המרכז הארצי לבחינות ולהערכה בהפרש של שנה בין שתי הבחינות. מתוך תשובותיהם של הנבחנים נלקחו לצורך המחקר תשובותיהם במבחני "ידע-כללי" ו-"חשיבה מתמטית". מאחר ועבור כל נבחן היו נתונים על שתי בחינות, ניתן היה לחשב את מהימנות המבחן החוזר של כל קבוצת פריטים שנבחרה משתי הבחינות. כלומר, ניתן היה לחשב את ציונו של כל נבחן אילו היו מוצגים לו רק 5 פריטים או 10, 15 וכו'. בדומה, ניתן לחשב עבור כל נבחן מה היה ציונו אילו היה נבחן במבחן גמיש המורכב מקבוצות הפריטים שהופיעו במבחנים הקונבנציונליים, מאחר ותשובותיו לפריטים אלו ידועות. במחקר של אללוף הושו, איפוא, מהימנויות המבחן החוזר עבור מבחנים גמישים וקונבנציונליים באורכים שונים. נבדקו שני סוגים עיקריים של מבחן גמיש: מבחן המבוסס על תת"ל ומבחן הרמה הגמישה (Flexilevel). הממצא החד-משמעי מן המחקר הוא, שעבור כל אורך מבחן (ממבחן בו 5 פריטים דרך מבחנים בעלי 10, 15 פריטים וכו', ועד למבחן באורך מלא), מהימנות המבחן החוזר של מבחן גמיש המבוסס על תת"ל ואשר שיטת ציינונו היא בייסיאנית (להלן: המבחן הבייסיאני) היתה תמיד גבוהה יותר או שווה למהימנותו של המבחן הקונבנציונלי. לחילופין, ניתן לנסח ממצא זה גם כך: כדי להגיע לרמת מהימנותו של המבחן הקונבנציונלי נדרש מבחן גמיש שאורכו 83% - 78% מן המבחן המקורי. יש מקום להניח שבתוצאה זו יש משום תת-הערכה של יעילות המבחן הגמיש. זאת בשל העובדה שהמבחן הגמיש התבסס על מאגר פריטים קטן יחסית - כל אותם פריטים שהופיעו במבחן הקונבנציונלי. יתרונו הגדול של מבחן גמיש הוא בכך שעבור כל נבחן נבחרת קבוצת פריטים קטנה יחסית מתוך מאגר פריטים גדול. וכפי שהראה Ree (1981), אומדן היכולת של הנבחנים הוא מדויק יותר ככל שמאגר הפריטים גדול יותר (עד לגודל של 200-300 פריטים), גם אם ספר הפריטים המוצגים בפועל לנבחן הוא קבוע.

במחקרו של אללוף, הודגמה עדיפותו של המבחן הגמיש על-פני המבחן הקונבנציונלי, כאשר המדד ליעילות הוא מהימנות מבחן חוזר. עם זאת, הסתבר מן המחקר כי תוצאה זו אינה ניתנת להכללה עבור כל סוגי המבחנים הגמישים; מהימנותו של מבחן הרמה הגמישה למשל, לא עלתה בדרך-כלל על מהימנותו של מבחן קונבנציונלי באורך דומה.

### 3.2.5 תקפות

העלאת מהימנותו של מבחן עשויה לשפר את תקפותו; לכן, בכל מקרה אין לצפות לכך שתקפותם של מבחנים גמישים תהיה נמוכה יותר. מחקר תוקף כזה בוצע, למשל, על-ידי Thompson & Weiss (1980). הם הישוו שני טיפוסים של מבחן גמיש (מבחן גמיש בייסיאני ומבחן גמיש שכבתי) עם מבחן בעל מבנה קבוע. כל סוגי המבחנים הועברו על-ידי מחשב, כך שהשפעת טכנולוגית הבחינה נשמרה ברמה קבועה. המבחנים היו מבחני אוצר מילים ונמצא כי המבחן הגמיש השכבתי ניבא טוב יותר את הקריטריון, שהוא ממוצע הציונים בלימודיהם האקדמיים של הנבחנים. תוצאה זו הושגה למרות שבמבחן הגמיש מספר הפריטים שנדרש היה קטן ב-25%. גם המבחן הגמיש הבייסיאני ניבא טוב יותר את הקריטריון, אך יתרון זה לא הגיע למובהקות סטטיסטית מספקת.

Sympson & Moreno (1985) הציגו סיכום של עשרים ושניים מחקרים בהם נבדקה שאלת תקפותם של מבחנים גמישים. במחקרים שנסקרו נבדקו מבחנים גמישים מסוגים שונים: מבחנים פירמידליים, דו-שלביים, מבחני רמה גמישה, מבחנים שכבתיים ומבחנים גמישים המבוססים על תת"ל. בשנים-עשר מן המחקרים נבדקה תקפות הניבוי של מבחנים גמישים, ובעשרה מתוך אלו ניתן היה להשוות בין מבחנים גמישים לבין מבחנים קונבנציונליים. רק במחקר אחד מתוך העשרה היה יתרון למבחן הקונבנציונלי. מבין התשעה הנותרים בחמישה מקרים היה יתרון כלשהו למבחנים הגמישים, ובאלו שלא היה בהם יתרון למבחנים הגמישים ציינו החוקרים במפורש כי המבחן הגמיש היה קצר מן המבחן הקונבנציונלי.

### 3.2.6 תהליך העברת המבחן

בתהליך העברת מבחן ממוחשב מחליף המחשב את הבוחן האנושי. ניתן לומר בבטחון, שעבור כל רמת דיוק בביצוע שאליה מסוגל להגיע בוחן במשימות שגרתיות, אפשר למצוא פתרון טכנולוגי טוב יותר. אין פירוש הדבר שהמחשב יכול למלא את כל התפקידים המוטלים על בוחן, אך בוודאי שבתחומים מסוימים ניתן לבצע תפקידים אלו טוב יותר באמצעות מחשב.

את האחידות והקביעות במתן הוראות, שהן המושכל הראשון של כל בוחן, ניתן להביא לרמה גבוהה ביותר באמצעות מחשב. המחשב לא ידלג על פיסקה בהוראות ולא ישנה את צורת מתן ההוראות מבחינה לבחינה. כמו כן, במבחן ממוחשב ניתן לוודא שהנבחן מבין את ההוראות ופועל לפיהן בזמן הבחינה עצמה. ניתן להציג פריטים לדוגמא ופריטי אימון, ולבדוק - עוד בטרם החל המבחן התפעולי - שהנבחן עונה עליהם כנדרש. במידת הצורך ניתן לחזור על ההוראות, להרחיב ולפרטן, ולהציג פריטי אימון נוספים. אחידות בתנאי הבחינה מושגת גם בכך שאפיוניו הביוגרפיים של הבוחן כמו גיל, מין ומוצא, אינם באים לידי ביטוי. יש עדויות לכך (ראה סיכומים אצל Hansen, Hedl & O'neil, 1973; Johnson & Mihal, 1973) שלגורמים אלה יכולה להיות השפעה על רמת הביצוע במבחן. מובן שהשפעת גורמים אלה היא רבה יותר ככל שהמגע בין הבוחן לנבחן הוא הדוק יותר, כמו במבחני משכל יחידניים, אך היא עלולה לבוא לידי ביטוי גם במבחנים קבוצתיים. אין ספק ששילוק השונות בין הבוחנים יוצרת האחדה בתנאי הבחינה.

במידה והמבחן המועבר הוא קצוב בזמן, המחשב יקצוב את הזמן בדייקנות גבוהה יותר מכל בוחן, ויוכל להתגבר על המקרים בהם נבחנים מנסים להמשיך בבחינה למרות ההוראה להפסיק. מצד שני, אם המבחן הוא מבחן עוצמה שאינו קצוב בזמן, במבחן הממוחשב ניתנת לנבחן האפשרות לעבוד ככל שירצה מבלי להתחשב במגבלות זמן שסיבתן מינהלית בלבד. במלים אחרות, מבחן עוצמה המועבר באמצעות מחשב הוא מבחן עוצמה אמיתי; זמן סיום הבחינה תלוי רק בקצב התקדמותו של הנבחן ולא במהירות עבודתם של נבחנים אחרים או של "הנבחן הממוצע".

כפי שהוזכר לעיל, המחשב אינו יכול למלא את כל תפקידיו של הבוחן, אך הוא יכול לשחררו ממשיות שגרתיות, ולהבטיח שיהיו לו זמן והזדמנות להתגבר על בעיות מיוחדות. מבחן ממוחשב - אין פירושו שהבוחן אינו צריך לדעת מה עושה הנבחן; המחשב הוא כלי שבעזרתו יוכל הבוחן לעקוב אחר התקדמותם של כל הנבחנים (ולא רק הקרובים אליו) ולראות את תגובותיהם במשך כל זמן הבחינה ומבלי להפריע להם.

בעשור האחרון החלו מתרבים המחקרים בתחום "מדידת התקינות" (appropriateness measurement), בהם מנסים לאתר בשיטות סטטיסטיות שונות את הנבחנים שציונם אינו תקין. אי-תקינות יכולה לנבוע מסיבות שונות; נבחן יצירתי למשל, יכול להתעמק בקריאת פריט קל, לפרשו בצורה שבונה המבחן לא התכוון אליה, ולמצוא דרך להצדיק את נכונות אחד המסחים לפריט למרות שרוב רובם של הנבחנים האחרים מוצאים את הפריט כקל ומשיבים עליו נכונה. מקרה אחר של אי-תקינות הוא המקרה בו הנבחן שוגה בהעקת תשובותיו לגליון התשובות ומסיט את תשובותיו, וכך, למרות שידע את התשובה הנכונה לרוב הפריטים הרי שהרישום השגוי מביא לציון נמוך בחלק זה של המבחן בעוד שציונו על חלקים אחרים במבחן הוא גבוה בהרבה. אי-תקינות יכולה להיגרם גם במקרים של מעשי רמיה: העתקת חלק מן התשובות מנבחן אחר או היכרות מראש עם חלק מפריטי הבחינה.

בתחום מדידת תקינות פעילים במיוחד Levine ו-Drasgow המבססים את שיטותיהם על תורת התגובה לפריט (ראה סקירה אצל - Hulin et al., ch.4, 1983 וכן: Drasgow & Levine, 1986; Trabin & Weiss, 1983, Levine & Drasgow, 1982; Birenbaum, 1985). לפי ההתקדמות בתחום, נראה שבעתיד הלא-רחוק ניתן יהיה ליישם שיטות כגון אלה בזמן אמיתי, כלומר, במהלך הבחינה הממוחשבת. במקרה שתתגלה אי-תקינות בתגובות הנבחן, ניתן יהיה להעמיק את הבדיקה על-ידי הצגת פריטים נוספים או להעיר את תשומת ליבו של הבוחן לכך ולבקש את התערבותו.

### 3.2.7 השפעת שיטת המבחן ואופנותו על תוצאותיו

מבחנים ממוחשבים ומבחנים גמישים אמורים ברוב המקרים להחליף מבחנים קונבנציונליים. משום כך, שאלה הנשאלת רבות היא, מהי השפעת מיחשוב ו/או הגמשת המבחן לא רק על הדיוק במדידה, אלא גם על התוצאות שמשגים הנבחנים; האם יש למיחשוב בצורתיו השונות השפעות דיפרנציאליות לגבי קבוצות אוכלוסיה שונות והאם קיימים הבדלים אינדיבידואליים בהשפעה שיש למעבר ממבחן קונבנציונלי למבחן ממוחשב.

במחקרים רבים נמצא שהמעבר למבחן ממוחשב כרוך בירידת הציון הממוצע של הנבחנים. למשל, במחקר שנערך על למעלה מ-500 מתגייסים לצבא ארה"ב, נמצא כי הציון הממוצע במבחן ממוחשב (קשיח) היה נמוך בכחמישית סטית תקן נמוך ממבחן זהה שהועבר בשיטה קונבנציונלית (Lee, Moreno & Sympson, 1986). זאת למרות העובדה שבשני אופני הבחינה לא הייתה מגבלת זמן ולמרות שנאמר לנבחנים להשיב על כל פריטי המבחן. לא

נמצא גם שום אפקט ליכולתם של הנבחנים על ההבדל בביצוע בשני אופני הבחינה, כלומר, יתרונו של המבחן הקונבנציונלי לא השתנה באופן עיקבי עבור רמות יכולת שונות. לעומת זאת קיימים גם ממצאים הפוכים; למשל, במחקר שהשווה את ביצועיהם של נערים שחורים ולבנים בשני סוגי מבחנים - קונבנציונלי וממוחשב (קשיח), נמצא כי שתי הקבוצות משפרות מעט את הישגיהן במבחן הממוחשב (אפקט לא מובהק) אך השיפור הגדול יותר הוא בהישגי קבוצת השחורים (Johnson & Mihal, 1973).

Lee (1986) הניחה שהירידה בביצוע במעבר ממבחן קונבנציונלי למבחן ממוחשב מקורה בחוסר הנסיון של הנבחנים בעבודה עם מחשב. לצורך בדיקת השערה זו בדקה באמצעות שאלון את נסיון העבודה במחשב של 92 סטודנטים, ובמקביל בדקה את ביצועיהם בשני מבחני יכולת אריתמטית: קונבנציונלי וממוחשב. את הנבדקים חילקה לשלוש קבוצות עפ"י נסיונם בעבודה עם מחשב ובדקה את ההבדלים בציוני שלוש הקבוצות על המבחן הממוחשב כאשר יכולתם האריתמטית של הנבחנים, כפי שנמדדה במבחן הקונבנציונלי, משמשת כמשתנה בקרה. נמצא הבדל מובהק בין ביצועיהם של בעלי הנסיון המועט ביותר לבין ביצועיהם של שאר הנבחנים במבחן הממוחשב. מסקנתה של Lee היא, כי מידת הנסיון של נבחנים בעבודה עם מחשב עשויה להשפיע על ביצועיהם. עם זאת יש לציין, כי קבוצת הנסיון המועט נבדלה מובהק מן הקבוצות האחרות גם במשתנה הבקרה ועל-כן יש להיזהר בהסקת מסקנות פסקניות מן המחקר, למרות שהניתוח הסטטיסטי אמור לטפל בבעיה זו. יש מקום להוסיף כי בקבוצת בעלי הנסיון המועט נכללו רק אותם נבחנים שנסיונם כלל אך ורק משחקי מחשב. נבחנים אשר כתבו לפחות מכתב אחד בעזרת מעבד תמלילים כבר נכללו בקבוצת בעלי הנסיון הבינוני, ולא נמצא כלל הבדל בכיוון הצפוי בין קבוצה זו ובין קבוצת בעלי הנסיון הרב. פירושו של דבר הוא, איפוא, שדי בנסיון זעום בעבודה עם מחשב כדי להתגבר על מגבלת חוסר הנסיון במעבר למבחן ממוחשב. למסקנה דומה הגיעו Johnson & White (1980) בניסוי שבו הראו כי אימון של כשעה בשימוש במחשב די בו כדי לשפר את ביצועיהם של אנשים מבוגרים במבחן משכל ממוחשב. יש להתייחס בזהירות לניסוי זה, למרות מובהקותו של האפקט, משום שלא הופעל משתנה בקרה: מדידה של המשכל באמצעים שווים בשתי קבוצות הניסוי.

מקור אחר להבדלים בביצוע יכול להיות המעבר ממבחן שבו יש אפשרות לחזור ולתקן תשובות, למצב שבו מוצג פריט אחד בכל פעם ואין אפשרות לשוב אליו לאחר מתן התשובה. יש עדויות לכך שהאפשרות לשוב אחורה ולתקן פריטים אכן מנוצלת על-ידי הנבחנים לטובתם (Sachar & Fletcher, 1978) והיא תלויה בין היתר גם בסוג הפריט (Hoffman & Lundberg, 1976); עם זאת, לא נמצא גורם שיסביר את ההבדלים האינדיבידואליים במידת השימוש שעושים נבחנים באפשרות זו.

שאלה מורכבת יותר היא, האם שינוי האופנות - ממבחן קונבנציונלי למבחן ממוחשב - עלול להביא לשינוי במהות המדידה. במלים אחרות: האם במבחן ממוחשב נמדדים דברים

שונים מאשר במבחן קונבנציונלי. תשובה אחת לכך ניתן למצוא אצל Moreno ועמיתיו (1984); הם הקרו את השפעת מיחשובו של מבחן הכושר של הצבא האמריקאי, ובין היתר, בדקו האם המבנה הגורמי של הציונים המופקים מן המבחן הממוחשב שונה מזה של המבחן הקונבנציונלי, והגיעו למסקנה שהכשרים הנמדדים בשתי צורות הבחינה הם זהים. אין איפוא שינוי איכותי של המדידה במעבר למבחן ממוחשב.

Pine & Weiss (1978), בדקו אם קיימים הבדלים במידת ההטייה כנגד קבוצות מיעוט חלשות במבחן גמיש לעומת מבחן קונבנציונלי. את בדיקתם ערכו על נתוני הדמיה לגבי 30 צירופים שונים של סוגי מבחנים, סוגי מאגרי פריטים, מידת ההבדל ביכולת בין קבוצת הרוב לבין קבוצת המיעוט, ואורכי המבחנים. אין זה פשוט להכליל את תוצאות הניסוי, אך ניתן לומר בפסקנות שסוג אחד של מבחנים גמישים (מבחן גמיש בייסיאני), הוא בהחלט עדיף על סוגים אחרים ועל מבחנים קשיחים; מידת ההטייה שיש בו כנגד קבוצת מיעוט היא פחותה, ועדיפות זו הולכת וגדלה ככל שאורכו של המבחן וטיב מאגר הפריטים עולים.

השימוש במחשב מעניק את האפשרות לתת משוב מיידי לנבחן על נכונות תשובותיו לפריטי המבחן. המשוב יכול להינתן לאחר כל תשובה או מיד עם סיום המבחן. Weiss ועמיתיו מאוניברסיטת מינסוטה הקרו את השפעת המשוב על ביצוע הנבחנים (Prestwood & Weiss, 1978; Betz & Weiss, 1976a, 1976b, Prestwood, 1978) ומצאו כי מתן המשוב מעלה את המוטיבציה של הנבחנים ועקב כך, לעיתים, גם את ביצועיהם. במחקריהם של Betz & Weiss נמצא גם כי מידת השיפור בביצועים היתה גדולה יותר ככל שמידת המשוב החיובי היתה רבה יותר. זוהי כנראה הסיבה לכך שהם מצאו במחקריהם כי נבחנים בעלי רמת יכולת נמוכה הראו שיפור בביצועיהם במבחן גמיש לעומת מבחן קשיח (כאשר שני סוגי המבחנים מועברים באמצעות מחשב), משום שבמבחן גמיש קושי הפריטים מותאם לרמת היכולת של הנבחן ועל-כן גם אלה שרמת יכולתם היא נמוכה, נבחנים למעשה במבחן שרמת הקושי שלו עבורם היא בינונית. ממצא זה מקבל תמיכה ממחקר שהשווה את השפעתו של מבחן גמיש ושל משוב לנבחנים על ביצועיהם של שחורים ולבנים במבחן יכולת מילולית (Pine, Church, Gialluca & Weiss, 1979). תוצאות המחקר מורכבות ולא תמיד הן מתיישבות עם ממצאים קודמים, אולם אחת המסקנות הפסקניות ממנו היא, כי המבחן הגמיש, בשל התאמתו לרמת היכולת האינדיבידואלית, מעודד את בעלי היכולת הנמוכה.

לסיכום אפשר לומר ששתי ההשפעות העיקריות של מיתשוב והגמשת מבחנים הן:  
א. מניעת האפשרות לחזור ולתקן תשובות לפריטים קודמים - ועל-כן ירידה בביצוע במבחן ממוחשב קשיח.

ב. מתן אפשרות לבחון לספק משוב לנבחנים ועל-ידי-כך להעלות את רמת המוטיבציה שלהם; הדבר נכון במיוחד במבחנים גמישים, בהם ניתן לנבחנים חלשים משוב חיובי רב יותר מאשר במבחנים קשיחים.

שתי ההשפעות מנוגדות זו לזו ובוודאי קיימים גורמים נוספים, שהשפעתם עדיין לא הובהרה, אשר יביאו לשינוי בביצוע במעבר ממבחן קונבנציונלי למבחן ממוחשב וממבחן קשיח למבחן גמיש. מסקנה אופרטיבית אחת היא ברורה למדי: לא ניתן להסתמך על נורמות ועל ניתוחי פריטים שהושגו בשיטה אחת, כאשר מיישמים שיטת בחינה או אופנות בחינה חדשה. Greaud & Green (1986), למשל, הראו כי שינוי מזערי בצורת הפריט כאשר הוא מועבר באמצעות מחשב יכול להשפיע בצורה קיצונית על ביצועיהם של הנבחנים, ובעיקר אמור הדבר במבחנים ביצועיים, בהם המטלה פשוטה יחסית ועל-כן איכות ומהות התגובה של הנבחנים תלויה במידה רבה באמצעי התגובה. תנאי מוקדם למיחשוב או הגמשת מבחן, אם כך, אינו רק פיתוח טכנולוגי, כי אם גם עבודת הכנה נרחבת באיסוף פרמטרים ונורמות של פריטים ומבחנים.

### 3.2.8 סוגי מבחנים חדשים והעשרת מדדי תגובה

שימוש במחשב, בין אם במבחן גמיש ובין אם לא, פותח פתח ליישום סוגים חדשים של מבחנים ופריטים, או שימוש מחודש במבחנים ישנים. כך, למשל, בתוכניות בחינה המוניות כמו אלו המועברות על-ידי ה-ETS בארה"ב או על-ידי המרכז הארצי לבחינות בישראל, משתמשים רק בפריטים רבי-ברירה. פריטים בהם הנבחן לא מסמן את התשובה הנכונה, אלא נותן תשובה מלאה לשאלה, אינם ישימים בבחינות מסוג זה, בשל המאמץ והזמן הרב הדרושים להערכת הבחינות וניתוחן הסטטיסטי.

המחשב מאפשר שימוש בפריטים פתוחים (free-response items), וזאת בתנאי שבדיקת התשובה אינה מחייבת תחום רב מדי. פריטים חשבוניים פשוטים או פריטים מילוליים כמו אלו המופיעים במבחני אוצר-מילים או מציאת הפכים (antonyms) הם מועמדים טבעיים להערכה וציינון ע"י תוכנית מחשב. Vale (1977; Vale & Weiss, 1978) הראה שניתן ליישם שיטה זו במבחן אוצר-מילים ממוחשב, והתוצאות שמתקבלות עולות במקרים מסוימים על השימוש בפריטים סגורים, זאת משום שמרחב התגובות האפשרי של המבחנים אינו מוגבל לארבע או חמש אפשרויות התגובה המופיעות במבחן סגור. ריבוי התגובות מאפשר לנצל באורח יעיל גם את התגובות הנכונות למחצה לשם אמידת היכולת של הנבחנים.

קיימים גם נסיונות התחלתיים לנצל את כוחו של המחשב להערכה וציינון של מטלות מורכבות כמו מטלות כתיבה קצרות (Reid, 1986) או קריאה בקול של קטע המוצג לנבחן בכתב. בניסוי מעבדתי של הערכת איכות הקריאה בקול נתקבל מתאם סודר של 0.93 בין הערכות המחשב לבין הערכותיו של מעריך מאומן לגבי 20 דוגמאות קריאה (Molholt & Presler; 1986). זוהי התחלה צנועה אך מבטיחה. אם ניתן יהיה לשמור על הישג כזה במעבר מן המעבדה אל השדה תהיה בכך משום פריצת דרך.



סוגים חדשים לגמרי של פריטים, שעד כה שימשו רק בפרדיגמות ניסוייות במעבדותיהם של פסיכולוגים קוגניטיביים, ניתנים גם הם למיחשוב. כך למשל, בעזרת מחשב, ניתן לחזור ולהשתמש במבחני זכרון לטווח קצר, שעד כה היו בשימוש רק במבחנים פרטניים כמו מבחן ווקסלר לילדים. סוגי פריטים אחרים שנוסו במבחנים ממוחשבים (Cory, 1978; Cory, Rimland & Bryson, 1977) הם בתחום סריקת זכרון (Sternberg, 1966) ועירנות תפסיתית (vigilance; ראה למשל: Holland, 1958) ולמידת מושגים (identifying concepts). התחום של למידת מושגים הוא מעניין במיוחד בחקש הנוכחי, שכן למרות שאין הוא מחייב שימוש באמצעי קלט ופלט שקשה לחקותם במבחני נייר ועפרון, הוא מחייב משוב מיידי לנבחן בתום כל צעד. במבחן זה, שמקורו במחקריהם של Bruner ועמיתיו (Bruner, Goodnow & Austin, 1956), מוצגים לנבחן שני אובייקטים ועליו לציין אם הם שייכים לאותה קטגוריה או לא. מובן שבצעד הראשון הנבחן אינו יודע את התשובה הנכונה ועליו לנחש, שכן הקטגוריה הוגדרה שרירותית על-ידי מפתח המבחן. לאחר שניחש, מקבל הנבחן את התשובה הנכונה ומוצג לו זוג אובייקטים נוסף. הנבחן ממשיך להגיב ולקבל משוב עד שהוא מצליח לגלות את המשותף לכל האובייקטים הנכללים באותה קטגוריה, או במילים אחרות - עד שהוא מזהה ולומד את המושג המגדיר את הקטגוריה.

לפחות לגבי מקצוע אחד (טכנאי סונאר), נמצא שמבחנים מסוג זה, שנכנה אותם לצורך הדיון כאן בשם "מבחנים קוגניטיביים", מיטיבים לנבא את הצלחת הנבחנים בביצוע עתידי מאשר מבחני נייר ועפרון מקובלים (Cory, 1977, 1978).

הספרות על הבדלים אינדיבידואלים במבחנים קוגניטיביים היא עניפה (ראה סקירות אצל Sternberg, 1982; Estes, 1982; Cooper & Regan, 1982, וכן אסופות מאמרים אצל Dillon & Schmek, 1983, ו- Dillon, 1985) ואין ספק כי ברבים מן המבחנים הקוגניטיביים נמצא הביצוע במתאם חיובי עם רמת-משכל ועם הצלחה בביצוע עתידי. כך, למשל, מצאו Kahneman ועמיתיו (Kahneman, Ben-Ishai & Lotan, 1973); Gopher & Kahneman, 1971) שהצלחה במשימות הדורשות חלוקת קשב עשויה לנבא את טיב הביצוע של טייסים ונהגי אוטובוסים.

אך שתי בעיות עקרוניות ניצבות בפני החוקרים המבקשים ליישם את המבחנים הקוגניטיביים כמבחנים כלליים ולא רק בתחום מקצועי צר. הבעיה הראשונה היא איסוף הוכחות אמפיריות בדבר תקפותם הרבה יותר של מבחנים קוגניטיביים בהשוואה למבחנים המקובלים, כלומר, הוכחה שמבחנים קוגניטיביים מאפשרים למדוד אוסף של כשרים שהם רלבנטיים לביצוע במגוון רחב של עיסוקים, ואינם מתמצים במבחנים המקובלים כמו מבחני כושר מילולי או מתמטי. למען הדיוק ההיסטורי, נציין שבתקופה המוקדמת של תקר האינטליגנציה כוונו המבחנים למדידת כשרים תחושתיים ומוטוריים אלא שתועלתם היתה מועטת (Berger, 1982).

הבעיה השניה היא הפיכת המבחנים הקוגניטיביים מכלי מחקר מעבדתי למבחן פסיכומטרי, על כל הכרוך בכך. משימות קוגניטיביות פשוטות לכאורה, ניתנות למדידה באופנים רבים. לעומת התגובה הפשוטה של נבחן לפריט במבחן רב-בריחה, הרי שבמבחנים הקוגניטיביים ישנה אפשרות למדידת זמני תגובה, אסטרטגיות פתרון, מדדי רגישות ואבחנה, וטעויות מסוגים שונים. על הפסיכומטריקאי לברר אילו מן המדדים הם רלבנטיים לניבוי, ומהי הדרך הטובה ביותר למדוד אותם. עליו גם לבנות כלים חדשים המקבילים לכלים הקלאסיים של ניתוח פריטים, ולנסח תורות מבחנים חדשות אשר ישכילו לטפל במבחנים קוגניטיביים. דוגמה למידת המאמץ הנדרש ניתן למצוא בעבודתם של Church & Weiss (1980), אשר טרחו על ניתוחה של משימה פשוטה יחסית - "חידת ה-15" - בה מתבקש הנבחן להביא מערך נתון של 15 ריבועים ממוספרים למצב מטרה מסויים, כאשר בכל צעד ניתן להזיז רק ריבוע אחד. הצגת ההוראות, הצגת הבעיות ותיקון טעויות הנבחנים במהלך פתרון הבעיה, בוצעו בעזרת מחשב. בין היתר ניתחו החוקרים שבעה מדדים לקושי הפריטים, המבוססים על הנתונים שנאספו בניסוי, ובנוסף להם עוד שישה מדדי קושי שחושבו על סמך המאפיינים של הבעיות עצמן - כמו המרחק בין המצב ההתחלתי ובין מצב המטרה של מערך הריבועים. זוהי רק דוגמה למידת הסיבוך הכרוכה בהפיכת מבחן קוגניטיבי לכלי פסיכומטרי.

הקירה ניסויית של מבחנים קוגניטיביים מלווה בפיתוח של מודלים פסיכולוגיים; מודלים המתארים את הפעולות המנטליות הכרוכות בהבנת הבעיה ובפתרונה. מבחינה זו, יש עדיפות למבחנים קוגניטיביים על-פני מבחנים מקובלים, שפריטיהם נבחרים לאחר מעשה רבים מן המקרים, בזכות תכונותיהם הסטטיסטיות, וללא רקע תיאורטי המסביר את תגובותיהם של הנבחנים (Egan, 1979). אחרי הכל, אחת ממטרותיה של הפסיכולוגיה הקוגניטיבית היא להנהיר את המנגנונים הקוגניטיביים העומדים בבסיס האינטליגנציה. נסיון להגיע למטרה זו נעשה, למשל, על-ידי Sternberg (1979), אשר ניסח מודל לכשרים מנטליים המבוסס על התיאוריה והידע שנצבר בתחום עיבוד מידע אנושי (human information processing). דוגמה לכוחה של התיאוריה בפיתוח מבחן ניתן למצוא במחקרם של Butterfield, Nielsen, Tangen & Richardson (1985) שחקרו תשיבה אינדוקטיבית. הם הראו שניתן לנבא ברמת דיוק גבוהה את רמת הקושי של פריטים על-סמך מודל פסיכולוגי. מאחר ואין צורך בבדיקה אמפירית של הפריט לפני השימוש בו, ניתן ליצר את הפריטים בעזרת תוכנית מחשב בכל רמת קושי דרושה.

אחד המדדים החשובים ביותר בפסיכולוגיה קוגניטיבית הוא החביון (latency) או זמן התגובה (reaction time). ברור שזמן תגובה אינו ניתן למדידה במבחני נייר ועפרון קבוצתיים, בעוד שמדידתו באמצעות מחשב היא מיידית ופשוטה. דוגמה ליישום ניתן למצוא במחקרם של Greud & Green (1986), בו הושוו מבחני מהירות שהועברו בשתי אופנויות: מבחן נייר ועפרון מול מבחן ממוחשב. בעוד שבמבחן נייר ועפרון נקצב זמן קבוע ונמדד מספר התגובות הנכונות ביחידת זמן קבועה לכל הנבחנים, הרי שבמבחן הממוחשב ניתן לחשב את זמן התגובה הממוצע או את מספר הפריטים ליחידת זמן, כאשר

מספר הפריטים מוחזק כקבוע. החוקרים מצאו כי מדד התגובה שניתן להפיק מן המבחן הממוחשב עולה באיכותו באופן ניכר על המדד המופק במבחן הקונבנציונלי.

בקונטקסט של פסיכולוגיה קוגניטיבית מוצאים בדרך-כלל יחס חליפין בין המהירות לבין הדיוק בביצוע משימה כלשהי. יחס חליפין זה נמצא עוד במאה שעברה, הוא מתועד עבור רוב המשימות הקוגניטיביות (ראה למשל Wickelgren, 1977; Pew, 1969; Fitts, 1966) ואף עמד לנגד עיניו של Thurstone (1937) בהגדירו את מושג היכולת. אם אמנם יחס החליפין הוא קבוע עבור כל האוכלוסיה, כי אז אין מקום למדידת דיוק ומהירות גם יחד. אך טיבו של יחס החליפין הוא כנראה מורכב יותר. על-פי מחקריהם של Hunt ועמיתיו (Hunt, Lunnenborg & Lewis, 1975) וכן של Sternberg (1977), יש קשר בין רמת משכל לבין מהירות עיבוד מידע. ממצאים כגון אלה הובילו אף את Eysenck (1982) לנסות ולהעמיד את מושג האינטליגנציה על בסיס פיזיולוגי של מהירות העברתו המדויקת של מידע בתוך האורגניזם. Sternberg בדק לעומק את מרכיביה הקוגניטיביים של משימת פתרון אנאלוגיות מילוליות, ומדד את הזמן שהקדישו הנבדקים לכל שלב בפתרון הבעיה. הוא מצא כצפוי, כי נבדקים מוכשרים יותר מגיעים לפתרון הבעיה מהר יותר, אך מהירותם אינה זהה בכל שלבי הפתרון; דווקא הנבדקים המוכשרים יותר הקדישו יותר זמן לשלבים הראשונים של פתרון הבעיה מאשר נבדקים מוכשרים פחות. בזכות זה, לנראה, הגיעו לפתרון מהיר יותר של הבעיה כולה. Hunt ועמיתיו מצאו כי כושר מילולי עומד במתאם עם המהירות שבה נבדקים עוברים מייצוג פיזי של גירוי למצב בו הם יודעים את משמעותו; למשל, המהירות בה נבדקים מזהים תבנית חזותית, מילה, או אות.

המגמות השיטיות שנתגלו במחקרים אלה עומדות בניגוד לחוסר הממצאים, או לממצאים הסותרים, של מחקרים אשר ניסו למצוא קשר בין מהירות הביצוע לבין הישגי הנבחנים במבחני הישג קונבנציונליים (ראה למשל Bridges, 1985), מחקרים אשר לא נעשה בהם נסיון לנתח את הביצוע למרכיביו, ואשר בבסיסם לא עמדה תיאוריה שהדריכה וכיוונה את המחקר.

הכלים הפסיכומטריים להתמודדות עם משימות קוגניטיביות מצויים כיום בשלב ראשוני של פיתוחם. הוצעו אמנם מודלים של תכונה חבויה עבור מרכיבים קוגניטיביים של פתרון בעיות (Whitely, 1980; Fischer 1973; Embretson, 1985b) וכן עבור גורם הזמן ומדידתו בפתרון בעיות (Thissen, 1983; Roskam, in press; White, 1973), אולם אף לא אחד מכלים אלה אינו מאפשר עדיין יישום מיידי של מבחנים קוגניטיביים. עם זאת, נראה ששילובה של הפסיכולוגיה הקוגניטיבית עם הפסיכולוגיה הדיפרנציאלית עשוי להיות פורה יותר מן השילוב שהציע Cgonbach (1957), ושילוב זה אפשרי כנראה רק במסגרת מבחנים ממוחשבים. מגמה שלמה של מחקר צמחה משילוב זה בשנים האחרונות (Eysenck, 1982; Embretson, 1985a) והיא נראית מבטיחה: אם תישא פרי, יהיה בכוחה לתת לפסיכומטריקאי מערכת שלמה של שיקולים תכניים בבניית מבחנים ותכנונם, מעבר לשיקולים המסורתיים של תוקף ניבוי.

### 3.3 תגובות הנבחנים והציבור

יחסם של הנבחנים ושל הציבור כולו עשוי להשפיע במידה רבה על אופן השימוש במבחנים הממוחשבים ועל כמות המשאבים שיוקצו לפיתוחם. בדומה להשפעה שהיתה לדעת הקהל על התקינה והחקיקה בנושא המבחנים הקונבנציונליים הן בארצות הברית (Bersoff, 1981); (Haney, 1981), הן בישראל (באשי, 1985). חשוב איפוא לברר, מהן תגובות הנבחנים למבחנים ממוחשבים ואילו מן הביקורות המופנות כלפי מבחנים קונבנציונליים עשויות להתחדד כאשר הן מופנות כלפי מבחנים ממוחשבים.

#### 3.3.1 הנבחנים

כפי שכבר הוזכר, אחד היתרונות של מבחנים ממוחשבים, בין אם גמישים ובין אם לאו, הוא ביכולתם לספק משוב לנבחן תוך כדי הבחינה. בניסוי שערכו Betz & Weiss (1976b), נבדקו תגובותיהם של סטודנטים אשר נבחנו במבחן אוצר מילים ממוחשב. לחלק מן הנבחנים ניתן משוב תוך כדי הבחינה, ונבחנים אלו נישאלו לאתר המבחן על עמדתם כלפי המשוב. רובם הגדול של הנבחנים (90%) אמרו שעמדתם היא חיובית ביותר. רובם גם העריכו שקבלת המשוב לא הפריעה לריכוזם המנטלי בעת הבחינה וציינו שקבלת המשוב הפכה את סיטואצית הבחינה למעניינת יותר. תוצאות דומות דווחו גם על-ידי Prestwood & Weiss (1978), אשר הראו, בנוסף, שסוג המבחן (גמיש או לא) ומידת הקושי של המבחן אינם משפיעים על מידת שביעות הרצון מן המשוב. מידת שביעות הרצון לפי מחקר זה, עומדת במתאם חיובי עם מידת המוטיבציה של הנבחנים ובמתאם שלילי עם מידת החרדה שלהם בזמן המבחן.

ידיעת תוצאות המבחן סמוך לסיומו היא גם הסיבה העיקרית להעדפת מבחנים ממוחשבים על-פני מבחני נייר ועפרון. כך נמצא במחקר של Schmidt, Urry & (1978, Gugel), אשר בדקו את תגובותיהם של נבחנים-מתנדבים למבחן גמיש ממוחשב הבודק כושר מילולי. יתרון אחר של המבחן הממוחשב שציינו הנבדקים במחקר זה, היה קיצור משך המבחן הממוחשב, שהוא פועל יוצא של מיחשובו. באופן כללי נמצא במחקר כי 83% מן הנבחנים העדיפו את המבחן הממוחשב; רק 10% העדיפו מבחן נייר ועפרון ו-7% היו אדישים לשיטת העברת המבחן. החסרון העיקרי של המבחן הממוחשב, לדעת הנבחנים, הוא חוסר האפשרות לשוב ולעיין בפריטי מבחן שכבר הוצגו על מנת לשקול שינוי התגובה. בנוסף לכך ציינו מספר נבחנים (20%) שכדאי לשפר את קריאותו של המבחן אשר הוקרן על גבי צג מחשב. בעיה זו, שהיא טכנית מעיקרה, הוזרה ומופיעה גם ממקורות אחרים. כך למשל נמצא שנבחנים מעדיפים שהתצוגה על-פני מסך המחשב תהיה בשחור על לבן, בדומה לחומר מודפס, ולא בלבן על שחור - כמקובל בצגי מחשב (Koch & Patience, 1978).

Green (1984) סקר את המחקרים שנעשו במסגרת פרויקט מיחשוב מבחן הכושר הכללי של צבא ארצות-הברית, והגיע למסקנה שתגובת הרוב הגדול של הנבחנים למבחן ממוחשב היא חיובית. גם כאן נמצא כי טענתם העיקרית של הנבחנים כלפי מבחן ממוחשב היא בדבר

חוסר יכולתם לחזור ולתקן פריטים שכבר ענו עליהם. במחקרים אלה גם לא נמצאה השפעה כלשהי של מידת הנסיון בעבודה עם מחשב על עמדות הנבחנים, מלבד ממצא אחד, והוא, שבעלי נסיון מועט הביעו יותר הסכמה עם הטענה שבבחינה באמצעות מחשב יש פחות יחס אישי לנבחן. יחד עם זאת, רוב הנבחנים מרגישים שבבחינה באמצעות מחשב יש יותר יחס אישי, ובלשונו של Green:

"After all, the computer console does react to test responses, which is more than can be said of an answer sheet".

כאשר נשאלו הנבחנים, במסגרת של שיחה חופשית לאחר המבחן, מהי עמדתם לגבי, התייחסו רובם לתוכן המבחן - סוגי הפריטים ורמת הקושי שלהם - ולא דווקא לצורת הצגתו.

תגובת הנבחנים למבחנים ממוחשבים היא, איפוא, חיובית ביותר; זהו הרושם הכללי המתקבל מן המחקרים שנסקרו לעיל, והוא מקבל תמיכה במחקרים נוספים (אספורמס, 1978; Garrison & Baumgarten, 1986; Johnson & Mihal, 1983); אך יש להזהיר מקבלת הממצא ללא סייגים, לפחות משני טעמים. הטעם הראשון הוא שבמחקרים אלה העמידו החוקרים את המבחן הממוחשב (ולא הקונבנציונלי) כמושא המחקר. לא היתה סימטריה בין שני אופני הבחינה, ואם הנבחנים ניחשו או קלטו נכונה את עמדות הנסיינים, אפשר שהטיות תגובה היטו את תוצאות המחקרים. הטעם השני הוא, שהעדפת המבחן הממוחשב נובעת גם מן החידוש שבו ולא רק מיתרונות אמיתיים שהוא מעניק לנבחן. בעוד שאת בעיית הטיית התגובה ניתן לפתור על ידי מערך מחקר מתאים, הרי ששאלת "ההנאה שבחידוש" תיפתר רק עם חלוף הזמן, כאשר יפוג החידוש.

על החסרון העיקרי שמוצאים הנבחנים במבחן ממוחשב - אי-היכולת לחזור, לעיין ולשנות את התשובה לפריטים קודמים - לא ניתן להתגבר בקלות. יש אמנם מערכות בחינה ממוחשבות מעטות שבהן הדבר אפשרי (Sachar & Fletcher, 1978), אך ע"פ רוב, ובייחוד במבחן גמיש, הדבר קשה או שאינו בר-ביצוע. דרוש מחקר נוסף על-מנת לקבוע אם הבדל זה בין מבחן ממוחשב למבחן נייר-ועפרון משפיע דיפרנציאלית על נבחנים שונים, או שהוא מתבטא רק בהפרש קבוע בציוניהם של הנבחנים במבחנים ממוחשבים. יש הטוענים (Schmidt et al., 1987) כי ניתן יהיה להתגבר על מכשול זה בהסברה נאותה לנבחנים. אחרים (Weiss, 1978) מוצאים את הבעיה כבלתי רלבנטית לחקר מבחנים ממוחשבים.

### 3.3.2 הציבור

אין עדיין מידע בדוק על תגובות הציבור למבחנים ממוחשבים. אך ניתן לזהות מוקדי ביקורת פוטנציאליים. מוקד הביקורת הראשון מקורו בדיעה הרווחת בציבור, שהמבחן הממוחשב מפלה לטובה את בעלי הנסיון בהפעלת מחשב. יהיה צורך בהדגמה מדעית משכנעת על-מנת להפריך טענות אלו. אופיינית היא אולי העובדה, שנבחנים שעברו מבחן ממוחשב מציינים שלהם עצמם לא היתה בעיה טכנית בהפעלת המחשב אך כאשר הם

נשאלים לגבי בעיות אפשריות הם טוענים שלנבחנים אחרים, בעלי רמת השכלה ותיכום נמוכים יותר, וודאי יהיו בעיות בתחום זה (Schmidt et al., 1978). סביר שעמדה זו מקורה בקריאה לא נכונה של המידה בה חדרה טכנולוגית המחשב לכל תחומי החיים, ויש מקום לבדוק אם היא אינה נחלתו של דור מפתחי המבחנים בלבד, ולא של דור הנבחנים שאולי יעמדו בבחינות ממוחשבות בעשורים הקרובים.

מוקד הביקורת השני ייחודי למבחנים גמישים. מאחר ובמבחנים גמישים, קבוצת הפריטים המהווה מבחן היא שונה מנבחן לנבחן, ותלויה ברמת היכולת שלו ובתגובותיו לפריטים קודמים, מתהווה מצב בו אין מבחן יחיד המודד את יכולת הנבחן. השאלה הציבורית שתעורר היא, כיצד ניתן לסמוך על התיאוריה הפסיכומטרית שתבטיח את שקילותם של נוסחי בחינה שמספרם עשוי להגיע למספר הנבחנים (יש כ-  $1.6 \times 10^{27}$  דרכים שונות ליצירת מבחן בן 20 פריטים מתוך מאגר בין 200 פריטים). באופן מעשי יהיה כמובן מספר קטן בהרבה של מבחנים שונים, אך גם מספר זה הוא עדיין גדול עד מאוד.

ביקורת נוספת, שהינה פחות ממוקדת, ניתן לכנות בשם "הביקורת הלודיטית". הכוונה היא לעמדה הכללית של התנגדות לטכנולוגיה באשר היא, אשר בהקשר למבחנים ממוחשבים תכוון כנגד העדר היחס האישי לנבחן וכנגד העובדה שמכונה מחליפה אדם בתפקיד שיש בכוחו לחרוץ גורלות. את עצמת הביקורת הזו לא תצמצם העובדה שבתפקידים רבים, החלטה המתקבלת על-ידי תוכנית מחשב היא מהימנה יותר מהחלטה אנושית, ואף לא העובדה שתגובותיהם של נבחנים למחשב עשויות להיות כנות יותר מאשר תגובותיהם למראיין בשר ודם (ראה סקירה של Hofer & Green, 1985).

אמת מסוימת ב"ביקורת הלודיטית" היא, שבדומה לכל טכנולוגיה אחרת, גם טכנולוגית המבחן הממוחשב מגבירה את כוחו של המשתמש בה. שימוש לא אחראי בטכנולוגיה עלול להגביר את הנזקים, כפי ששימוש אחראי ונבון עשוי להגדיל את התועלת. תמיד יהיה נכון לומר שמבחן קונבנציונלי הבנוי כהלכה והנתון בידיים נאמנות, יהיה מועיל יותר ממבחן ממוחשב חובבני ומרושל. אך כפי שצוין כבר על-ידי Madaus (1986), חשיבותה של הביקורת הלודיטית היא בכך שהיא ממקדת את תשומת הלב אל הסכנות הטמונות בטכנולוגיה, מאפשרת להימנע מהן על-ידי עצם חשיפתן, ובכך מפנה את המקום לניצול היתרונות הטכנולוגיים. בהקשר של מבחנים, דווקא המחשב הוא זה שיאפשר לשוב אל הגישה היחידנית ויתן לנבחן את האפשרות להתבטא בצורה חופשית, ולא רק על-ידי סימון תשובה אחת מתוך אלה המוצעות לו.

התפתחות המבחנים הממוחשבים בתחום המיון, הייעוץ והאבחנה הקלינית, תביא בעתיד הקרוב - על-פי דעתו של עד מומחה (Matarazzo, 1983) - להתרבות התביעות המשפטיות על רקע שימוש לא מקצועי בטכנולוגיה החדשה. כפי שהדבר קרה כבר ברפואה, ולמרות שהנזקים יהיו קטנים יותר, הרי שבכל מקרה שבו יתברר כי התקבלה החלטה מוטעית שיש

עמה נזק לפרט (דחיית מועד לעבודה, דחיית מועד ללימודים, החלטה על טיפול או אישפוז), והחלטה זו אירעה כתוצאה מכשל חומרה או תוכנה, תהיה למקרה תהודה ציבורית עצומה.

#### 4. המעבר למבחנים ממוחשבים

בעת שקילת מעבר ממבחן קונבנציונלי למבחן ממוחשב, יש להבחין בין המקרה שבו המבחן הממוחשב אמור להחליף את המבחן הקונבנציונלי ולבוא תחתיו, לבין המקרה בו המבחן הקונבנציונלי אמור להמשיך ולהתקיים במקביל למבחן הממוחשב. המקרה השני הוא מורכב יותר, שכן הוא מחייב נקיטת מספר צעדים להבטחת ההקבלה בין שתי אופנויות הבחינה. בהמשך הדברים, נציין מהם הצעדים המיוחדים למהלך כזה, בנוסף לצעדים אותם יש לנקוט בכל מקרה של מיחשוב מבחן.

לגבי מבחן ממוחשב, כמו לגבי כל מבחן שהוא, יש להבטיח את מהימנות המדידה ותקפותה. כאשר מדובר במבחן גמיש המבוסס על תת"ל, מתעוררות לפחות שתי בעיות שמקורן במודל של תת"ל. הראשונה קשורה למאגר הפריטים שעליו מבוסס המבחן, ונובעת מכך שלצורך מבחן גמיש לא בהכרח יתאימו כל הפריטים שהועברו במבחן הקונבנציונלי. יש להקפיד, בבניית מאגר הפריטים, על חד-מימדיות, או לפחות על דומיננטיות של מימד אחד, ועל-כן לא כל פריט יכול להיכלל בו. מתעוררת איפוא השאלה, אם תקפותה של המדידה נשמרת. אם הפריטים המושמטים מן המאגר מודדים גורמים נוספים, שהם רלבנטיים למדידה, אזי אין ספק שתקפותה של המדידה תיפגע. מאידך, אם רב המימדיות נובעת מגורמים שאינם קשורים לקריטריון שהציון במבחן אמור לנבא, או לתוכן שהמבחן אמור למדוד, השמטת פריטים תגרום להעלאת תקפות המדידה.

הבעיה השניה נעוצה באי-היכולת ליישם את המדדים המסורתיים של מהימנות לגבי מבחנים המבוססים על תת"ל. בתורת התגובה לפריט לא קיים מדד יחיד, המתאר את מהימנות המבחן, כי אם אוסף של ערכים המציינים את טעות המדידה (או את ההופכי לה - כמות האינפורמציה) עבור כל רמה ורמה בסולם היכולת. מקבוצת ערכים זו ניתן לחשב מדד המקביל לערך המסורתי של מהימנות עבור כל רמה בסולם היכולת, ועל-ידי סכימה, או מיצוע של ערכים אלה, להגיע למדד מהימנות יחיד (Green, Bock, Humphreys, Linn & Reckase, 1984) אך יש לזכור כי פירושו המדויק של ערך יחיד כזה, אינו זהה למדד המהימנות המסורתי.

בעוד שבמבחן המסורתי מחושבת מהימנות המדידה של הציון הכולל (מספר התשובות הנכונות) במבחן, הרי שבמבחן גמיש המבוסס על תת"ל יתן ערך כזה אומדן למהימנות המדידה של היכולת <sup>8</sup>. מאחר וסולם היכולת וסולם הציונים המקובל במבחן קונבנציונלי, אינם עומדים בקשר ליניארי זה עם זה, לא ניתן להשוות את שני מדדי המהימנות בצורה פשוטה.

מאותה הסיבה גם לא ניתן להגיע לכיול מושלם של ציוני מבחן גמיש, המבוסס על תת"ל, עם ציוני מבחן קונבנציונלי - צעד שהוא הכרחי אם שתי צורות הבחינה אמורות להתקיים זו בצד זו. ניתן למצוא טרנספורמציה בין שני הציונים, כך שנבחן יקבל ציון דומה בשני המבחנים, אך לא תתקיים הדרישה החזקה הנדרשת מפרוצדורות כיול - טענות המדידה של הציונים תהיה זהה בשני הסולמות.

מבחינה טכנית, לשם השגת מדידה ברמה טובה, דרוש במבחן הממוחשב הגמיש מאגר פריטים גדול יחסית, שבו כל פריט נבדק בצורה נסיונית על מאות (ולפעמים אלפי) נבחנים. קשה לקבל אומדנים טובים של פרמטרים על מספר נבחנים הקטן מ-500. אם המבחן הממוחשב מבוסס על תת"ל, יש לבדוק שהמודל של תת"ל אומנם מתקיים לגבי הפריטים הנדונים כאשר הם מועברים לאוכלוסיה הרלבנטית. הבדיקה צריכה לכלול את צורת העקומה האופיינית של הפריטים ואת חד-מימדיות המאגר כפי שצוין לעיל בסעיפים 3.2.1 ו-3.2.2.

במעבר ממבחן נייר-ועפרון למבחן ממוחשב יש להבטיח שהמיכשור מתאים לנבחנים. הבעיות בתחום זה הן קריאות הפריטים כאשר הם מוצגים על צג המחשב, ומידת הנוחיות של אמצעי התגובה העומדים לרשות הנבחן (למשל התאמת האמצעים לאיטרי יד-ימין). בנוסף, יש להבטיח שהנבחנים יודעים להשתמש במחשב - בין אם על-ידי אימון מוקדם ובין אם באמצעות למידה במהלך הבחינה עצמה. כאמור, (ס' 3.2.6) דווקא המבחן הממוחשב מאפשר גמישות רבה באימון הנבחנים ובבדיקה אם הבינו את ההוראות. עבור סוגי פריטים מסוימים (ר' לעיל ס' 3.2.7), יש להבטיח שמיחשובם אינו משנה את אופיים - במיוחד במקרים בהם המבחן הממוחשב אמור לפעול במקביל למבחן הקונבנציונלי.

ההפעלת מבחן ממוחשב אין לשכוח את עמדות הנבחנים. יש להתכונן לביקורת אפשרית, שתכוון בעיקר כנגד העובדה שבמבחן גמיש כל נבחן עשוי להיבחן על קבוצת פריטים שונה מזו שיקבלו עמיתיו. על כך יש להשיב בהסבר קצר על הדרך בה נבחרים פריטים עבור כל נבחן, ואפשר להיעזר בדוגמה; הדוגמה שבדרך כלל ניתנת על-מנת לסבר את האוזן, היא זו המשווה את המבחן הגמיש לתחרות קפיצה לגובה. בתחרות כזו המתחרה יכול להחליט מה יהיה גובה הרף - וזאת בהתאם ליכולתו כפי שהוכחה בעבר או במהלך התחרות. את גובה הרף משנים עבור כל מתחרה, ולמרות זאת ברור לכל, שהדירוג הסופי של המתחרים הוא טוב לא-פחות (ואולי אף יותר) מזה שהיה מושג לו כולם היו מנסים את קפיצותיהם עבור כל רמות הגובה של הרף.

מן הסקירה הכללית על הצעדים שאותם יש לנקוט במעבר ממבחן קונבנציונלי למבחן ממוחשב, מתברר כי המעבר אינו קל ואינו מיידי. עם זאת, לאור האפשרויות החדשות הנפתחות על-ידי מיחשוב מבחנים, נראה כי המאמץ עשוי להיות כדאי בתנאים מסוימים. עבור תוכניות מיון והשמה רחבות היקף, סביר ששכרו של המאמץ בצידו. עם זאת,



לגבי בחינות הנכתבות לצורך חד-פעמי והמיושמות בקנה מידה זעיר או קטן, קרוב לוודאי שהמבחן הקונבנציונלי ישאר בשימוש עוד עשרות שנים.

## 5. סיכום

במאמר זה נסקרו היתרונות, המגבלות, והחסרונות של מבחנים ממוחשבים, תוך הדגשה על הגמשת מבחנים - התועלת שבה, והבעייתיות שהיא מעוררת. עמדנו על היתרונות המינהליים הגלומים במיחשוב מבחנים, שהחשובים בהם הם שיפור הבקרה והניהול, הפחתת שיעורי הטעויות והשגת הסכון בזמנם של הבוחן והנבחן. יתרונות אלו יש לעמת כנגד ההשקעה הטכנולוגית והכלכלית הכרוכה במיחשוב, אך צויין שהשקעה זו תלך ותרד עם חלוף הזמן ועם ההתקדמות הטכנולוגית.

הצגנו את התועלת הפסיכומטרית העיקרית הצפויה ממיחשוב מבחנים: האפשרות להגמשת המבחן והתאמתו לנבחן, והאפשרות להעשיר את סוגי המדדים, את סוגי המבחנים ואת התיאוריות העומדות בבסיסם. אלו הם הסיכויים הגדולים הצפונים במיחשוב מבחנים, ובצידם השאלות והסייגים.

הזכרנו, כי בדומה לכל חידוש טכנולוגי, כך גם בתחום מיחשוב מבחנים יש צורך בחשיבה ובהסברה של המשמעויות, ואולי אף הסיכונים, הטמונים ביישום הטכנולוגיה; אך קיים הסיכוי, כי בעבודת מתקר ויישום זהירה, ניתן יהיה להגיע לבחינה ממוחשבת שתשלב בתוכה מדידה מדויקת ויעילה, התייחסות לצרכיו האישיים ויכולתו של כל נבחן ונבחן, ואשר תוכל לשקף נאמנה את כל מרחב כשריו וידיעותיו של הנבחן.

### רשימת מקורות

- אוסין, ל., נשר, פ., הוראת חשבון בסיוע מחשב בבית הספר היסודי. עיונים בחינוך, 1979, חוברת מס' 24, 93-108.
- אלוף, א., השוואת טיב המדידה של מבחנים אדפטיביים וקונבנציונליים. עבודת גמר לתואר מוסמך, האוניברסיטה העברית, ירושלים, 1987.
- אנקווה, ד., מדריך לצינון, מסמך פנימי, המרכז הארצי לבחינות ולהערכה, 1985.
- אספורמס, י., עמדות נבחנים כלפי מבחני מיון ממוחשבים וקונבנציונליים כפונקציה של משתני אישיות. הרצאה שניתנה בכינוס המדעי ה-21 של הסתדרות הפסיכולוגיה בישראל, תל אביב, 1987.
- באשי, י. (יו"ר), דו"ח הוועדה לבדיקת השימוש במבחני כושר קבוצתיים במערכת החינוך. צורף לחוזר מיוחד של מנכ"ל משרד החינוך והתרבות מיום 26.12.1985.
- בודסקו, ד., ניתוח גורמים של מערכת מבחני הכניסה הכלל-אוניברסיטאיים תשמ"ה. דו"ח טכני ראשוני, המרכז הארצי לבחינות ולהערכה, דו"ח מס' 21, 1985.
- בלר, מ., יצוגים עציים ומרחביים ליחסי קירבה ויחסי דומיננטיות בין פריטים ומבחנים. חיבור לשם קבלת תואר ד"ר, האוניברסיטה העברית, 1982.
- בן-דור, י., מבדקים ממוחשבים - יתרונות. הרצאה שניתנה בכנס החטיבה הייעוצית-תעסוקתית של הסתדרות הפסיכולוגים, נתניה, 1985.

- Anastasi, A. (1986). Personal communication to B. Nevo.
- Ansley, T.N., & Forsyth, R.A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. Applied Psychological Measurement, 9, 37-48.
- American Psychological Association (1986a). Guidelines for computer-based tests and interpretations. Washington, DC: Author.
- American Psychological Association. (1986b). The APA Monitor, Vol. 17, no. 10.
- Baker, F.B. (1987). Methodology review: Item parameter estimation under the one-, two-, and three-parameter logistic models. Applied Psychological Measurement, Vol. 11, 2, 111-141.
- Balla, J.R., & McDonald, R.P. (1985). Latent trait item analysis and facet theory - a useful combination. Applied Psychological Measurement, 9, 191-198.
- Bejar, I.I. (1978). A comparison of conventional and computer-based achievement testing. In D.J. Weiss (Ed.), Proceedings of the 1977 Computerized Adaptive Testing Conference. Minn.: University of Minnesota.
- Berger, M. (1982). The "scientific approach" to intelligence: An overview of its history with special reference to mental speed. In H.J. Eysenck (Ed.), A model for intelligence. New York: Springer-Verlag.
- Bersoff, D.N. (1981). Testing and the law. American Psychologist, 36, 1047-1056.
- Betz, N.E., & Weiss, D.J. (1976a). Effect of immediate knowledge of results and adaptive testing on ability test performance (Research report 76-3). Minn.: University of Minnesota.
- Betz, N.E., & Weiss, D.J. (1976b). Psychological effects of immediate knowledge on results and adaptive ability testing. (Research report 76-4). Minn.: University of Minnesota.
- Birenbaum, M. (1985). Comparing the effectiveness of several IRT based appropriateness measures in detecting unusual response patterns. Educational and Psychological Measurement, 45, 523-534.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord, & M.R. Novick. (Eds.), Statistical theories of mental test scores.

- Reading, MA: Addison-Wesley.
- Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46, 443-459.
- Boring, F.G. (1942). Sensation and perception in the history of experimental psychology. New York: Appleton-Century-Crofts.
- Bridges, K.R. (1985). Test-completion speed: Its relationship to performance on three course-based objective examinations. Educational and Psychological Measurement, 45, 29-35.
- Bruner, J.S., Goodnow, J., & Austin, G.A. (1956). A study of thinking. New York: Wiley.
- Butcher, J.N., Keller, L.S., & Bacon, S.F. (1985). Current developments and future directions in computerized personality assessment. Journal of Consulting and Clinical Psychology, 53, 803-815.
- Butterfield, E.C., Nielsen, D., Tangen, K.L., & Richardson, M.B. (1985). Theoretically based psychometric measure of inductive reasoning. In S.E. Embretson (Ed.), Test design. Orlando, FL: Academic Press.
- Cannale, M. (1986). The promise and threat of computerized adaptive assessment of reading comprehension. In C.W. Stansfield (Ed.), Technology and language testing. Washington, DC: TESOL.
- Carroll, J.B. (1982). The measurement of intelligence. In R.J. Sternberg (Ed.), Handbook of human intelligence. Cambridge: Cambridge University Press.
- Church, A.T., & Weiss, D. (1980). Interactive computer administration of a spatial reasoning test (Research report 80-2). Minn.: University of Minnesota.
- Cleary, T.A., Linn, R.L., & Rock, D.A. (1968). An exploratory study of programmed tests. Educational and Psychological Measurement, 28, 345-360.
- Cooper, L.A., & Regan, D.T. (1982). Attention, perception and intelligence. In R.J. Sternberg (Ed.), Handbook of human intelligence. Cambridge: Cambridge University Press.
- Cory, C.H. (1977). Relative utility of computerized versus paper-and-pencil tests for predicting job performance. Applied Psychological Measurement, 1, 551-564.

- Cory, C.H. (1978). Interactive testing using novel item formats. In D.J. Weiss (Ed.), Proceedings of the 1977 CAT Conference. Minn.: University of Minnesota.
- Cory, C.H., Rimland, B., & Bryson, R.A. (1977). Using computerized tests to measure new dimensions of abilities: An exploratory study. Applied Psychological Measurement, 1, 101-110.
- Cronbach, L.J. (1957). The two disciplines of scientific psychology. American Psychologist, 12, 671-684.
- DeAyala, R.J., & Koch, W.R. (1986). A computerized implementation of a Flexilevel test and its comparison with a Bayesian computerized adaptive test. Paper presented at the meeting of NCME, San Francisco, 1986.
- Dillon, R.F., & Schmeck R.R., (Eds.), (1983). Individual differences in cognition (Vol. 1). New York: Academic Press.
- Dillon, R.F. (Ed.), (1985). Individual differences in cognition, Vol. 2. New York: Academic Press.
- Divgi, D.R. (1986). Does the Rasch Model really work for multiple choice items? Not if you look closely. Journal of Educational Measurement, 23, 283-298.
- Dorans, N.J., & Kingston, N.M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating to the GRE verbal scale. Journal of Educational Measurement, 22, 249-262.
- Dover, S. (1986). Reactions to testing by computer. Paper presented at the 21st International Congress of Applied Psychology. Jerusalem, Israel, 1986.
- Dragow, F., & Levine, M.V. (1986). Optimal detection of certain forms of inappropriate test scores. Applied Psychological Measurement, 10, 59-67.
- Dragow, F., & Lissak, R.I. (1983). Modified parallel analysis: A procedure for examining latent dimensionality of dichotomously scored item responses. Journal of Applied Psychology, 68, 363-373.
- Dragow, F., & Parsons, C.K. (1983). Application of unidimensional item response theory to multidimensional data. Applied Psychological Measurement, 7, 189-199.
- Egan, D.E. (1979). Testing based on understanding: Implications from studies of spatial ability. Intelligence, 3, 1-15.
- Embretson, S.E. (1984). A general latent trait model for response processes. Psychometrika, 49, 175-186.

- Embretson, S.E. (Ed.), (1985a). Test design. Orlando, FL: Academic Press.
- Embretson, S.E. (1985b). Multicomponent latent trait models for test design. In Test design. Orlando, FL: Academic Press.
- English, R.A., Reckase, M.D., & Patience, W.M. (1977). Application of tailored testing to achievement measurement. Behavior Research Methods and Instrumentation, 9, 158-161.
- Estes, W.K. (1982). Learning memory and intelligence. In Sternberg, R.J. (Ed.), Handbook of human intelligence. Cambridge: Cambridge University Press.
- Eysenck, H.J. (Ed.), (1982). A model for intelligence. New York: Springer-Verlag.
- Fischer, G.H. (1973). The linear logistic test model as an instrument in educational research. Acta Psychologica, 37, 359-374.
- Fischer, G.H. (1983). Some latent trait models for measuring change in qualitative observations. In D.J. Weiss (Ed.), New horizons in testing. New York: Academic Press.
- Fitts, P.M. (1966). Cognitive aspects of information processing: III. Set for speed versus accuracy. Journal of Experimental Psychology, 71, 849-857.
- Fowler, R.D. (1985). Landmarks in computer-assisted psychological assessment. Journal of Consulting and Clinical Psychology, 53, 748-759.
- Garrison, W.M., & Baumgarten, B.S. (1986). An application of computer adaptive testing with communication handicapped examinees. Educational and Psychological Measurement, 46, 23-35.
- Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. British Journal of Mathematical and Statistical Psychology, 33, 234-246.
- Gopher, D., & Kahneman, D. (1971). Individual differences in attention and the prediction of flight criteria. Perceptual and Motor Skills, 33, 1335-1342.
- Gould, S.J. (1981). The mismeasure of man. New York: Norton.
- Greaud, V.A., & Green, B.F. (1986). Equivalence of conventional and computer presentation of speed tests. Applied Psychological Measurement, 10, 23-34.

- Green, B.F. (1984). Computer-based adaptive testing: The state-of-the-art in 1984. An address presented at the annual meeting of the APA in Toronto, 1984.
- Green, B.F., Bock, R.D., Humphreys, L.G., Linn, R.L., & Reckase, M.D. (1984). Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement, 21, 347-360.
- de Gruijter, D.N.M. (1986). Small N does not always justify Rasch model. Applied Psychological Measurement, 10, 187-194.
- Hambleton, R.K., & Murray, L.N. (1983). Some goodness of fit investigations for item response models. In R.K. Hambleton (Ed.), Applications of Item response theory. BC: Educational Research Institute of BC.
- Hambleton, R.K., & Rovinelli, R.J. (1986). Assessing the dimensionality of a set of test items. Applied Psychological Measurement, 10, 287-302.
- Hambleton, R.K., & Swaminatan, H. (1985). Item response theory: Principles and applications. Dordrecht, The Netherlands: Kluwer.
- Haney, W. (1981). Validity vaudeville and values: A short history of social concerns over standardized testing. American Psychologist, 36, 1021-1034.
- Hattie, J. (1985). Methodological review: Assessing unidimensionality of tests and items. Applied Psychological Measurement, 9, 139-164.
- Hedl, J.J., O'Neil, H.F., & Hansen, D.N. (1973). Affective reactions toward computer-based intelligence testing. Journal of Consulting and Clinical Psychology, 40, 217-222.
- Hicks, M.M. (1986). Computerized multilevel ESL testing, a rapid screening methodology. In C.W. Stansfield (Ed.), Technology and language testing. Washington, DC: TESOL.
- Hofer, P.J., & Green, B.F. (1985). The challenge and creativity in computerized psychological testing. Journal of Consulting and Clinical Psychology, 53, 826-838.
- Hoffman, K.I., & Lundberg, G.D. (1976). A comparison of computer-monitored group tests with paper-and-pencil test. Educational and Psychological Measurement, 36, 791-809.
- Holland, J.G. (1958). Human vigilance. Science, 128, 61-67.
- Holland, P.W., & Rosenbaum, P.R. (1985). Conditional association and unidimensionality in monotone latent variable models (Technical report 85-65). Princeton, NJ: ETS.

- Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). Item response theory, applications to psychological measurement. Homewood, IL: Dow Jones-Irwin.
- Hunt, E., Lunneborg, C., & Lewis, J. (1975). What does it mean to be high verbal? Cognitive Psychology, 7, 194-227.
- Jannarone, R.J. (1986). Conjunctive item response theory kernels. Psychometrika, 51, 357-373.
- Johnson, D.F., & Mihal, W.L. (1973). Performance of blacks and whites in computerized versus manual testing environments. American Psychologist, 694-699.
- Johnson, D.F., & White, C.B. (1980). Effects of training on computerized test performance in the elderly. Journal of Applied Psychology, 65, 357-358.
- Kahneman, D., Ben-Ishai, R., & Lotan, M. (1973). Relation of a test of attention to road accidents. Journal of Applied Psychology, 58, 113-115.
- Kingston, N.M., & Dorans, N.J. (1982). The feasibility of using item response theory as a psychometric model for the GRE Aptitude Test (GRE Board professional report 19-12). Princeton, NJ: ETS.
- Kingston, N.M., & Dorans, N.J. (1984). Item location effects and their implications for IRT equating and adaptive testing. Applied Psychological Measurement, 8, 147-154.
- Kingston, N.M., & Dorans, N.J. (1985). The analysis of item-ability regressions: An exploratory IRT model fit tool. Applied Psychological Measurement, 9, 281-288.
- Koch, B.R., & Patience, W.M. (1978). Student attitudes toward tailored testing. In: D.J. Weiss (Ed.), Proceedings of the 1977 CAT Conference. Minn.: University of Minnesota.
- Krug, S.E. (Ed.), (1984). Psychware: A reference guide to computer-based products for behavioral assessment in psychology, education and business. Kansas City, MO: Test Corporation of America.
- Lee, J.A. (1986). The effects of past computer experience on computerized aptitude test performance. Educational and Psychological Measurement, 46, 727-733.
- Lee, J.A., Moreno, K.E., & Sympson, J.B. (1986). The effects of mode of test administration on test performance. Educational and Psychological Measurement, 46, 467-474.



- Levine, M.V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. British Journal of Mathematical and Statistical Psychology, 35, 42-56.
- Levine, M.V., & Drasgow, F. (1983). Appropriateness measurement: Validating studies and variable ability models. In D.J. Weiss (Ed.), New Horizons in Testing. New York: Academic Press.
- Linn, R.L. (1986). Educational testing and assessment; research needs and policy issues. American Psychologist, 41, 1153-1160.
- Lord, F.M. (1970). Item characteristic curves estimated without knowledge of their mathematical form - a confrontation of Birnbaum's logistic model. Psychometrika, 35, 43-50.
- Lord, F.M. (1971a). A theoretical study of two-stage testing. Psychometrika, 36, 227-224.
- Lord, F.M. (1971b). The self scoring flexilevel test. Journal of Educational Measurement, 8, 147-151.
- Lord, F.M. (1977). A broad-range tailored test of verbal ability. Applied Psychological Measurement, 1, 95-100.
- Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.
- Lord, F.M. (1983). Small N justifies Rasch Model. In D.J. Weiss (Ed.), New horizons in testing. New York: Academic Press.
- Lord, F.M. (1984). Conjunctive and disjunctive item response functions (Research report 84-85). Princeton, NJ: ETS.
- Lord, F.M., & Wingersky, M.S. (1985). Sampling variances and covariances of parameter estimates in IRT. In D.J. Weiss (Ed.), Proceedings of the 1982 IRT and CAT Conference. Minn.: University of Minnesota.
- Madaus, G.F. (1986). The perils and promises of new tests and new technologies: Dick and Jane and the Great Analytical Engine? In The redesign of testing for the 21st century. Proceedings of the 1985 ETS invitational conference. Princeton, NJ: ETS.
- Matarazzo, J.D. (1983). Computerized psychological testing. Science, 221, 323.
- McBride, J.R., & Martin, J.T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D.J. Weiss (Ed.), New horizons in testing. New York: Academic Press.
- McBride, J.R., & Symson, J.B. (1985). The computerized adaptive testing system development project. In D.J. Weiss (Ed.),

- Proceedings of the 1982 IRT and CAT conference. Minn.: University of Minnesota.
- McDonald, R.P. (1981). The dimensionality of tests and items. British Journal of Mathematical and Statistical Psychology, 34, 100-117.
- McKinley, R.L., & Mills, C.N. (1985). A comparison of several goodness-of-fit statistics. Applied Psychological Measurement, 9, 49-57.
- Mislevy, R.J. (1986). Bayes modal estimation in item response models. Psychometrika, 51, 177-195.
- Mislevy, R.J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. Applied Psychological Measurement, 11, 81-91.
- Molholt, G., & Presler, A.M. (1986). Correlation between human and machine ratings of Test of Spoken English reading passages. In C.W. Stansfield (Ed.), Technology and language testing. Washington, DC: TESOL.
- Moreno, K.E., Wetzel, C.D., McBride, J.R., & Weiss, D.J. (1984). Relationship between corresponding Armed Service Vocational Aptitude Battery (ASVAB) and Computerized Adaptive Testing (CAT) subtests. Applied Psychological Measurement, 8, 155-163.
- Osin, L. (1984). TOAM: CAI on a national scale. Proceedings of the 4th Jerusalem Conference on Information Technology, IEEE Computer Society.
- Pew, R.W. (1969). The speed accuracy operating characteristic. Acta Psychologica, 30, 16-26.
- Phillips, S.E., & Mehrens, W.A. (1987). Curricular differences and unidimensionality of achievement test data: An exploratory analysis. Journal of Educational Measurement, 24, 17-39.
- Pine, S.M., Church, A.T., Gialluca, K.A., & Weiss, D.J. (1979). Effects of Computerized Adaptive Testing on black and white students (Research report 79-2). Minn.: University of Minnesota.
- Pine, S.M., & Weiss, D.J. (1978). A comparison of the fairness of adaptive and conventional testing strategies (Research report 78-1). Minn.: University of Minnesota.

- Prestwood, J.S. (1978). Effects of knowledge of results and varying proportion correct on ability, test performance and psychological variables. In D.J. Weiss (Ed.), Proceedings of the 1977 CAT Conference. Minn.: University of Minnesota.
- Prestwood, J.S., & Weiss, D.J. (1978). The effects of knowledge of results and test difficulty on ability, test performance and psychological reactions to testing (Research report 78-2). Minn.: University of Minnesota.
- Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. Chicago: The University of Chicago Press.
- Ree, M.J. (1981). The effects of item calibration sample size and item pool size on adaptive testing. Applied Psychological Measurement, 5, 11-19.
- Reid, J. (1986). Using the Writer's Workbench in composition teaching and testing. In C.W. Stansfield (Ed.), Technology and Language Testing. Washington, DC: TESOL.
- Rogers, J.H., & Hattie, J.A. (1987). A Monte Carlo investigation of several person and item fit statistics for item response models. Applied Psychological Measurement, 11, 47-57.
- Rosenbaum, P.R. (1984). Testing the local independence assumption in item response theory (Technical report 84-45). Princeton, NJ: ETS.
- Roskam, E.E. (in press). Toward a psychometric theory of intelligence. In E.E. Roskam, & R. Suck (Eds.), Progress in mathematical psychology. Amsterdam: North-Holland.
- Sachar, J., & Fletcher, J.D. (1978). Administering paper-and-pencil tests by computer, or the medium is not always the message. In D.J. Weiss (Ed.), Proceedings of the 1977 CAT Conference. Minn.: University of Minnesota.
- Samejima, F. (1974). Normal Ogive model on the continuous response level in the multidimensional latent space. Psychometrika, 39, 111-121.
- Scheiblechner, H. (1985). Psychometric models for speed-test construction: The linear exponential model. In S.E. Embretson (Ed.), Test design. Orlando, FL: Academic Press.
- Schmidt, F.L., Urry, V.W., & Gugel, J.F. (1978). Computer assisted tailored testing: Examinee reactions and evaluations. Educational and Psychological Measurement, 38, 265-273.

- Skinner, H.A., & Pakula, A. (1986). Challenge of computers in psychological assessment. Professional Psychology: Research and Practice, 17, 44-50.
- Space, L.G. (1981). The computer as psychometrician. Behavior Research Methods and Instrumentation, 13, 595-606.
- Sternberg, R.J. (1977). Intelligence, information processing and analogical reasoning: The componential analysis of human abilities. NJ: Erlbaum.
- Sternberg, R.J. (1979). The nature of mental abilities. American Psychologist, 34, 214-230.
- Sternberg, R.J. (1982). Reasoning, problem solving and intelligence. In R.J. Sternberg (Ed.), Handbook of human intelligence. Cambridge: Cambridge University Press.
- Sternberg, S. (1966). High speed scanning in human memory. Science, 153, 652-654.
- Swaminathan, H. (1983). Parameter estimation in item response models. In R.K. Hambleton (Ed.), Applications of item response theory. BC: Educational Research Institute of BC.
- Swaminathan, H., & Gifford, J. (1980). Estimation of parameters in latent trait models. In D.J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minn.: University of Minnesota.
- Sympson, J.B. (1978). A model for testing with multidimensional items. In D.J. Weiss (Ed.), Proceedings of the 1977 CAT Conference. Minn.: University of Minnesota.
- Sympson, J.B., & Moreno, K.E. (1985). Validity of adaptive testing: A summary of research results. Paper presented at the APA convention.
- Thissen, D. (1983). Timed testing: An approach using item response theory. In D.J. Weiss (Ed.), New horizons in testing. New York: Academic Press, 1983.
- Thissen, D., & Wainer, H. (1985). Confidence envelopes for item response functions. In D.J. Weiss (Ed.), Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference. Minn.: University of Minnesota.
- Thompson, J.G., & Weiss, D.J. (1980). Criterion-related validity of adaptive testing strategies (Research report 80-3). Computer Adaptive Testing Laboratory, Psychometric Methods Program, University of Minnesota.

- Thurstone, L.L. (1937). Ability, motivation and speed. Psychometrika, 2, 249-254.
- Trabin, T.E., & Weiss, D.J. (1983). The person response curve: Fit of individuals to item response theory models. In D.J. Weiss (Ed.), New horizons in testing. New York: Academic Press.
- Traub, R.E. (1983). A priori considerations in choosing an item response model. In R.K. Hambleton (Ed.), Applications of item response theory. BC: Educational Research Institute of BC.
- Tsutakawa, R.K. (1985). Estimation of item parameters and the GEM algorithm. In D.J. Weiss (Ed.), Proceedings of the 1982 IRT and CAT Conference. Minn: University of Minnesota.
- Tsutakawa, R.K., & Hsin, Y.L. (1986). Bayesian estimation of item response curves. Psychometrika, 51, 251-267.
- Tung, P. (1986). Computerized Adaptive Testing: Implications for language test developers. In C.W. Stansfield (Ed.), Technology and language testing. Washington, DC: TESOL.
- Urry, V.W. (1977). Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement, 14, 181-196.
- Vale, C.D. (1978). Computerized administration of free-response items. In D.J. Weiss (Ed.), Proceedings of the 1977 Computerized Adaptive Testing Conference. Minn.: University of Minnesota.
- Vale, C.D. (1985). Design of a microcomputer-based adaptive testing system. In D.J. Weiss (Ed.), Proceedings of the 1982 IRT and CAT Conference. Minn.: University of Minnesota.
- Vale, C.D., & Weiss, D.J. (1977). A comparison of information functions of multiple choice and free response vocabulary items (Research report 77-2). Minn.: University of Minnesota.
- van de Vijver, F.J.R. (1986). The robustness of Rasch estimates. Applied Psychological Measurement, 10, 45-57.
- Waters, B.K. (1977). An empirical investigation of the stratified adaptive computerized testing model. Applied Psychological Measurement, 1, 141-152.
- Weiss, D.J. (1973). The stratified adaptive computerized ability test (Research report 73-3). Minn.: University of Minnesota.
- Weiss, D.J. (1978). Discussion of the Sacher and Fletcher paper. In Proceedings of the 1977 CAT Conference. Minn.: University of Minnesota.

- Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing. Applied Psychological Measurement, 6, 473-492.
- White, P.O. (1973). Individual differences in speed, accuracy and persistence: A mathematical model for problem solving. In H.J. Eysenck (Ed.), The measurement of intelligence. Lancaster, England: MTP.
- Whitely, S.E. (1980). Multicomponent latent trait models for ability tests. Psychometrika, 45, 479-494.
- Wickelgren, W.A. (1977). Speed-accuracy tradeoff and information processing dynamics. Acta Psychologica, 41, 67-85.
- Wildgrube, W. (1985). Computerized testing in the German Federal Armed Forces (FAF): Empirical approaches. In D.J. Weiss (Ed.), Proceedings of the 1982 IRT and CAT conference. Minn.: University of Minnesota.
- Wingersky, M.S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test modes. In R.K. Hambleton (Ed.), Applications of item response theory. BC: Educational Research Institute of BC.
- Wingersky, M.S., Barton, M.A., & Lord, F.M. (1982). LOGIST user's guide. Princeton, NJ: ETS.
- Winsberg, S., Thissen, D., & Wainer, H. (1984). Fitting item characteristic curves with spline functions (Technical report 84-52). Program statistics research. Princeton, NJ: ETS.
- Yen, W.M. (1981). Using simulation results to choose a latent trait model. Applied Psychological Measurement, 5, 245-262.