שילוב מטה-קוגניציה במבחני שמישות

פרידה נסאר
הטכניון – מכון טכנולוגי לישראל
fareda.nassar@gmail.com

רקפת אקרמן
הטכניון – מכון טכנולוגי לישראל
ackerman@ie.technion.ac.il

אברהם שטוב
הטכניון – מכון טכנולוגי לישראל
shtub@ie.technion.ac.il

Incorporating Metacognition into Usability Testing

Fareda Nassar
Technion – Israel Institute of Technology

Rakefet Ackerman
Technion – Israel Institute of Technology

Avraham Shtub
Technion – Israel Institute of Technology

Abstract

Usability testing is an important phase in the development of any software product, and of those used for learning, in particular. Usually, objective measures, like response time and success rates, are collected, together with global subjective measures, such as satisfaction. In this study, we adapted measures from the metacognitive approach to generate a comprehensive set of measures allowing more detailed analysis of users' subjective experience and work efficiency. We compared two user interfaces of a software tool designed to support project management learning in an academic course. In addition to measuring fluent work with the system and global satisfaction, the participants performed a set of focused tasks and rated their confidence in their success in each one. Triangulation of response time, success, and confidence was highly informative in exposing differences between the user interfaces, that were not exposed by global performance and satisfaction measures. Importantly, better outcomes were found when reliable confidence was experienced. This finding suggests that a product that eliminates overconfidence produces better outcomes. Overall, the study offers an applicable methodology for usability tests that takes into account metacognitive considerations for delving into the subjective experience and learning process of the users in more detail than done before.

Keywords: metacognition, overconfidence, e-learning, usability testing.

תקציר

פיתוח תוכנה שנועדה לליווי תהליך למידה מחייב שימוש במבחני שמישות המותאמים למטרה זו. בדרך כלל, במבחני שמישות נאספים מדדים אובייקטיבים, כמו זמן ביצוע מטלה ואחוזי הצלחה, ובנוסף נאספים מדדים סובייקטיבים כלליים, כמו שביעות רצון. כדי להתעמק יותר בתהליך הלמידה ובויסות המאמצים הקוגנטיביים, במחקר הנוכחי בדקנו את התרומה של מדדים מתחום המטה-קוגניציה להערכת שימושיות התוכנה. הבדיקה נעשתה על ידי השוואה בין שני ממשקים של אותה תוכנה, שנועדה לתמוך בלמידה של ניהול פרוייקטים בקורס אקדמי. מעבר לניהול פרויקט באופן שוטף, הנבדקים ביצעו סט של משימות נקודתיות, כגון העסקת עובד והקצאת תקציב למשימה. עבור כל משימה הנבדקים התבקשו לדרג ביטחון בתשובה שסיפקו. דירוג הביטחון ביחד עם זמן ביצוע מטלה ונכונות התשובה אפשרו יצירת סט מדדים מורחב להערכת הלמידה. מדדי הביצוע ושביעות הרצון הכלליים לא הצביעו על הבדל בין שני הממשקים, בעוד המדדים המפורטים הצביעו על הבדלים רבים. הממצא החשוב הוא שנמצא כי קיים קשר בין מידת ביטחון היתר לבין הביצוע, כך שבממשק שהישלה את המשתמשים פחות הביצוע היה טוב יותר. המחקר

מציע שילוב של מדדים מטה-קוגניטיביים במבחני שמישות לבחינה והערכה יותר מעמיקה של השימוש בתוכנות מחשב המלוות למידה.

מילות מפתח: מטה-קוגניציה, ביטחון יתר, למידה באמצעות מחשב, מבחני שמישות, טכנולוגיות למידה.

מבוא

במהלך פיתוח תוכנות המלוות תהליך למידה, יש לבצע בדיקות שמישות המותאמות למטרה זו. על פי ההגדרה הכי נפוצה שלה, שמישות היא המידה שבה מוצר מאפשר לקבוצת משתמשים נתונה להשיג מטרות קבועות מראש בצורה יעילה ומשביעת רצון (ISO 9241-11, Guidance on Usability, 1998) מדדי שמישות מתחלקים למדדים אובייקטיבים ומדדים סובייקטיביים. מדדים אובייקטיבים כוללים למשל מידת ההצלחה ומשך ביצוע המשימה. לעומת זאת, מדדים סובייקטיביים מבטאים את חויית המשתמש, כמו דירוג קלות השימוש במוצר ועמדות כלפי המוצר (Hornbaek, 2006). Hornbaek (2006) הראה קורלציה נמוכה בין המדדים האובייקטיבים והמדדים הסובייקטיביים, מה שמעיד על חשיבות השילוב של שני סוגי המדדים כדי לקבל תמונה מלאה של שמישות המוצר. מדד סובייקטיבי נוסף הנפוץ בספרות של מבחני שמישות הינו ה-System Usability Scale (SUS). זהו שאלון המספק מבט סובייקטיבי גלובלי לגבי שמישות של מוצר. התוצר של SUS הינו ציון הנע בין 0 ל 100 שמאפשר השוואה של שימושיות בין מוצרים שונים. במחקר הנוכחי אנחנו מציעים להוסיף למבחני שמישות קבוצת מדדים הנגזרת מתחום המטה-קוגניציה ומאפשרת התעמקות מפורטת יותר בחווית המשתמש תוך כדי תהליך הלמידה בעזרת מוצר תוכנה המלווה קורס אקדמי.

המחקר בתחום המטה-קוגניציה עוסק בניהול ההשקעה של משאבים קוגניטיביים בזמן ביצוע משימות כגון למידה, מענה לשאלות ידע ופתרון בעיות. התהליך המטה-קוגניטיבי מחולק לניטור ושליטה. הניטור כולל הערכה סובייקטיבית של איכות ביצוע המשימה ואילו השליטה היא ההחלטה שמתקבלת בעקבות הניטור. לדוגמה, בתחום הלמידה, בזמן למידת טקסט התלמיד מעריך את רמת הידע שצבר ולפי זה מקבל החלטה לגבי המשך השקעת הזמן בלימוד (Nelson & Narens, 1990).

הניטור המטה-קוגניטיבי שבו התמקדנו במחקר הנוכחי הינו ביטחון בתשובה שהנבדק מספק (בסקלה של 0-100%). קיימות שתי תופעות נפוצות בהשוואה בין מידת הביטחון לרמת ההצלחה בפועל: ביטחון יתר וביטחון חסר. תופעת ביטחון היתר מתגלה בעת שממוצע הביטחון שהנבדק מספק גבוה מאחוז התשובות הנכונות בפועל במבחן. לעומת זאת, ביטחון חסר מתגלה בעת שממוצע הביטחון נמוך מאחוז התשובות הנכונות. התופעה הנפוצה יותר היא ביטחון יתר (e.g., Ackerman & Zalmanov, 2012). מחקרים הראו כי דיוק הניטור המטה-קוגניטיבי משפיע על קבלת החלטות לגבי ויסות תהליך הלמידה ועל איכות הביצוע במבחן בתום הלמידה (e.g., Thiede, Anderson, &Therriault, 2003). בנוסף, לתרומת המטה-קוגניציה לתחום הלמידה ניתן להבחין בספרות בתרומה בתחום אינטראקציות אדם-מחשב. Vu ועמיתיו (2000) הראו כי ההערכות הסובייקטיביות של הנבדקים לגבי רמת המומחיות בשימוש בתוכנות מחשבים מנבאות בצורה אמינה יותר את איכות השימוש בתוכנות בפועל מאשר תדירות השימוש. כמו כן, מחקרים הראו תופעת ביטחון יתר עקבית בעת למידת טקסטים ממסך מחשב לעומת למידה מנייר (Ackerman & Goldsmith, 2011; Ackerman & Lauterman, 2012; Lauterman & Ackerman, 2014). מחקרים קודמים שאספו ביטחון בהצלחה כחלק ממדדי שמישות, התיחסו רק להשוואה ברמת הביטחון בין תנאים שונים (Vu et.al, 2000), אך לא דנו בביטחון יתר ובמדדים נוספים המקובלים בספרות המטה-קוגנטיבית ומאפשרים התעמקות רבה יותר בתהליך הלמידה.

במחקר הנוכחי נערכה השוואה של שמישות בין שני ממשקים למשתמש עבור מוצר תוכנה אחד שנועד לתמיכה בלימוד הנושא של ניהול פרויקטים, הידוע במורכבותו הרבה. המשתתפים במחקר נבחרו להיות בעלי שתי רמות ידע בתחום ניהול פרויקטים. מחקרים מראים כי בעיות שמישות שונות מתגלות אצל קבוצות נבדקים ברמות ידע שונות (e.g., Sauer, Seibel & Ruttinger, 2010). המשתתפים כללו סטודנטים לפני ואחרי הקורס האקדמי "תכנון פרוייקטים וניהולם", אך אף סטודנט לא התנסה עם הממשקים של תוכנת לימוד ניהול פרוייקטים לפני המחקר.

השערות המחקר

ההשערה הראשונה הייתה שממשק שמאפשר ניטור מטה-קוגניטיבי יותר אמין יניב ביצוע יותר טוב. במילים אחרות, הממשק שיאפשר מדדים מטה-קוגניטיביים יותר מדוייקים יניב למידה יותר טובה. ההשערה השניה היתה שמדדים מטה-קוגנטיביים מאפשרים לחשוף הבדלי שמישות שמדדים

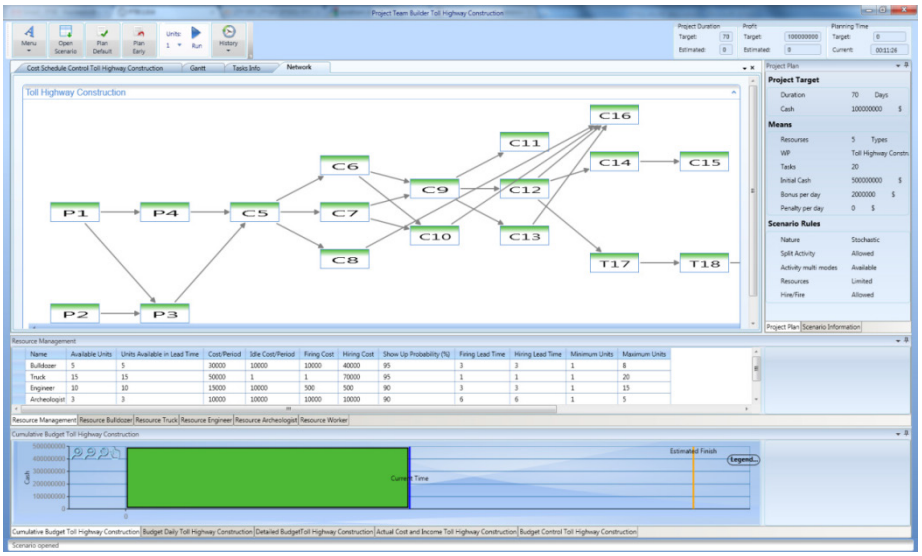
מקובלים בתחום לא מגלים. כך מפתחי התוכנה יכולים לגלות בעיות בממשק למשתמש שכדאי לשפר ולהסיק איזה ממשק מתאים יותר לליווי למידה.

מתודולוגיה

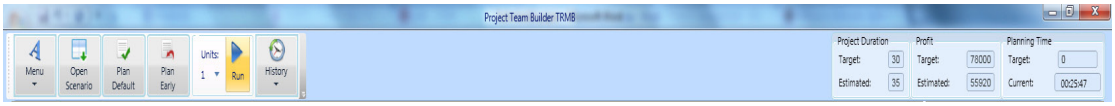
נבדקים. במחקר השתתפו 99 סטודנטים מהפקולטה להנדסת תעשייה וניהול (46% נשים). מחצית הסטודנטים סיימו את הקורס "תכנון פרויקטים וניהולם" בפקולטה להנדסת תעשיה וניהול והם היוו את קבוצת בעלי הידע הקודם ($N = 49$). המחצית השנייה כללה סטודנטים שעוד לא לקחו את הקורס והם היוו את קבוצת חסרי הידע הקודם.

חומרים. בנוסף לשני הממשקים של התוכנה שנועדה לתמיכה בלימוד נושא ניהול פרוייקטים היו גם א) דף הדרכה כללית לשימוש בתוכנה, ב) דף לרישום תשובות למשימות שניתנו לנבדקים שכלל גם כן סקלה לסימון ביטחון בתשובה וסקלה לציון קלות ביצוע משימה. ג) סקלתSUS להערכה כללית של שמישות התוכנות.

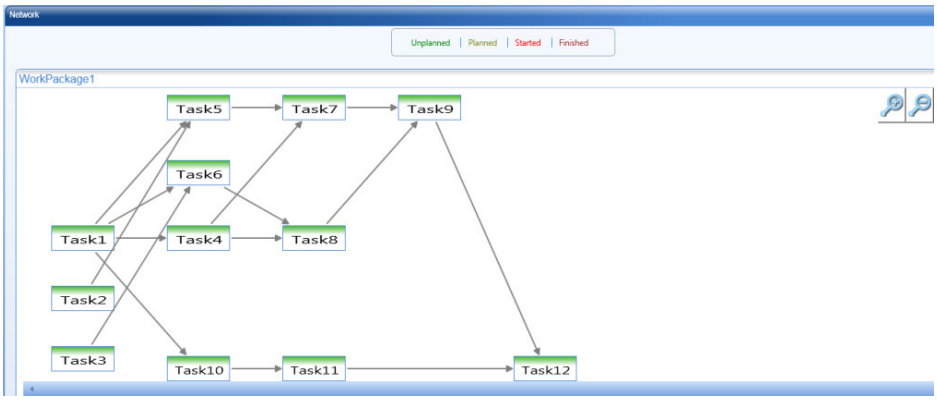
הליך. ההרצה של הניסוי התבצעה באופן יחידני. כל נבדק התנסה בשימוש בשני הממשקים למשתמש. הממשקים שונים זה מזה באופן הצגת המידע כך שבממשק הראשון כל סוגי המידע מוצגים בצורה במקביל במסך הראשי (ראו איור 1), דבר המאפשר סרגל כלים פחות עמוס (ראו איור 2). לעומת זאת, בממשק 2 הנבדק צריך לנווט בין מסכים שונים על מנת לקבל סוגי מידע שונים (משימות, משאבים, כספים). לדוגמא, איור 3 מראה את המשימות על כל המסך ועל מנת לעבור לסוג מידע אחר (משאבים, כספים, תרשימים) צריך לעבור למסך אחר ע"י ניווט דרך סרגל הכלים הראשי בממשק. דבר שהופך את סרגל הכלים ליותר עמוס (ראו איור 4) ומצריך להפעיל את הזיכרון של המשתמשים בניווט.



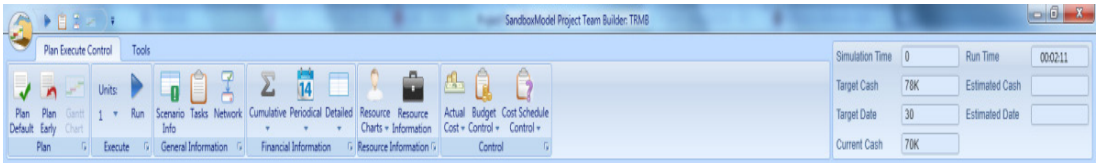
איור 1. המסך הראשי בממשק 1



איור 2. סרגל הכלים בממשק 1



איור 3. המסך הראשי בממשק 2 – בעת הצגת המשימות



איור 4. סרגל הכלים בממשק 2

הנבדק ביצע בכל ממשק עשר משימות אשר נקבעו מראש בהסתמך על ניתוח משימות (Task analysis) לשני הממשקים. המשימות חולקו לשני סטים, כך שסט ראשון כלל משימות שלאור ניתוח המשימות היו צפויות להיות קלות יותר לביצוע באמצעות הממשק הראשון. הסט השני כלל משימות שצפויות להיות קלות יותר לביצוע באמצעות ממשק 2. עבור כל משימה אספנו את המדדים הבאים:

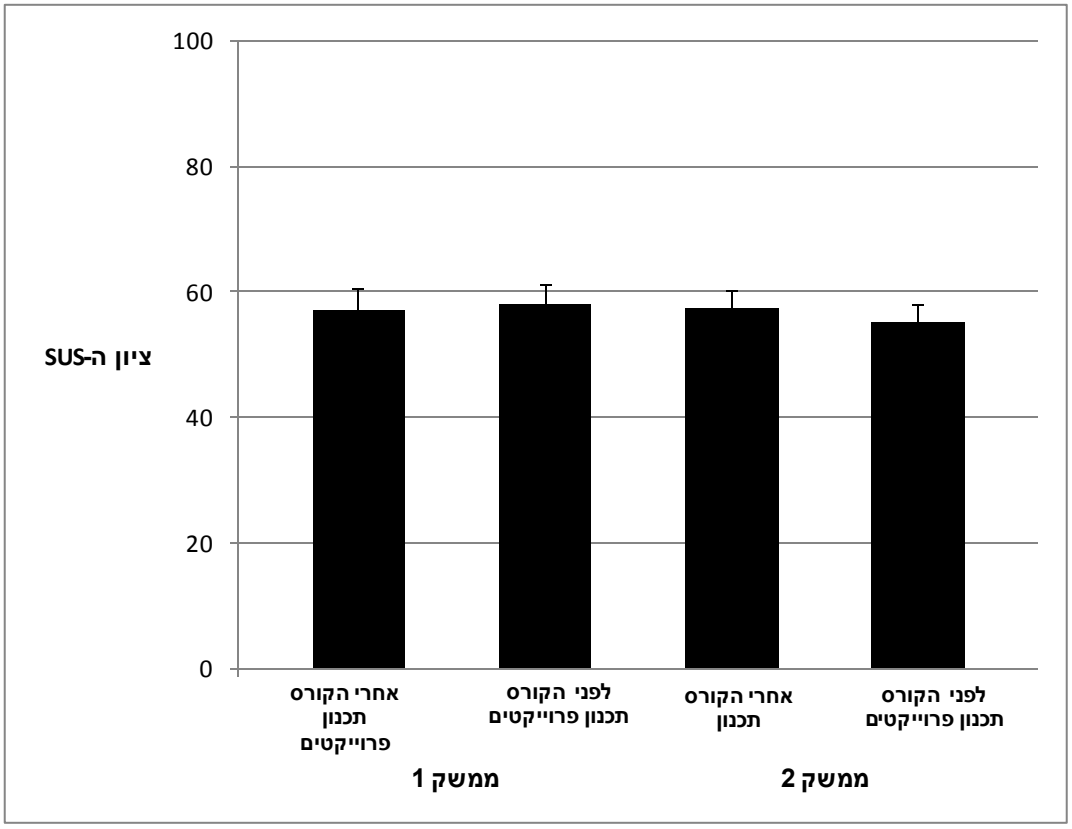
- אחוז הצלחה – נכונות התשובה שהנבדק מספק בעקבות ביצוע המשימה.
- זמן תגובה – ביצוע המשימה (בשניות).
- ביטחון של הנבדק בתשובה שהוא סיפק (סקלה 0-100).
- קושי נתפס – דירוג של קושי ביצוע המשימה (סקלה 1-7)

מייד אחרי שהנבדק סיים לבצע את המשימות בעזרת הממשק הראשון הוא התבקש למלא את שאלון ה- SUS, המסכם את החויה שלו עם הממשק זה, ולעבור לביצוע אותן משימות בעזרת ממשק 2 ולמלא את סקלת ה-SUS לגבי השימוש בממשק 2. סדר השימוש בממשקים אוזן מעבר לנבדקים. המפגש עם כל נבדק נמשך כשעתיים.

תוצאות

שילוב מדדים מטה-קוגנטיביים ואובייקטיביים בניתוח מפורט אפשר איתור הבדלים בין הממשקים לעומת שאלון ה SUS שלא הצביע על הבדלים בין שני הממשקים:

SUS. הציונים של ה SUS נעו בין 0-100. ניתוח שונות נעשה על ציוני ה-SUS שהתקבלו עבור שני הממשקים. לא נמצאו הבדלים בין שני הממשקים וגם לא בין שתי קבוצות הנבדקים השונות ברקע שלהם בניהול פרוייקטים, $F < 1$. ראו איור 5.



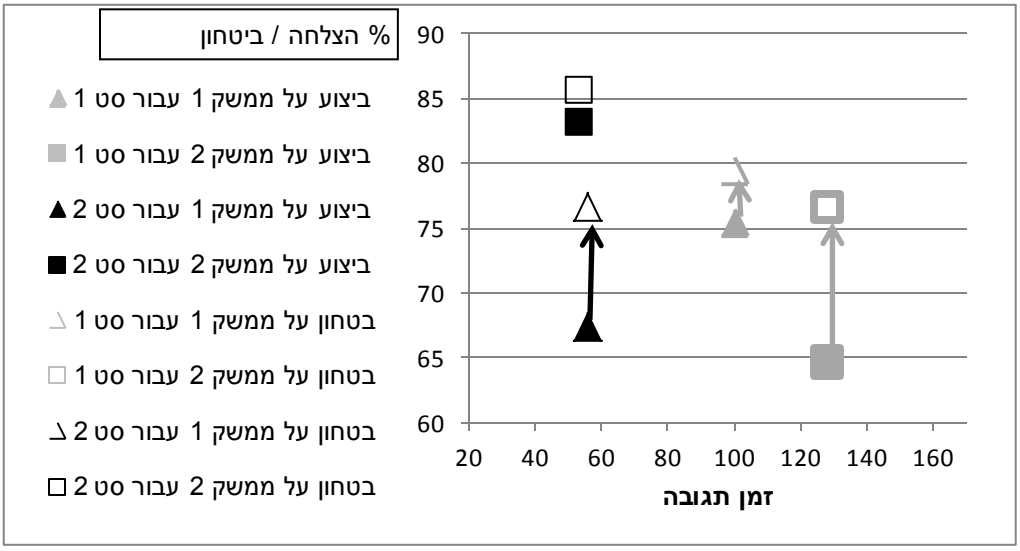
איור 5. תוצאות ה-SUS

מדדים אובייקטיבים ומטה-קוגניטיביים:

בנוסף למדדים שנאספו ישירות, נכונות התשובה, זמן תגובה, ביטחון, קלות נתפסת, בניתוח הזה התווספו מדדים מחושבים המתבססים על המדדים הישירים: א) יעילות: אחוז תשובות נכונות שנצברו לדקה ב) ביטחון יתר: הפער בין ממוצע הביטחון בתשובה לבין אחוז התשובות הנכונות בפועל ג) רזולוציה: קורלציה תוך נבדקית בין הביטחון בתשובה לבין נכונות התשובה.

המדדים האובייקטיבים הצביעו על הבדלים בביצוע בין שני הממשקים עבור כל סט, תוצאות התומכות בחלוקה לשני סטים.

אחוז הצלחה. עבור כל נבדק חישבנו את אחוז התשובות הנכונות שלו מעבר לכל 9 המשימות (0-100). לא היה הבדל באחוז ההצלחה בין שני הממשקים, $F < 1$. לעומת זאת, כן נמצא הבדל באחוז ההצלחה בין שני הסטים כך שאחוז ההצלחה בסט הראשון ($M = 70.0$; $SD = 44.5$) היה קטן יותר מאחוז ההצלחה בסט השני ($M = 75.2$; $SD = 42.3$), $F(1,196) = 6.3$, $p < .01$. האינטראקציה המובהקת של הממשק עם הסט $F(1,196) = 51.9$, $p < .01$, הראתה כי אחוז ההצלחה בסט הראשון היה גבוה יותר בממשק הראשון, $t(98) = 4.1$, $p < .01$, מצד שני האפקט היה הפוך עבור הסט השני, $t(98) = 5.8$, $p < .01$. ראו איור 6. ההבדלים האלה בין הסטים מוכיחים כי חלוקת המשימות לשני סטים תאמה את ניתוח המשימות.



איור 6. גרף המראה את הביטחון, אחוז ההצלחה וביטחון היתר על ציר הזמן

זמן תגובה. זמן התגובה נמדד בשניות. הנבדקים ביצעו את המשימות בזמן קצר יותר בעזרת הממשק הראשון ($M = 78.1$; $SD = 44.0$) לעומת ממשק 2 ($M = 90.9$; $SD = 59.3$), $F(1,196) = 4.9$, $p = .03$. בנוסף, נמצא כי האפקט של הסט היה מובהק, כך שזמן התגובה בסט הראשון קצר יותר מזמן התגובה בסט השני, $F(1,196) = 334.9$, $p < .01$. האינטראקציה המובהקת $F(1,196) = 23.7$, $p < .01$, הראתה כי זמן התגובה בסט הראשון היה קצר יותר על הממשק הראשון, $t(98) = 3.8$, $p < .01$, מצד שני לא נמצא אפקט עבור הסט השני, $t(98) < 1$. ראו איור 6.

יעילות. מדד היעילות חושב על ידי חלוקת אחוז ההצלחה בזמן הביצוע הממוצע למטלה. נמצא הבדל מובהק ביעילות בין הסטים. הנבדקים היו פחות יעילים בביצוע הסט הראשון ($M = 70.8$; $SD = 77.3$) לעומת הסט השני ($M = 222.6$; $SD = 335.8$), $F(1,196) = 271.3$ $p < .01$. האינטראקציה המובהקת בין הסט לממשק, $F(1,196) = 12.6$, $p < .01$, מצביעה כי עבור הסט הראשון הנבדקים היו יותר יעילים על הממשק הראשון, $t(98) = 4.2$, $p < .01$. לעומת זאת, עבור הסט השני הנבדקים היו יותר יעילים על ממשק 2. ראו איור 6. התוצאות עבור מדדי הביצוע האובייקטיבים, כלומר, אחוז הצלחה, זמן תגובה ויעילות היו טובות יותר עבור הסט המתאים לכל ממשק ביחס לסט השני.

גם המדדים המטה-קוגניטיביים הראו הבדלים בין הממשקים לפי הסטים, כך שהממשק שאפשר ניטור מטה-קוגניטיבי יותר אמין הביצוע על אותו ממשק היה יותר טוב:

ביטחון בתשובה (0%-100%). נמצא כי הביטחון על הממשק הראשון ($M = 77.9$; $SD = 16.3$) היה נמוך יותר מהביטחון על ממשק 2 ($M = 81.1$; $SD = 17.7$), $F(1,196) = 4.9$, $p = .03$. לא נמצא הבדל בביטחון בין שני הסטים. לעומת זאת, האינטראקציה בין הממשק לסט נמצאה מובהקת $F(1,196) = 64.4$, $p < .01$ כך שעבור הסט הראשון ההבדל בביטחון בשני הממשקים לא היה מובהק $t(98) = 1.6$, $p = .10$. מצד שני, עבור הסט השני הביטחון על הממשק הראשון היה נמוך יותר מהביטחון על ממשק 2, $t(98) = 6.1$, $p < .01$. ראו איור 6.

ביטחון יתר. מדד ביטחון היתר נמדד ע"י חישוב הפער בין אחוז ההצלחה לביטחון. לא נמצאו אפקטים עיקריים, אך האינטראקציה בין הסט לממשק היתה מובהקת, $F(1,196) = 18.1$, $p < .01$. עבור הסט הראשון ביטחון היתר על הממשק הראשון ($M = 4.0$; $SD = 20.2$) היה נמוך יותר לעומת ממשק 2 ($M = 12.0$; $SD = 21.5$), $t(98) = 3.3$, $p < .01$. לעומת זאת, עבור הסט השני נמצא אפקט הפוך, $t(98) = 2.5$, $p < .05$, כך שביטחון היתר על הממשק הראשון ($M = 9.1$; $SD = 24.3$) היה גבוה יותר מביטחון היתר על ממשק 2 ($M = 2.5$; $SD = 17.9$). ראו איור 6. כלומר, תחושת הביטחון היתה מכוילת באופן מדויק יותר עבור הסט המתאים לממשק.

רזולוציה. מכיוון שחישוב הרזולוציה על ידי קורלציה תוך-נבדקית מותנה בשימוש בלפחות 6 פריטים בעוד שכאן יש 4-5 פריטים בכל סט, חישבנו את הרזולוציה מעבר לכל 9 המשימות. התוצאות הראו אפקט מובהק שולית של הממשק $F(1,196) = 3.04$, $p = 0.08$. הרזולוציה של הנבדקים נטתה להיות נמוכה יותר על הממשק הראשון ($M = .34$; $SD = .63$) לעומת ממשק 2

($M = .49$; $SD = .52$). רזולוציה גבוהה מעידה על כך שהנבדק בטוח יותר במשימות שבהן גם הצליח בפועל יותר, כלומר מבחין היטב בין טיב הלמידה במשימות השונות. הרזולוציה הגבוהה יותר בממשק 2 יכולה לרמוז על עיבוד עמוק יותר בממשק זה. חשוב לציין שבכל התנאים הרזולוציה היתה שונה מאפס באופן מובהק, $p < .01$.

קושי השימוש. הנבדקים מצאו את ממשק 1 יותר קשה לשימוש ($M = 3.1$; $SD = 1.1$) לעומת ממשק 2, ($M = 2.9$; $SD = 1.2$), $F(1,196) = 3.4$, $p = .07$. בנוסף, הנבדקים מצאו את הסט הראשון יותר קשה לשימוש, $F(1,196) = 45.9$, $p < .01$, האינטראקציה בין סט לממשק נמצאה מובהקת $F(1,196) = 78.1$, $p < .01$. עבור הסט הראשון, ממשק 1 נמצא קל יותר לשימוש. לעומת זאת, עבור הסט השני ממשק 1 נמצא יותר קשה לשימוש. גם עבור המדד הזה, התוצאות תואמות לחלוקת לשני סטים.

דיון

תוצאות המחקר הראו כי ה-SUS אינו יכול לשמש כמדד יחיד לבחינת שמישות מכיוון שכן התגלו הבדלים באחוז ההצלחה, זמן ביצוע, דירוג קלות הביצוע ובביטחון יתר בין שני הממשקים, בעוד הוא לא הראה הבדלים בין הממשקים.

המחקר מדגיש את החשיבות של בחינה מפורטת של מדדים אובייקטיבים וסובייקטיביים בעת עריכת מבחן שמישות. הגישה המטה-קוגניטיבית מספקת מדדים המקשרים בין המדדים האובייקטיבים לבין המדדים הסובייקטיביים על ידי בחינת מדד הביטחון ומדדים הנובעים משילוב שלו עם הצלחה בפועל ועם זמן תגובה. מהמחקר עולה שניתוח מפורט, לאור סט מדדים מגוון המתייחס לביצוע משימות ספציפיות, עשוי לחשוף בעיות שמישות, שלא היו מתגלות בשימוש בסקאלות כלליות להערכה סובייקטיבית, כגון ה-SUS שמקובלת מאוד בתחום מבחני השמישות.

אחד היתרונות החשובים של שיטת בחינת השמישות שמוצעת במחקר הזה היא היכולת לאסוף מדדים אובייקטיבים וסובייקטיביים מפורטים מבלי להפריע לנבדק ומבלי להשפיע על האינטראקציה בין הנבדק לממשק, בעיה שהוצגה במחקרים קודמים (Van den Haak et al., 2003). דבר המאפשר הערכה אמינה של השימוש בתוכנה.

ראוי להדגיש את הקשר שנמצא בין ההטיה בביטחון לבין אחוז ההצלחה עבור אותו סט של משימות. החשיבות של הממצא היא שעבור נבדקים שביצעו פחות טוב סט מסוים של משימות נוצרת אשליה שאחוז ההצלחה יהיה דומה לסט מטלות בו אחוז ההצלחה בפועל היה טוב יותר. הסיכון הוא בהשפעת האשליה הזאת על שימושים עתידיים בממשק. משתמש שהושלה שביצע טוב, לא יבדוק את הביצוע שלו שנית וגם לא יחשוב שיש צורך בלמידה או הדרכה נוספת לשימוש עתידי בממשק.

במבט כללי יותר, מחקרים שעסקו בתרומה של המדדים המטה-קוגניטיביים והשפעת הדיוק של מדדים אלה על הליך הלמידה עסקו עד עכשיו במטלות זיכרון של צמדי מילים, למידה מטקסטים, ומענה על שאלות ידע כללי (e.g., Ackerman &Goldsmith, 2008; Metcalfe & Finn, 2008; Thiede et al., 2003) או בפתרון בעיות מתמטיות ומילוליות (ראו סקירה Ackerman & Thompson, in press). במחקר הנוכחי יישמנו את אותה מתודולוגיה לבחינת הביצוע של המשתמשים בעת שימוש בתוכנה לימודית. אנו מקווים לראות בעתיד מחקרים נוספים המיישמים שיטה זו לצורך מבחני שמישות ולצורך חקר ביצוע מטלות מורכבות בכלל.

מקורות

Ackerman, R., & Goldsmith, M. (2008). Control over grain size in memory reporting – with and without satisficing knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1224-1245.

Ackerman, R., & Goldsmith, M. (2011). Metacognitive regulation of text learning: On screen versus on paper. *Journal of Experimental Psychology: Applied*, *17*(1), 18-32.

Ackerman, R., & Lauterman, T. (2012).Taking reading comprehension exams on screen or on paper? A metacognitive analysis of learning texts under time pressure. *Computers in Human Behavior*, *28*, 1816-1828.

Ackerman, R. & Thompson, V. (in press). Meta-reasoning: What we can learn from meta-memory. To appear in A. Feeney, & V. Thompson (Eds.), *Reasoning as Memory*. Hove, UK: Psychology Press.

Ackerman, R., & Zalmanov, H. (2012). The persistence of the fluency-confidence association in problem solving. *Psychonomic Bulletin & Review*, *19*(6), 1187-1192.

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies, 64*(2), 79-102.

ISO/IEC. "9241-11 Ergonomic Requirements for Of Tice Work with Visual Display Terminals (VDT) s-Part II Guidance on Usability," ISO/IEC 9241-11,1998 (E).

Lauterman, T., & Ackerman, R. (2014). Overcoming screen inferiority in learning and calibration. *Computers in Human Behavior*, *35*, 455-463.

Metcalfe, J., & Finn, B. (2008).Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review*, *15*(1), 174-179.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *The Psychology of Learning and Motivation, 26*, 125-141.

Sauer, J., Seibel, K. &Ruttinger, B. (2010).The influence of user expertise and prototype fidelity in usability tests. Applied Ergonomics, 41, 130-140.

Thiede, K. W., Anderson, M., &Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology*, *95*(1), 66-73.

Van den Haak, M., De Jong, M., &Schellens, P. J. (2003). Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behavior & Information Technology*, *22*(5), 339-351.

Vu, K.P.L., Hanley, G. L., Strybel, T. Z., & Proctor, R. W. (2000). Metacognitive processes in human–computer interaction: Self-assessments of knowledge as predictors of computer expertise. *International Journal of Human–Computer Interaction*, *12*, 43-71.