

על התוקף של

מבחנים רחבי היקף בחינוך

פרופ' מיכל בלר

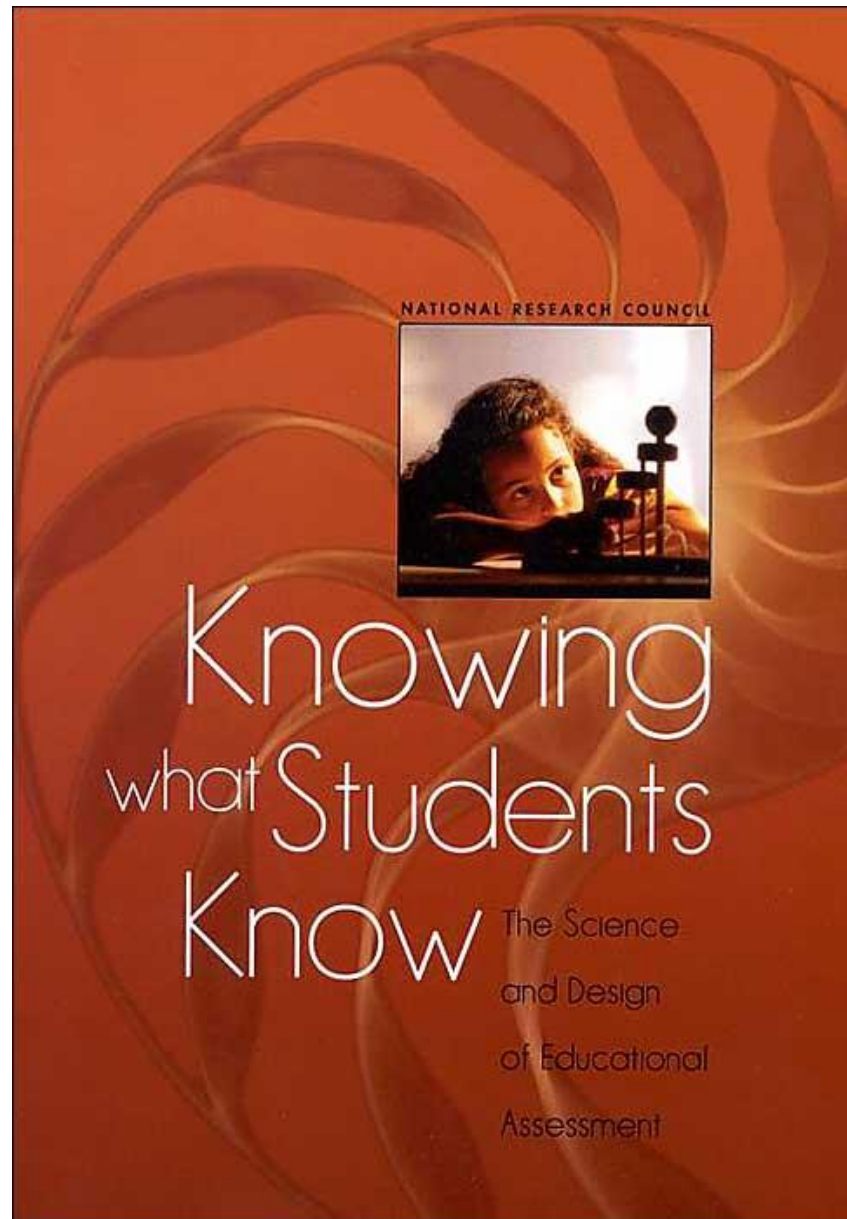
מנכ"לית ראמ"ה

הרשות הארצית למדידה והערכה בחינוך

יום עיון בנושא : כיצד לתקף? סוגיות במדידה והערכה

מכון הנרייטה סאלד , איל"ת

20.06.2007



הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

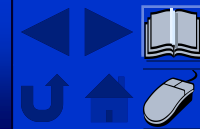
בניית מבחנים
תקפים

הערות לסיכום



רשות ארצית
למידה
והערכה
בחינוך

© 2007 # 2



**מרבית הויכוחים הנטושים כיום על בחינות,
כמו הויכוח על מבחנים "עתירי סיכון",
מקורם בבסיס באי הסכמות לגבי התוקף**

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

הערות לסיכום



רשות ארצית
למדידה
והערכה
בחינוך

NCES STANDARD: 2-6

PURPOSE :

To ensure that educational tests used in NCES surveys for measuring and making inferences about education-related domains are **valid, technically sound, and fair**.

To ensure that

the administration and scoring of educational tests are **standardized**,

that scales used over time are **stable**,

and that the results are reported in a **clear unbiased manner**

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפוט/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

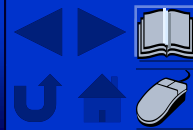
הערות לסיכום



רשות ארצית
למידה
והערכה
בחינוך

© 2007 # 4

NCES Validity Standard



STANDARD 2-6-2: Validity - All test instruments used in NCES surveys must meet the purpose(s) stated in the instrument specifications. All intended interpretations and proposed uses of raw scores, scale scores, cut scores, equated scores, and derived scores, including composite scores, sub-scores, score differences, and profiles, must be supported by evidence and theory.

GUIDELINE 2-6-2A: Evidence of validity should be based on analyses of the content, response processes (i.e. the thought processes used to produce an answer), internal structure of the instrument, and/or the relationship of scores to a criterion.

GUIDELINE 2-6-2B: The rationale for each intended use of the test instruments and test proposed interpretations of the scores obtained should be explicitly stated.

GUIDELINE 2-6-2C: When judgments occur in the validation process, the selection process for the judges (experts/observers/raters) and the criteria for judgments should be described.

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

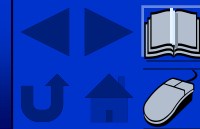
מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

הערות לסיכום



רשות ארצית
למידה
והערכה
בחינוך



- קבלת תלמידים שסיכויי הצלחתם באקדמיה הגבוהים ביותר
- איתור חוזקות וחולשות אצל תלמידים בבתי ספר
- הערכת איכות המורים ושיפוט ביצועי בתי ספר

■ השאלה היא באיזו מידה המסקנות הנגזרות מתוצאות המבחנים שנבנו למטרות לעיל הן מסקנות נכונות?

- NCLB requires that the assessments required by the law “be used for purposes for which such assessments are valid and reliable”

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

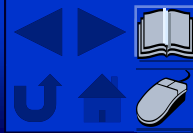
מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

הערות לסיכום



רשות ארצית
למידה
והערכה
בחינוך



משמעות מושג התוקף

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

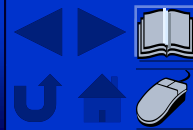
הערות לסיכום



רשות ארצית
למידה
והערכה
בחינוך

© 2007 # 7

משמעות המושג תוקף



תוקף אינו מאפיין של המבחן עצמו אלא:

- המושג תוקף משמש בתחום של מדידה חינוכית כדי לתאר **היסק הוא מסקנה המבוססת על ציון מבחן**. ניתן לגזור ממבחן סידרת מסקנות שאולי רק חלקן תקפות
- יש לאסוף **עדויות לאורך זמן** כדי לאושש או להפריך את הטענה **בדבר תקפותו של מבחן**
- מסגרת מארגנת של העדויות תחת **validity framework** יכולה לסייע
- השאלה הראשונה היא **מהו השימוש הנעשה בתוצאותיו ומהן המסקנות הנגזרות מהן**. מבחן יכול לשמש מספר מטרות שונות. עבור כל שימוש עשויות להיגזר ממנו מסקנות שונות שחלקן תקפות ואחרות פחות ולכן **שימוש אחד תקף והאחר פחות**.
- לשימושים שונים של מבחנים עשויות להיות השלכות שונות וגם לכך יש השפעה על התוקף – **consequential validity** (Cronbach, Messick)

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

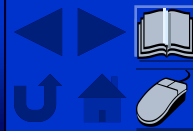
הערות לסיכום



רשות ארצית
למדידה
והערכה
בחינוך

© 2007 # 8

השלכות חיוביות של המיצ"ב בשנים תשס"ב-תשס"ו



■ כלי מקצועי ואחיד למדידת:

א- הישגים לימודיים בארבעה מקצועות יסוד (שפת אם, אנגלית, מתמטיקה, ומדעים וטכנולוגיה)

ב- אקלים בית-ספרי וסביבה לימודית

■ כלי לבתי הספר לבחינת הטמעת תוכנית הלימודים, עמידה ביעדים, והגדרת תוכניות עבודה מסודרות

■ סינכרון כל מערכת החינוך – מטה, מחוזות, רשויות ובתי-ספר – להלימה בין תהליכים לימודיים להישגים

■ יכולת לנתח את הישגים באוכלוסיות שונות, ולאתר פערים ומגמות

■ יכולת לנטר השפעה של התערבויות שונות

■ דיאלוג עם הפיקוח על נתונים במקום רשמים

...

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

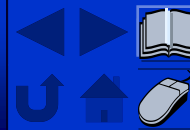
הערות לסיכום



רשות ארצית
למדידה
והערכה
בחינוך

© 2007 # 9

השלכות שליליות שנצפו מהפעלת המיצ"ב (בחלק מבתי הספר)



■ הסטת משאבים בשנת המיצ"ב:

- השקעת זמן במקצועות המיצ"ב על חשבון מקצועות לימוד אחרים
- העמדת המורים הטובים יותר ל"טובת" כיתות המיצ"ב

■ "למידה בשירות הבחינה":

- צמצום ההוראה לפי המפרט
- הכנה מוגזמת ואינפלציה בציונים
- חזרה על חומר של שנה קודמת על חשבון חודשי הוראת החומר של השנה הנוכחית

■ הפרת טוהר הבחינות:

- סיוע של מורים בעת הבחינה
- הרחקת תלמידים חלשים מהבחינה
- היעדר מוטיבציה וזלזול של תלמידים במבחנים, בעיקר בכיתות ח'

■ אחר:

- "חרדת בחינות" אצל ילדים לגבי מבחנים שלא שימשו בפועל למדידת הישגים אישיים
- דיון בלתי-מקצועי בתקשורת שהסיט את תשומת הלב מהעיקר לטפל
- השוואה בין שנים של נתונים שאינם מכוילים במלואם וגזירת מסקנות מוטעות לגבי מצב המערכת

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

הערות לסיכום



רשות ארצית
למדידה
והערכה
בחינוך

גורמים המגבילים תוקף

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

הערות לסיכום



רשות ארצית
למידה
והערכה
בחינוך



קשה למצוא דוגמאות לתוקף מלא
בשל כמות עצומה של ממצאים

הגורמים המגבילים יכולת לבנות מבחן שתקפות מסקנותיו מלאה:

שימוש בתוצאות המבחן
למטרות שלא להן נועד

מידה של נושאים שלא
צריכים להימדד –

Construct-Irrelevance
Variance

חוסר יכולת למדוד בצורה
טובה את כל הנדרש –

Construct Under
Representation

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

הערות לסיכום



רשות ארצית
למידה
והערכה
בחינוך

דוגמא:

בדיקה של **אוריינות שפה** מתקיימת לא אחת **ללא**
דרישה לעדות לגבי יכולת **כתיבה** ו**יכולת דיבור**
(בעקר בשל סיבות טכניות ותקציביות)

■ בחינות ה- **TOEFL**, שהן בחינות רחבות היקף
המועברות באינטרנט ברחבי העולם, מדגישות
מדידה של כל האלמנטים הנדרשים לבחינת השליטה
בשפה האנגלית כשפה זרה (קריאה, כתיבה, דיבור)

■ הרחבת המסגרת התיאורטית של בחינת אוריינות
שפת אם בפיזה לכלול מרכיב של **אוריינות שפה**
דיגיטלית

■ שיטה להרחבת דגימת התוכן - **matrix sampling** -
נהוגה במבחנים הבינלאומיים

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

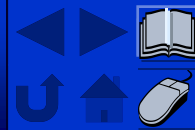
מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

הערות לסיכום



רשות ארצית
למדידה
והערכה
בחינוך



דוגמאות:

- **אינטראקציה בלתי רלוונטית בין המבחן למאפייני הנבחנים** (למשל מבחן מודפס באותיות רגילות הניתן לנבחנים קשיי ראייה)
- **שפה גבוהה** ובלתי מובנת במבחני מתמטיקה ומדעים (בעייתי בפרט לאוכלוסיות עולים)
- **הדילמה סביב מתן תנאים מותאמים ללקויי למידה** (כמו הקראה) במבחן שנועד לבדוק יכולת הבנת הנקרא

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

הערות לסיכום



רשות ארצית
למידה
והערכה
בחינוך

כיצד קובעים תוקף

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

הערות לסיכום



רשות ארצית
למידה
והערכה
בחינוך

■ ניתוח תוכן

■ ניתוח סטטיסטי של הקשרים בין ציוני המבחן לבין משתנים אחרים

■ ניתוח סטטיסטי של הביצועים במבחן

■ מהימנות הנתונים

■ ניתוח עומק של תהליך פתרון הפריטים על ידי נבחנים שונים

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

הערות לסיכום

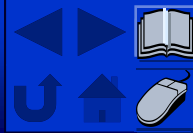
אף אחת מהעדויות האלה אינה מספיקה כשלעצמה כדי להוכיח שמבחן הוא תקף

עם זאת, עדות שלילית לגבי אחת מהן – יכולה לפסול מבחן



רשות ארצית
למידה
והערכה
בחינוך

ניתוח תוכן



- מבחן הוא מדגם של עולם התוכן הנמדד, ועליו לייצגו נאמנה

- דרך שטחית לבדוק זאת נקראת **"תוקף נראה" (face validity)**. בשל שטחיותה של שיטה זו יש מי שכינה אותה **faith validity**.

- שיטה זו שמשה רבות בתקופה בה הייתה התלהבות performance assessment' כאשר **האמונה הייתה שמשימות מורכבות הן בהכרח משימות הבודקות מיומנויות חשיבה גבוהות טוב יותר מאשר מבחנים סגורים**. השדה חיפש משימות עשירות, מעניינות ומאתגרות וזה כשלעצמו סיפק עדות לכאורה לתוקף. זה לא מספיק.

- הכנסת **מומחים** לשיפוט תוקף התוכן היא **הכרחית** אך כשלעצמה גם היא **אינה מספקת**. לעיתים אלמנטים טכניים קטנים כמו בחירת מסיחים לא טובה, או שימוש לא מכוון בשפה גבוהה מדי יכולים לשנות את אופי הפריט. זאת ועוד, המיומנויות הנדרשות לפתרון פריט תלויות במידה רבה באימון ובידע המוקדם שיש לקבוצת הנבחנים

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

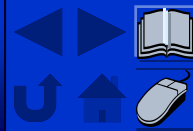
בניית מבחנים
תקפים

הערות לסיכום



רשות ארצית
למידה
והערכה
בחינוך

ניתוח הקשרים בין המבחנים ומדדים אחרים



הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

הערות לסיכום



רשות ארצית
למידה
והערכה
בחינוך

© 2007 # 18

■ **קריטריון חיצוני** - criterion related evidence.

נדיר מאד בתחום המדידה החינוכית הבית ספרית

□ מתאים לבניית מבחן חדש שיחליף מבחן מוכר ומשומש

■ **מדידת הקשר בין המבחן לבין מדדים אחרים:**

□ של אותו מבנה – **תוקף מתכנס** – convergent validity

□ של מבנים אחרים – **תוקף מבחין** – discriminant validity

■ **יש חשיבות במדידת שני סוגי הקשרים לעיל ובחינת השונות בעוצמות הקשר:**

□ אמפירית ידוע כי בין ציוני מבחנים המודדים משתנים שונים יש בפועל מתאם חיובי גבוה יחסית גם אם אינם מודדים אותו מבנה

- **ניתוחי תפקוד דיפרנציאלי (DIF) של פריטים במבחן עבור תת-קבוצות באוכלוסיה (למשל, מגדר).**

□ הניתוח נועד לאתר פריטים שבהחזקה קבוע של הציון הכללי במבחן מתגלה בהם תפקוד דיפרנציאלי בין הקבוצות הרלוונטיות.

- **לגבי פריטים שאותרו בניתוח DIF קיימת עדות להטיה בשל העובדה שהם מודדים דבר שונה לגבי כל אחת מתת הקבוצות – וזו עדות הפוגעת בתוקף התוכן של המבחן עבור הקבוצה בכללותה**

- **השאלה הזו עולה בעוצמה רבה בכל הקשור לבחינה של תלמידים עם לקויות למידה באמצעות מבחנים רגילים**

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

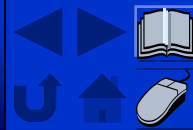
בניית מבחנים
תקפים

הערות לסיכום



רשות ארצית
למדידה
והערכה
בחינוך

על הקשר בין מהימנות לתוקף



■ **מהימנות היא תנאי הכרחי אך לא מספיק לתוקף.** במילים אחרות, יכול להיות מבחן מהימן אך לא תקף אולם לא יכול להיות מבחן לא מהימן וכן תקף

■ כאשר מדובר במהימנות מסוג "מבחן מקביל" – אזי בעצם מדובר בסוגיה של תוקף. **אם נוסחים מקבילים מובילים למסקנות שונות – הדבר יהווה עדות לפגיעה בתוקף של המבחן**

■ לעתים יש **תחליפיות** (trade-off) מסוימת בין מהימנות לתוקף בפרט כשמדובר בסוגיה של **construct under representation**

■ **אין תשובה מדעית להיקף התחליפיות – התשובה תלויה במטרת המבחן ובשימוש שנעשה בו**

- אם מדובר במבחן חורץ גורלות של פרטים יש להקפיד על מהימנות גבוהה
- אפשר לוותר מעט כאשר מדובר במבחן שהוא ברמת כחה, או מבחן פרטני תיאורי או דיאגנוסטי

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

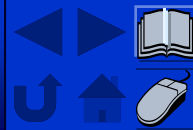
בניית מבחנים
תקפים

הערות לסיכום



רשות ארצית
למידה
והערכה
בחינוך

ניתוח עומק של תהליך פתרון הפריטים



■ החוקר יושב עם הנבחן ומבקש ממנו לנסות ולהסביר מה הוא עושה בבואו לפתור פריט במבחן

□ מהן המיומנויות והידע בהם השתמש לפתרון הפריט. למשל, האם הוא נשען על זיכרון וידע קודם או מפעיל מיומנות חשיבה?

■ גישה זו כמעט ואינה משמשת בפועל במבחנים רחבי היקף שכן היא מורכבת, גוזלת זמן ויקרה מאד

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

הערות לסיכום



רשות ארצית
למדידה
והערכה
בחינוך

תקפות "שיפור" / "ירידה"

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

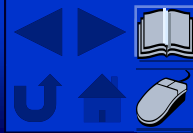
בניית מבחנים
תקפים

הערות לסיכום



רשות ארצית
למידה
והערכה
בחינוך

תקפות ציוני שיפור/ירידה



- לא אחת אנו עדים לדיווחים בעיתונות על עליה או ירידה בציונים ברמת הארצית
- דיווחים אחרים נוגעים לסוגיית הפער בין מגזרים באוכלוסיה ולשאלה האם קטנו או גדלו
- **האם יש ממש בדיווחים אלה?**

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

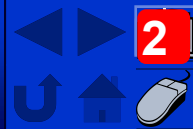
בניית מבחנים
תקפים

הערות לסיכום



רשות ארצית
למדידה
והערכה
בחינוך

"הכישלון" בתשס"ו...



15:05 , 12.06.06

news חדשות



מבחן המיצ"ב: נמשכת הירידה בהישגים

שלחו להדפסה

מה עושים



השרה תמיר. טעויות מדידה? צילום: גיל יוחנן

לחצו כאן להגדיל הטקסט

משרד החינוך פרסם היום את תוצאות מבחן המיצ"ב (מדדי יעילות וצמיחה בית ספריים) לשנת תשס"ו, שנערך בכ-1,200 בתי ספר. מחישוב ממוצעי המקצועות לתלמידי כיתות ה' וח' עולה כי במרבית המקצועות ישנה ירידה בציונים, למעט עברית שם נותר הממוצע 77. אך טענות קשות נשמעו כלפי ניתוח הנתונים: "אין כלי מקצועי שמודד רמת קושי של בחינה אחת לעומת הבחינה בשנה שעברה" מורן זליקוביץ

11 אחוזים מתלמידי כיתות ב' בחינוך הממלכתי נכשלים במבחני הבנת הנקרא. בקרב תלמידי החינוך המיוחד המשולבים בבתי הספר הרגילים הנתון חמור אף יותר - 44 אחוזים אינם יודעים לקרוא. כך עולה מניתוח ראשוני של נתוני בחינות המיצ"ב (מדדי יעילות וצמיחה בית ספריים) לשנת תשס"ו, שמשרד החינוך מפרסם היום (ב'). מנכ"ל משרד החינוך, שמואל אבואב, אמר כי "אנו מודאגים מכמות הנכשלים בקריאה. מאתר תלמידים אלה ונעבוד איתם פרטנית בכיתות בקבוצות מצומצמות של עד שלושה תלמידים".



הקדמה

משמעות מושג ה"תוקף"

גורמים המגבילים תוקף

כיצד קובעים תוקף

תקפות שיפור/ירידה

מדד ב"ת לבקרת שיפור

בניית מבחנים תקפים

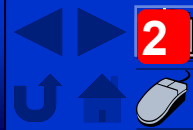
הערות לסיכום



רשות ארצית למדידה והערכה בחינוך

© 2007 # 24

אחרי החגיגה בתשס"ה....



4 מעריב יום א', ה' בסיון תשס"ה 12.6.2005

נחנו המיץ ב' החדשים: עלייה גדולה בהישגי התלמידים בישראל



מצב המיץ ב':

מקצוע:	ציון במבחן הקודם:	ציון במבחן השנה:
עברית	70	80
מתמטיקה	55	70
מדע	70	80
אנגלית	75	78

* הנתונים מתייחסים לתוצאות הביניים של תלמידי כיתות ח' בבתי הספר הממלכתיים-יהודיים כפי שנאספו עד כה

צילום: התחנה החדשה, לעיתים איל דרור

תעודת כבוד

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

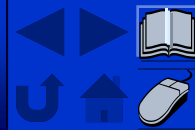
הערות לסיכום



רשות ארצית
למדידה
והערכה
בחינוך

© 2007 # 25

סוגית הכיול



- האם ציון 90 בבחינה (קלה) השנה מלמד על שיפור לעומת ציון של 75 בבחינה (קשה) אשתקד?
- ראמ"ה פועלת להטמיע בקרב כל השותפים לתהליך החינוכי ובקרב הציבור את ההבנה:
 - שלדירוגים יש משמעות, אם בכלל, רק בקרב אוכלוסיות רלוונטיות
 - להשוואות על-פני שנים יש ערך רק כאשר סולמות הציונים מכוילים. כיול נועד לנטרל את ההשפעה של קושי המבחן בשנה מסוימת.
- ההשוואות לאורך שנים, שנערכו בין ציוני המיצ"ב והבגרות, **אינן תקפות בשל העדר כיול בין הסולמות משנים שונות**
- ראמ"ה פועלת לכייל באופן מקצועי את סולמות ציוני המיצ"ב **כך ששנת תשס"ז תשמש כשנת בסיס להשוואות בנות-משמעות בעתיד.**

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיוד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

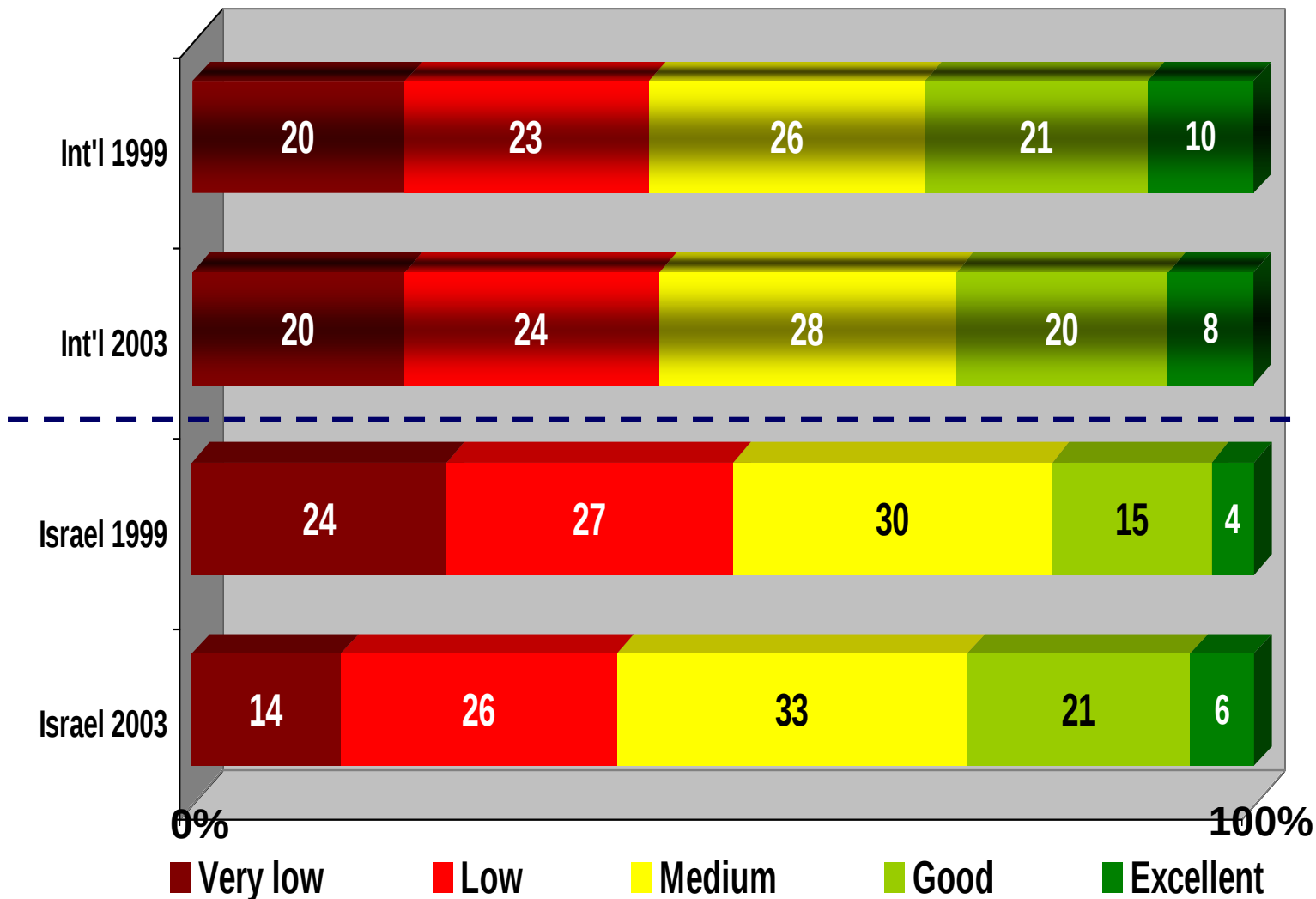
הערות לסיכום



רשות ארצית
למידה
והערכה
בחינוך

TIMSS – Math

2003 vs. 1999, Israel vs. International Performance



הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

הערות לסיכום

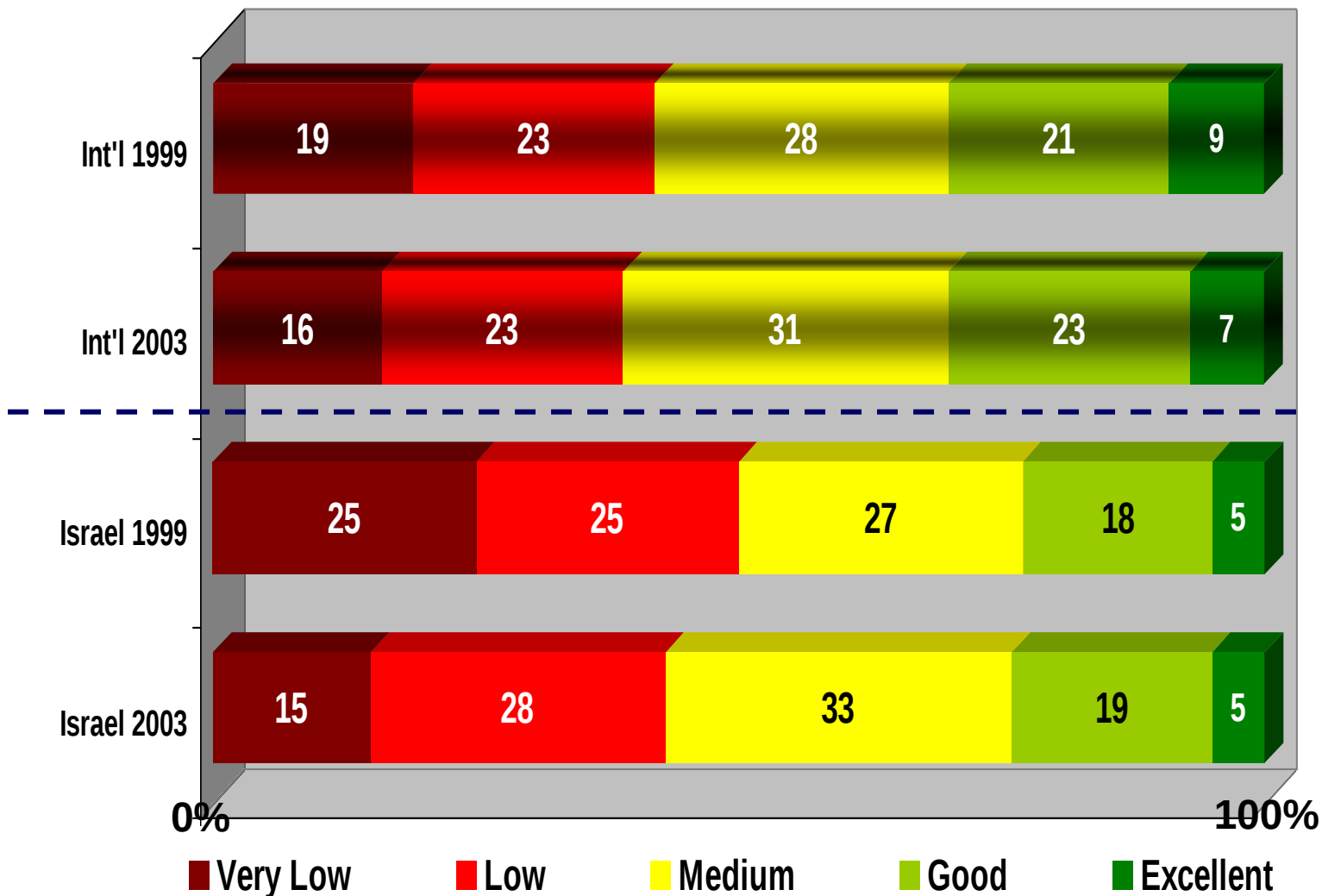


רשות ארצית
למידה
והערכה
בחינוך

© 2007 # 27

TIMSS – Science

2003 vs. 1999, Israel vs. International Performance



הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

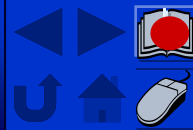
הערות לסיכום



רשות ארצית
למידה
והערכה
בחינוך

© 2007 # 28

Achievement Gap on TIMSS Reduced



Improvement (Std) on TIMSS - 2003 vs. 1999

	Hebrew Speakers	Arabic Speakers
Math	0.18	0.80
Science	0.15	0.80

Gaps (Std) on TIMSS - Hebrew vs. Arabic Speakers

	1999	2003
Math	1.0	0.50
Science	0.9	0.40

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

הערות לסיכום



רשות ארצית
למידה
והערכה
בחינוך

- בעקבות העובדה שמבחנים הפכו קובעים ועתירי סיכון (high stakes) עבור מנהלים, מורים ואחרים מתפתחת סביבם התנהלות המובילה בין השאר **לאינפלציה בציונים** [דן קורץ]

- **תופעה זו:**

- מורידה את תוקף הבחינות
- פוגמת ביכולת לערוך בדיקה תקפה של **מגמות לאורך זמן**
- מונעת יכולת לבחון **פערים** באוכלוסיה

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

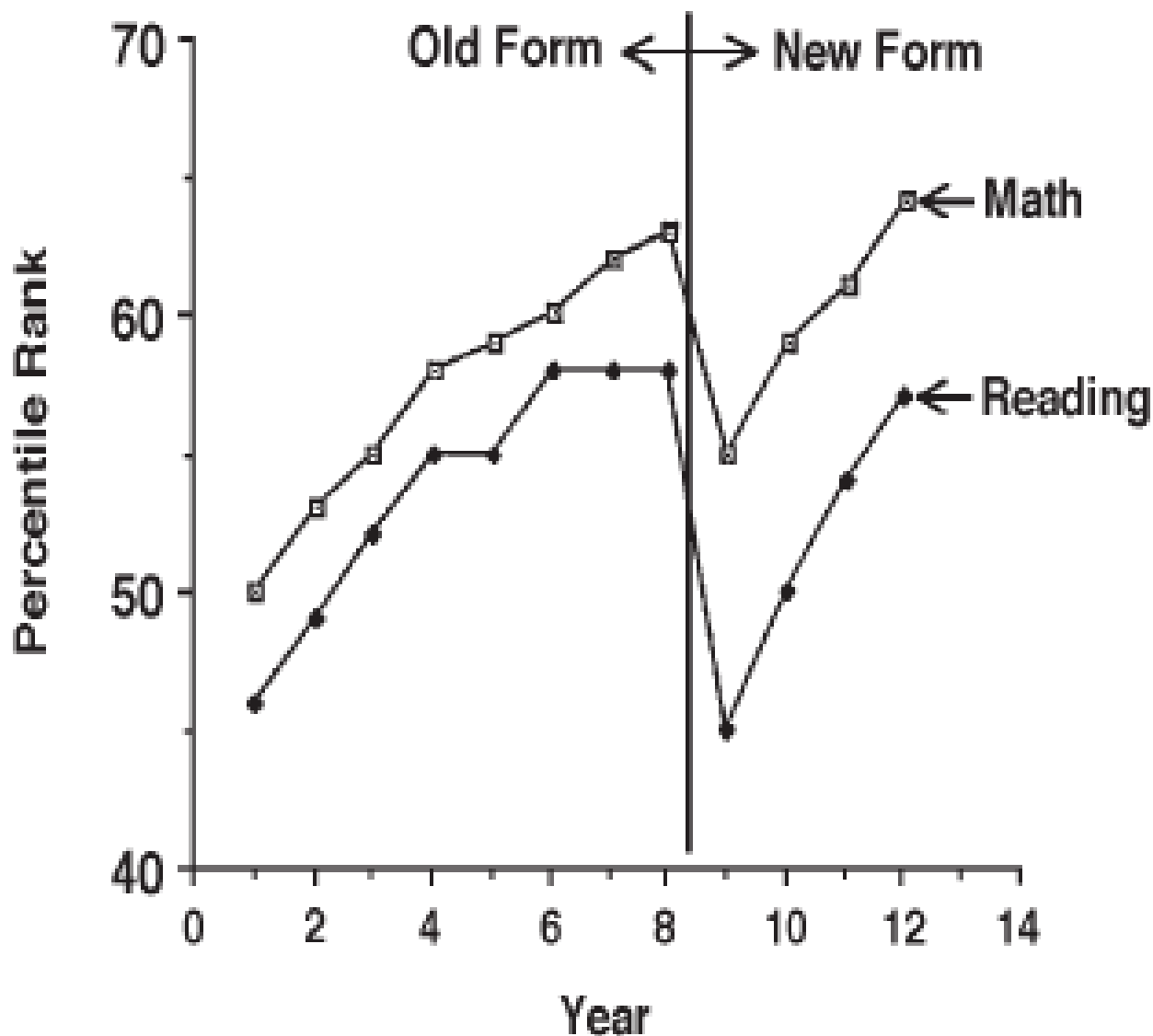
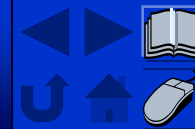
בניית מבחנים
תקפים

הערות לסיכום



רשות ארצית
למידה
והערכה
בחינוך

Results of Changing to a New Test Form



הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

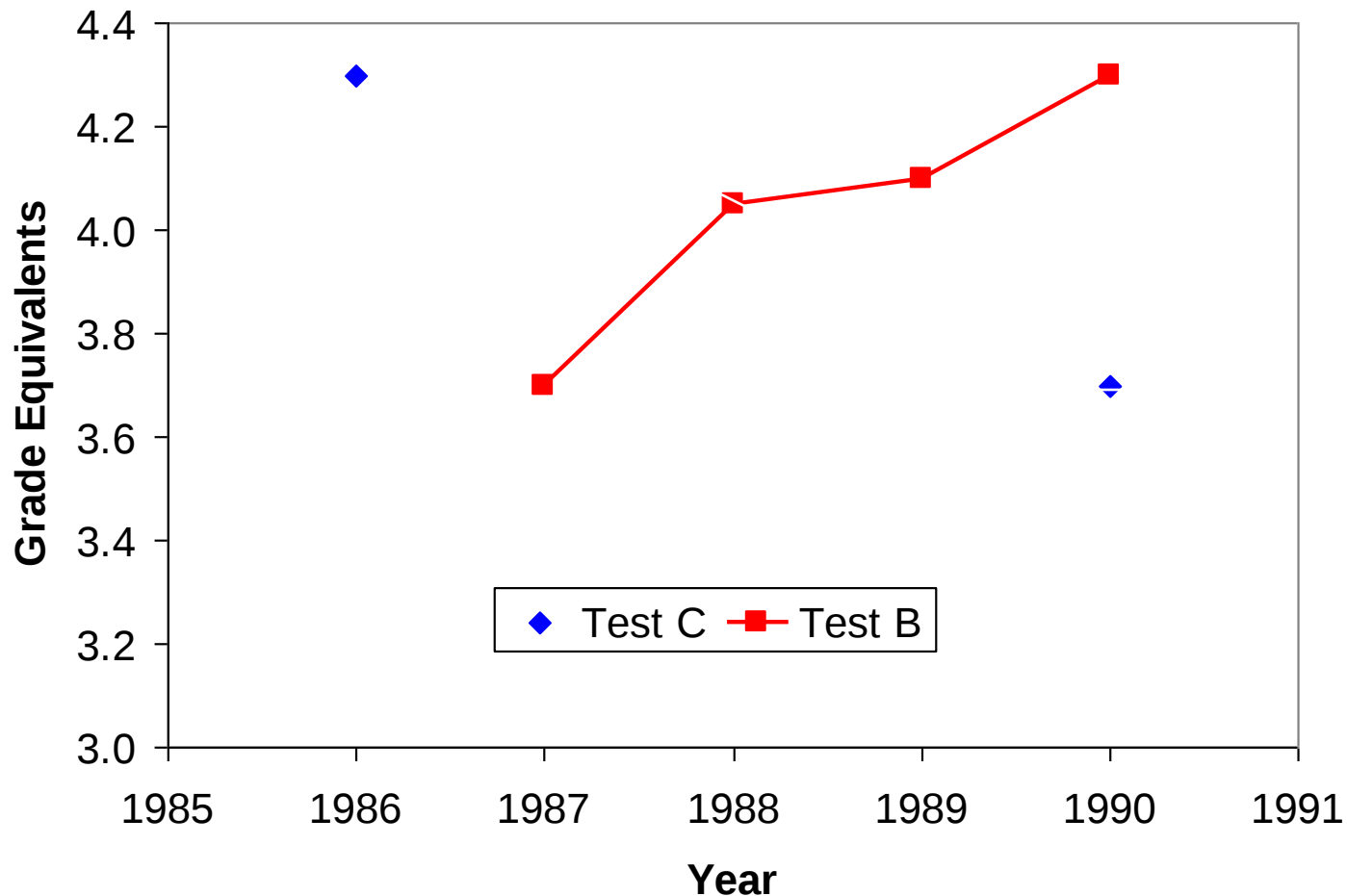
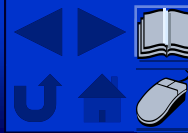
הערות לסיכום



רשות ארצית
למידה
והערכה
בחינוך

© 2007 # 31

Performance on Coached and Uncoached Tests



SOURCE: Adapted from Koretz, Linn, Dunbar, and Shepard (1991)

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

הערות לסיכום



רשות ארצית
למידה
והערכה
בחינוך

© 2007 # 32

מדד בלתי-תלוי לבקרת השיפור

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

הערות לסיכום



רשות ארצית
למידה
והערכה
בחינוך

■ **השוואה בין נתוני ה-NAEP, שהם ארציים ואינם נתפסים כעיתירי סיכון, לבין הבחינות המונהגות בכל אחת מהמדינות בארה"ב בנפרד**

Gains on State Reading Tests Evaporate on 2005 NAEP Has a "Race to the Bottom" Begun? Gains on State Reading Tests Evaporate on NAEP

Decline in 8th Grade Scores Points to "Middle School Slump"

הקדמה

משמעות מושג ה"תוקף"

גורמים המגבילים תוקף

כיצד קובעים תוקף

תקפות שיפור/ירידה

מדד ב"ת לבקרת שיפור

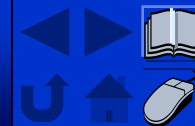
בניית מבחנים תקפים

הערות לסיכום

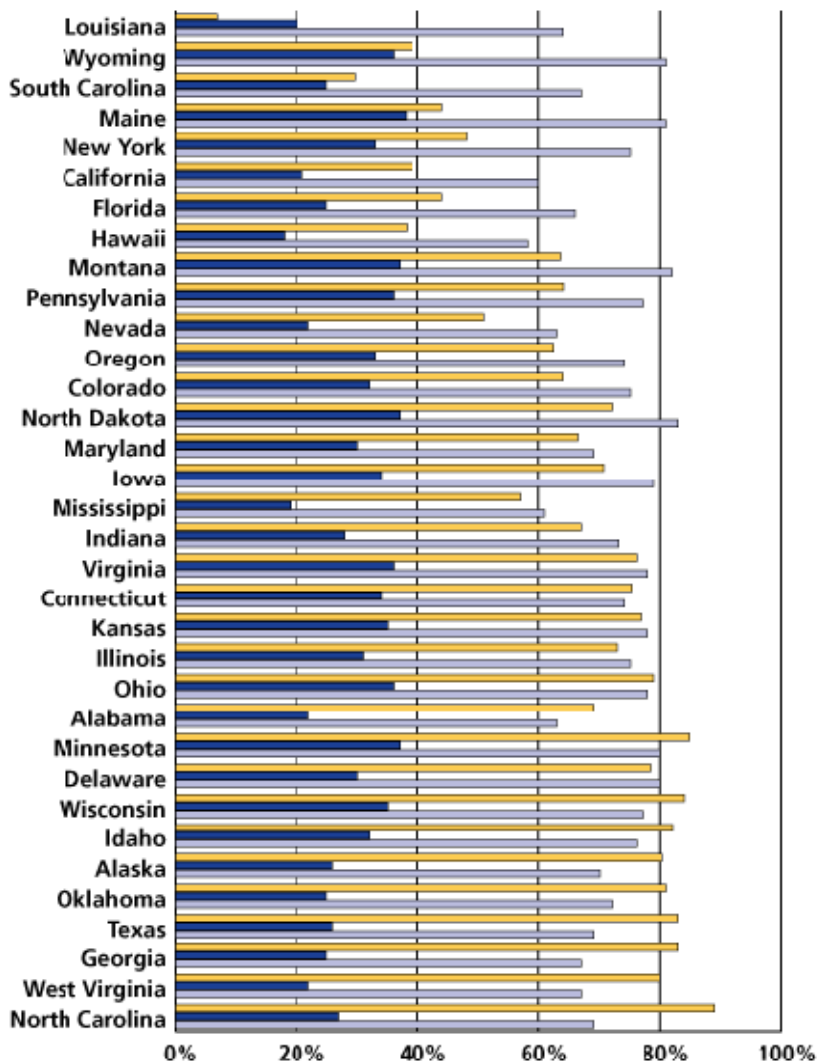


רשות ארצית למדידה והערכה בחינוך

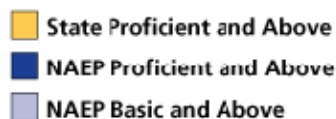
2005 State vs. NAEP Results 8th Grade



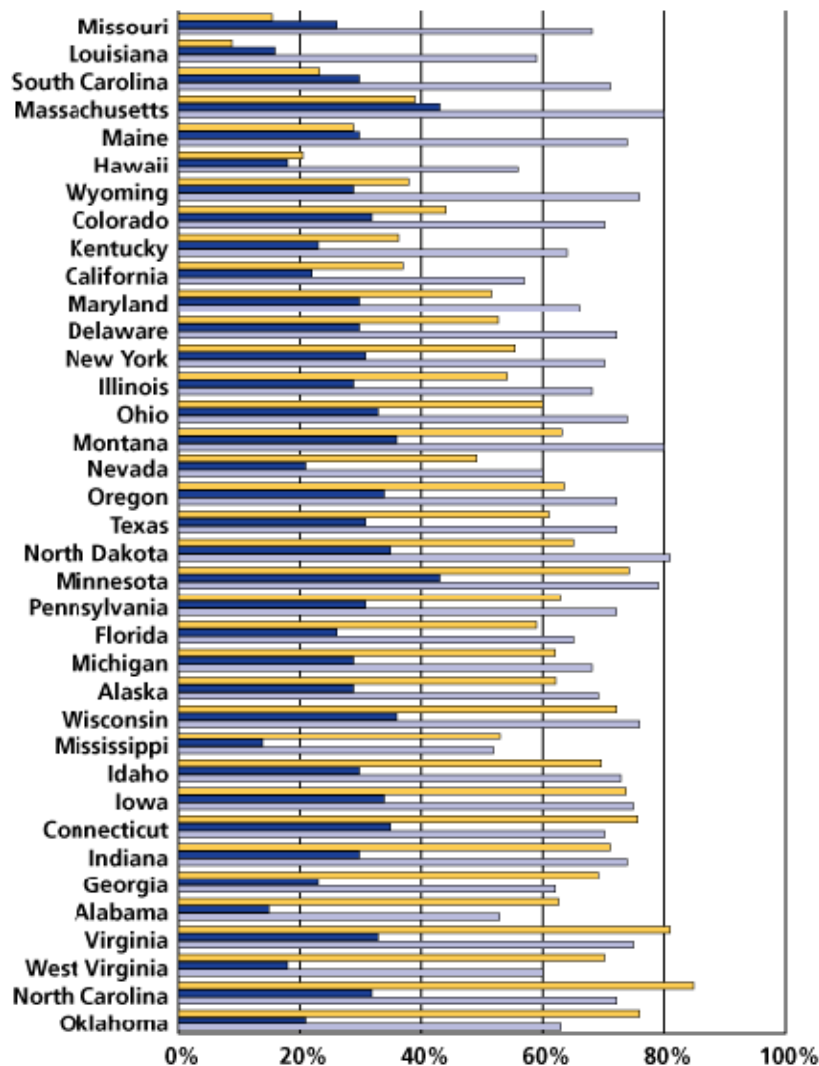
Reading



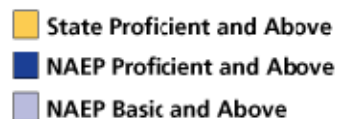
Percentage of Students



Math



Percentage of Students



הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיבוד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

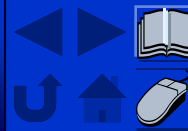
הערות לסיכום



רשות ארצית
למידה
והערכה
בחינוך

© 2007 # 35

Contrasts of Percent Proficient or above on NAEP and State Assessments (Grade 8 Mathematics) - From R. Linn



	NAEP	State Assessments
Missouri	21%	16%
Tennessee	26%	87%

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

הערות לסיכום



רשות ארצית
למידה
והערכה
בחינוך

בניית מבחנים תקפים

הקדמה



משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

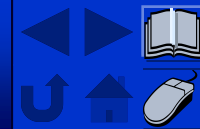
מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

הערות לסיכום



רשות ארצית
למידה
והערכה
בחינוך



תכנון שמשיא תוקף

- הגדרה רחבה של תחום הדעת הנמדד
- פיקוח על חשיפת פריטים וסוגי פריטים
- אלגוריתם ציינון יציב

תכנון שמשיא מהימנות

- שיפור רמת הדיוק בציינון
- אורך בחינה מספיק

תכנון שמשיא יעילות

- שימוש בטכנולוגיה: להעברה, לעיבוד, לניתוח וציינון

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

הערות לסיכום



רשות ארצית
למידה
והערכה
בחינוך

■ הגישה הרווחת כיום בהערכת הישגים היא של בניית מבחנים מבוססי סטנדרטים של תוכן (content standards)

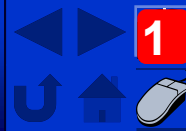
■ סטנדרטים (אם הוגדרו היטב) מספקים **מפרט** בהיר לפיתוח מבחן

■ **הלימה בין המבחן לבין הסטנדרטים** היא תנאי הכרחי להבטחת ההטמעה של הסטנדרטים בהוראה באופן המבוקש

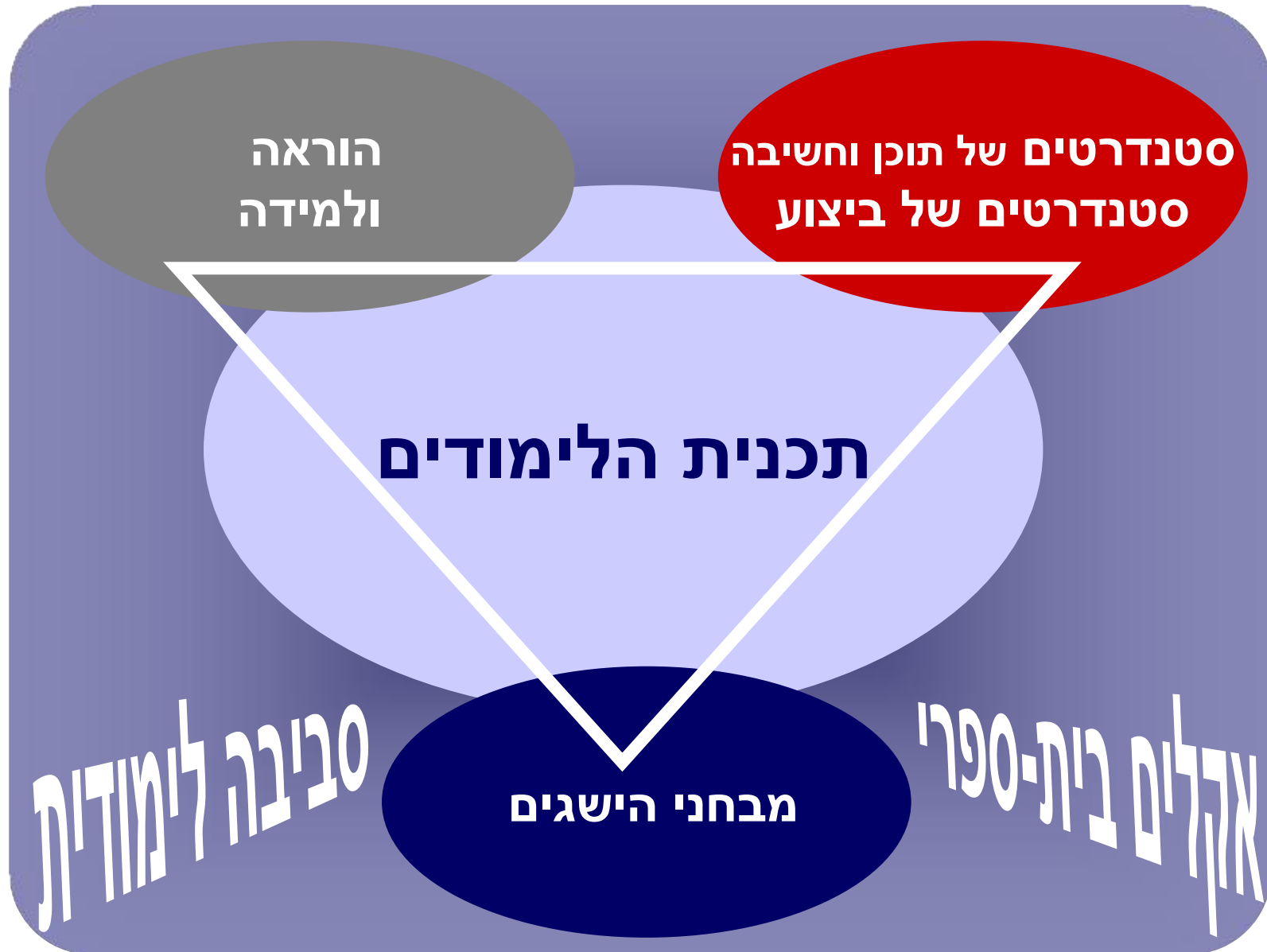
□ יש עדויות רבות לכך שלמבחנים עתירי סיכון יש השפעה על **מה המורים ילמדו**

□ סביר שהשפעת המבחנים על מערך ההוראה של המורה אף גדולה מזו של הסטנדרטים ולכן חייבת להיות הלימה בין השניים אחרת ייווצר **עיוות של "הצרת" תוכנית ההוראה בפועל**

הלימה בין מרכיבי היסוד של מערכת הלמידה



1



הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

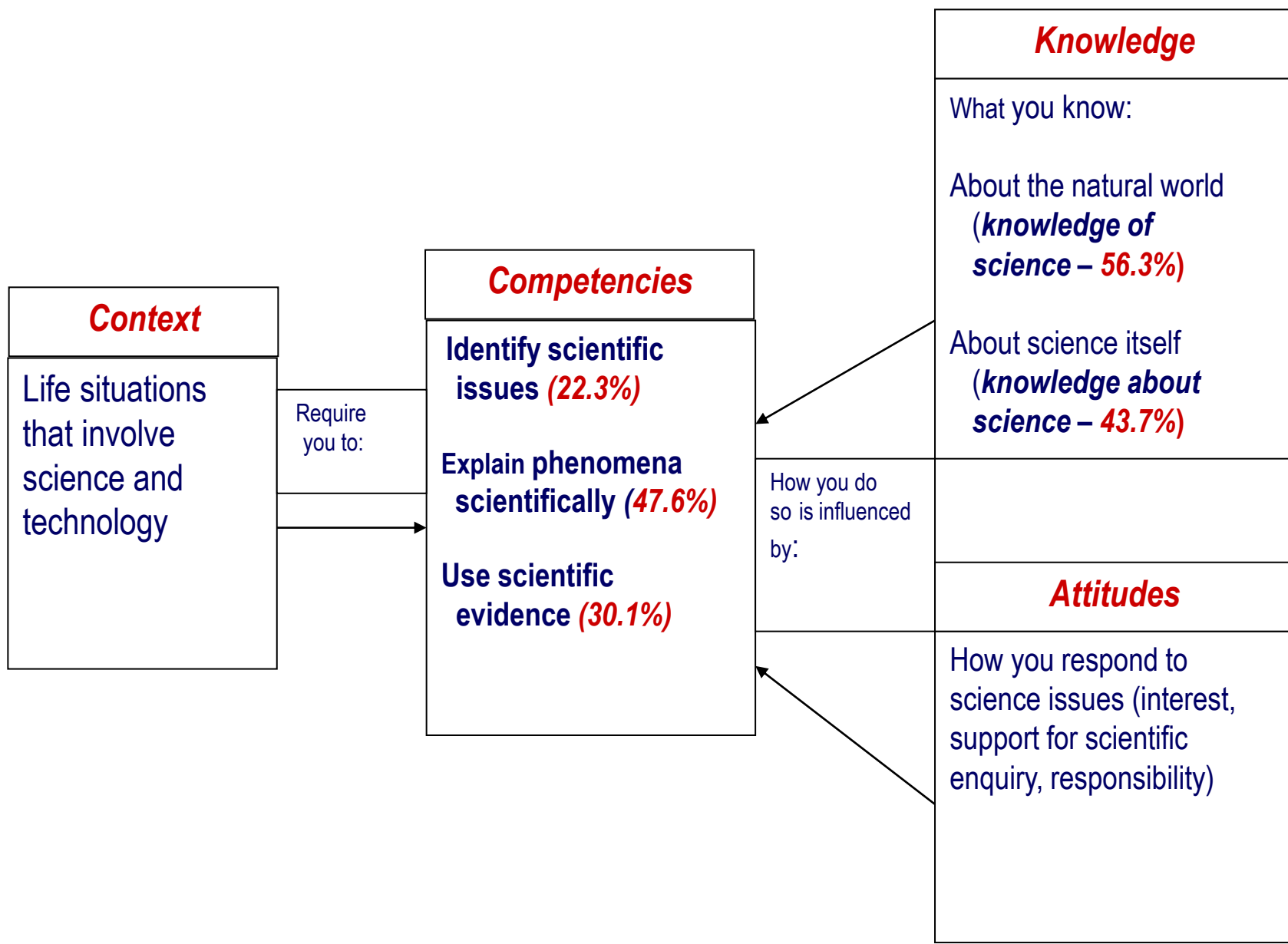
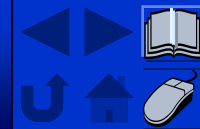
הערות לסיכום



רשות ארצית
למדידה
והערכה
בחינוך

© 2007 # 40

The PISA framework for scientific literacy



הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

הערות לסיכום

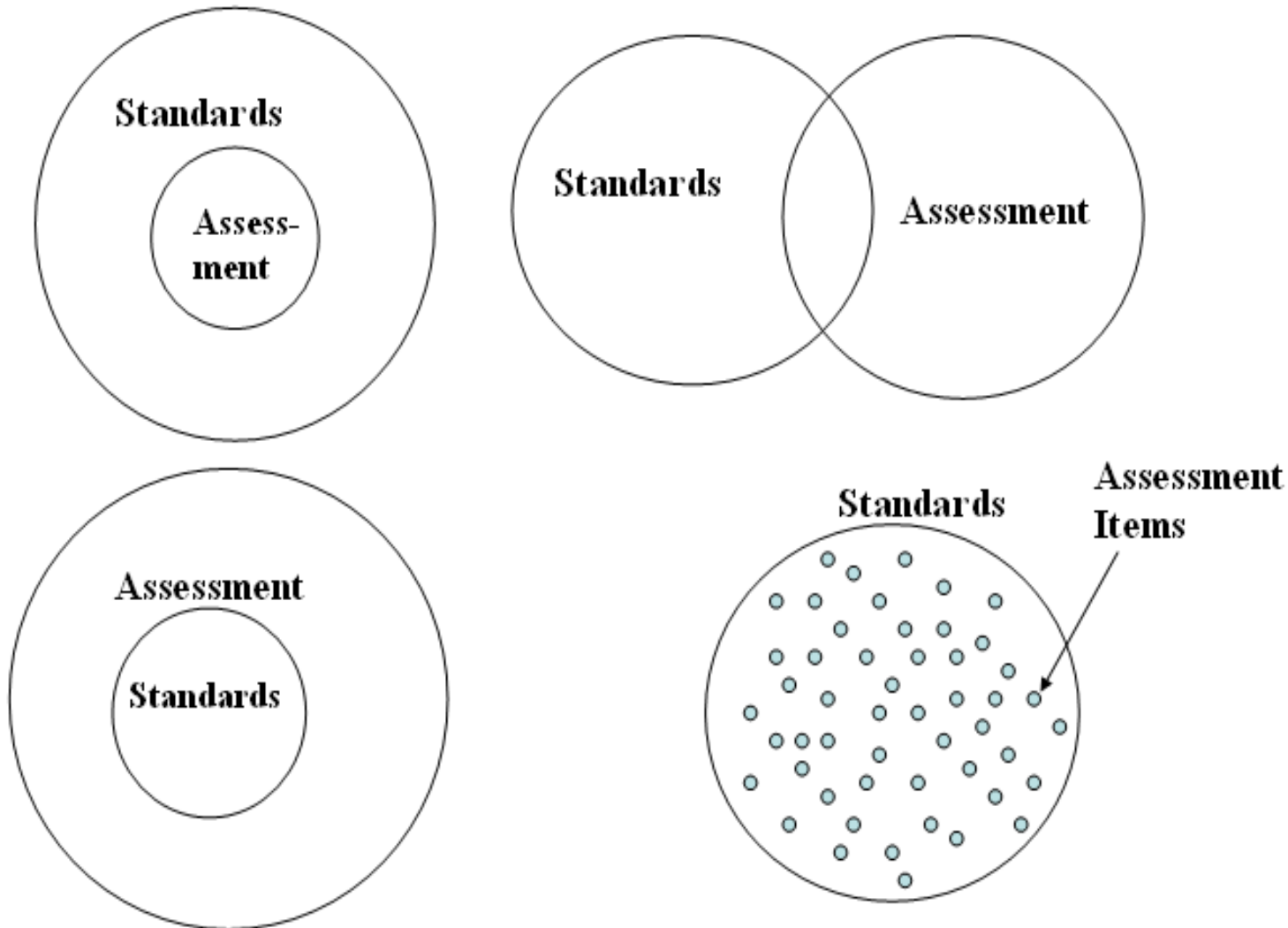


רשות ארצית
למידה
והערכה
בחינוך

© 2007 # 41

Degrees of Alignment

<http://www.wcer.wisc.edu/WAT/index.aspx>



הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

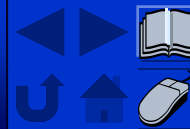
הערות לסיכום



רשות ארצית
למידה
והערכה
בחינוך

© 2007 # 42

רוחב יריעה למול אחידות



■ מבחן הישגים בכתה מוגבל בזמן (60-90 דקות)
ומכאן נגזרת **מגבלה על מספר השאלות לכל ילד**

■ במבחן כמו **מיצ"ב** – יש ערך לאחידות בנוסח המבחן
הניתן לכל ילד - ולכן מספר השאלות מוגבל וכתוצאה
מכך **מוגבלת דגימת התוכן**

■ **במבחנים בינלאומיים** כמו טימס ופיזה – בהם אין
חשיבות לבדיקה ברמת הפרט – נערכת דגימת תוכן
רחבה מאד וילדים שונים מקבלים נוסחים שונים של
מבחן כך שסה"כ מתקבלת **דגימת תוכן רחבה מאד**

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

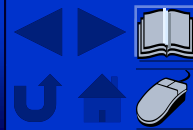
מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

הערות לסיכום



רשות ארצית
למדידה
והערכה
בחינוך



Improving Validity of Large-scale Tests: Universal Design and Student Performance

NCEO Technical Report 37

**Published by the National Center
on Educational Outcomes**

Prepared by:

Christopher J. Johnstone

December 2003

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

הערות לסיכום

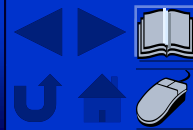


רשות ארצית
למידה
והערכה
בחינוך

© 2007 # 44

<http://www.education.umn.edu/nceo/OnlinePubs/Technical37.htm>

Summary of Major Validity Studies by Assessment Aspect



What is being measured?

1. Alignment with state standards
2. Construct definition (a) What is being measured
3. Construct definition (b) What students do on the test
4. Construct definition (c) Comparison with curriculum/what is taught in the classroom

How is it being measured?

5. Contaminations
6. Accommodations

Who is being tested?

7. School list completeness
8. Representation of non-participating schools
9. Student list completeness
10. Representation of absent students
11. Combined studies of population bias (studies 7, 8, 9, and 10)
12. Representation of excluded SD and LEP students

Are there better ways to analyze the data?

13. Direct estimation with minimal conditioning
14. Omission of subscales from primary analysis
15. Analyses using school as the unit
16. Estimation of item domain sampling error

How can appropriate uses be maximized and inappropriate uses minimized?

17. Meaning of “confirming state results”
18. Limits on NAEP’s capacity to evaluate state results
19. Evaluation of audience interpretations
20. Controls/supports for secondary analysis

What special information does valid measurement of trends require?

21. Bridge studies for changes in constructs, measurement, sample, administration
22. Estimation of multi-time-point trends

NAEP VALIDITY STUDIES: AN AGENDA FOR NAEP VALIDITY RESEARCH

<http://nces.ed.gov/pubs2003/200307.pdf>

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

הערות לסיכום



רשות ארצית
למידה
והערכה
בחינוך

© 2007 # 45

■ תרבות חדשה:

- יצירת אורה של "תודעת שיפור" תוך הפחתת ה"איומים" החיצוניים
- "מדידה בשירות למידה"
- שיתוף כל בעלי העניין

■ המדידה עצמה:

- התמקדות באיתור פערים ובניתוח מגמות לאורך זמן – יותר מאשר במדידת הישגים מוחלטים
- העלאת רמת המדידה (כולל כיול)
- לימוד מהניסיון הנצבר בעולם

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

הערות לסיכום



רשות ארצית
למדידה
והערכה
בחינוך



תודה!

שאלות? הערות?

הקדמה

משמעות מושג
ה"תוקף"

גורמים
המגבילים תוקף

כיצד קובעים
תוקף

תקפות
שיפור/ירידה

מדד ב"ת לבקרת
שיפור

בניית מבחנים
תקפים

הערות לסיכום



רשות ארצית
למידה
והערכה
בחינוך

© 2007 # 47