

תורת ההכללה

נעמי גפני
בלה פבזנר

יוני 2012



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

דוח מרכז 381
ISBN:978-965-502-165-3

תורת ההכללה

נעמי גפני, בלה פבזנר

יוני 2012

**ייעוץ: פרופ' רות בייט-מרום, האוניברסיטה הפתוחה
ד"ר אבנר כספי, האוניברסיטה הפתוחה**

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

תוכן העניינים

1.	דברי פתיחה	5
2.	מבוא	5
3.	מסגרת לאמידה של רכיבי טעות ומקדמי הכללה	7
4.	שלבים בביצוע מחקר הכללה	8
4.1	החלטה על מערך התצפיות	8
4.2	דגימת שטחות (מערך האמידה)	11
4.3	הגדרה של מטרת המדידה	13
4.4	אמידה של טעות התקן של המדידה (דיוק המדידה)	14
4.5	חישוב של מקדמי הכללה (הערכה של מערך המחקר)	15
4.6	ביצוע של מחקרי אופטימיזציה	16
5.	דוגמאות נוספות	18
5.1	דוגמה למערך תצפיות עם שטחה קבועה	18
5.2	דוגמה למערך תצפיות עם מערך שטחות מקונן	19
6.	סיכום	20
7.	תוכנות זמינות לניתוח הכללה	21
22	ביבליוגרפיה	
23	שאלות תרגול	
24	תשובות לשאלות תרגול	

1. דברי פתיחה

בכתיבת פרק זה הנחנו כי לקורא אין ידע נרחב בפסיכומטריקה, אבל יש לו היכרות עם תורת המבחנים הקלאסית ועם המושגים: מהימנות, ציון אמיתי, טעות מדידה, מהימנות מבחן-מבחן חוזר, מהימנות של נוסחים מקבילים ועקביות פנימית. נוסף על כך אנו מניחות כי לקורא יש היכרות עם ניתוח שונות (ANOVA). תורת ההכללה עושה שימוש במערכי ניתוח שונות שבהם עיקר הגורמים הם מקריים (כלומר, מסקנות המחקר ניתנות להכללה על תנאים נוספים שלא נבדקו במחקר) ומרחיבה את משמעותם של כמה מהמושגים מתורת המבחנים הקלאסית.

בהמשך הפרק נכיר מושגים בסיסיים של תורת ההכללה. נלמד מהו מחקר הכללה ומהם תוצאותיו, בעיקר אומדנים לשונות של מקורות הטעות השונים, אומדנים לדיוק המדידה – טעויות התקן של המדידה ומקדמי הכללה. מקדמי ההכללה הם מדדים המסכמים את התוצאות של מחקר הכללה: הם מהווים אינדיקציה למידה שבה כלי המדידה או תהליך המדידה מבחינים באופן מהימן בין נשואי המדידה (למשל, תלמידים, כיתות, בתי ספר, שיטות לימוד) ללא קשר למאפיינים הספציפיים של הכלים או התהליך (למשל, איזו קבוצת פריטים ואילו מערכים שימשו במדידה). מקדמי ההכללה מאפשרים לנו לדעת באיזו מידה תוצאות המחקר ניתנות להכללה (למשל, על קבוצת פריטים אחרת ו/או על מערכים אחרים).

2. מבוא

בדרך כלל תהליכי המדידה במדעי הרוח, החברה ואף במדעי הטבע אינם מושלמים, ולכן תוצאות המדידה אינן מדויקות, ויש בהן טעות – טעות מדידה. טענה זו נכונה, למשל, כאשר אנו מודדים עמדות בנוגע למתמטיקה, לכישורי ניהול, לתפיסת כאב או ללחץ דם. הטעות במדידה נובעת, בין השאר, בגלל הקושי להגדיר במדויק את התכונות שמודדים ומשום שתכונות אלה אינן נצפות באופן ישיר. אנו מפתחים כלי מדידה שיספקו לנו עדויות על אודות התכונות שאנו מבקשים למדוד, אבל תהליך המדידה מושפע מגורמים רבים, ולכן בתוצאות המדידה יש מידה מסוימת של טעות.

תורת ההכללה – GT = Generalizability Theory (Rajaratnam, 1972 Cronbach, Gleser, Nanda, &) היא הרחבה של תורת המבחנים הקלאסית (Test Theory CTT = Classical). היא מספקת מסגרת תאורטית רחבה לבדיקת הדיוק והמהימנות של מדידת התנהגויות שונות. בעזרתה אפשר לזהות ולהבין את המקורות לטעות המדידה ואפשר לתכנן תהליכי מדידה יעילים וטובים. בפרק זה נציג את המונחים המרכזיים של התאוריה בדרך פשוטה ככל האפשר ונתאר את הקשר בינה לבין תורת המבחנים הקלאסית.

על-פי תורת המבחנים הקלאסית הציון הנצפה הוא סכום של שני מרכיבים : (1 ציון אמיתי, 2) טעות. ההנחה ב-CTT היא שהטעות היא מקרית, ולכן שונות הציון הנצפה היא הסכום של שונות הציון האמיתי ושל שונות הטעות. האומדנים לטעות המדידה ולמהימנות (היחס שבין שונות הציון האמיתי לשונות הציון הנצפה) משתנים בהתאם למערך המחקר ולשיטה של איסוף הנתונים. בחישוב מהימנות של מבחן-מבחן חוזר (test-retest) שונות הטעות היא השונות בין ביצועים בזמנים שונים (אך לא השונות שנובעת מדגימת הפריטים, שהרי בשתי הפעמים אנו משתמשים באותם הפריטים). בחישוב מהימנות פנימית (internal consistency) שונות הטעות היא השונות שנובעת מדגימת הפריטים (אך לא שונות בין ביצועים בזמנים שונים, שהרי אנו מעבירים את המבחן רק פעם אחת). בחישוב מהימנות של נוסחים חלופיים (alternate forms) שונות הטעות היא השונות שנובעת הן מדגימת הפריטים והן מהביצועים בזמנים שונים (אך לא נוכל להפריד ביניהן).

תורת ההכללה גמישה הרבה יותר מתורת המבחנים הקלאסית מכיוון שאפשר להגדיר במסגרתה את הציון האמיתי ואת הטעות ביותר מדרך אחת. תורת ההכללה מספקת מודלים ושיטות שמאפשרים לחוקר לפרק את הטעות למקורותיה, לאמוד בו-זמנית את השונות של כמה מקורות טעות ולהעריך את תרומתו היחסית של כל מקור לטעות. GT עושה זאת בעזרת מודלים של ניתוח שונות (ANOVA). המידע על הרכיבים של שונות הטעות מאפשר לתכנן מחקרים עתידיים שבהם מיישמים את התוצאות שהתקבלו וכך מצמצמים את טעות המדידה.

מונח בסיסי ב-GT הוא **מרחב ההכללה** (universe of generalization). כאשר מודדים תכונה מסוימת אצל אדם, מתעדים בעצם רק מדגם של התנהגות. הציון שמתקבל הוא רק אחד מציונים רבים שהיו יכולים לשמש את אותה המטרה. החוקר לא מעוניין בדרך כלל בתגובות של הנבחן לשאלות מסוימות במבחן מסוים או במועד מסוים, אלא מבקש להכליל ממנו על מצבים אחרים. התועלת של השימוש בציון תלויה במידה שבה תתאפשר ההכללה. אוסף המצבים האפשריים הוא מרחב ההכללה. קרונבך ועמיתיו (1972) טוענים כי קיים **מרחב תצפיות קבילות** (universe of admissible observations), שכל אחת מהן הייתה יכולה לשמש בתהליך של קבלת ההחלטות. הציון האידאלי לצורך קבלת ההחלטה הוא ממוצע הביצוע מעבר לכל מרחב התצפיות הקבילות. ממוצע זה נקרא **ציון המרחב** (universe score). בידי החוקר נמצא הציון הנצפה בלבד, והוא משתמש בו כאילו היה ציון המרחב, ולמעשה הוא מכליל ממדגם על מרחב. השאלה היא עד כמה ההכללה מדויקת.

המונח **מהימנות** – מונח מרכזי בתורת המבחנים הקלאסית – מוחלף בתורת ההכללה במונח הרחב והגמיש יותר **הכללה** (generalizability). במקום לשאול, כמו ב-CTT, עד כמה הציונים הנצפים משקפים את הציונים האמיתיים, ב-GT נשאל עד כמה הציונים הנצפים מאפשרים לנו להכליל מהם על התנהגות הנבחן במרחב מוגדר של מצבים (במרחב ההכללה). למרות ההבדלים בין שתי הגישות יש להדגיש כי GT איננה סותרת את CTT, אלא מרחיבה אותה.

3. מסגרת לאמידה של רכיבי טעות ומקדמי הכללה

המטרה המרכזית של תורת ההכללה היא להעריך את האפיונים של תהליך מדידה מסוים ולאמוד את דיוק המדידה (Cardinet, Johnson, & Gianreto, 2010). לשם כך יש לזהות תחילה את הגורמים התורמים לטעות ולאמוד את גודלם באמצעות ניתוח שונות. תרומתם היחסית של מקורות הטעות לשונות הציון נגזרת מאומדנים אלה. אומדנים של רכיבי השונות משמשים לחישוב טעויות המדידה ומקדמי ההכללה (G coefficients). רכיבי שונות אלה משמשים גם לניתוחי אופטימיזציה של תהליך המדידה, שמטרתם לנבא כיצד הטעויות והמקדמים ישתנו כפונקציה של שינויים במערך המחקר.

קיימת הקבלה בין דרך העבודה ב-GT לדרך העבודה ב-ANOVA. ב-GT עיקר הדגש הוא בזיהוי הגורמים התורמים לטעות המדידה, וב-ANOVA הדגש הוא בזיהוי האפקטים הקשורים למשתנים בלתי תלויים. בשתי השיטות החוקר מפעיל מניפולציות על הגורמים שמעניינים אותו, מפקח על חלקם או מתעלם מחלקם.

שבסון ועמיתיו (Shavelson, Webb, & Rowley, 1989; Shavelson & Webb, 1991) טוענים שההרחבה של CTT ל-GT דומה (אם כי לא זהה) להרחבה של ניתוח שונות חד-גורמי לניתוח שונות רב-גורמי. בניתוח שונות חד-גורמי החוקר מפרק את השונות הכללית לשני מרכיבים בלבד: שונות בין קבוצות (between) ושונות בתוך קבוצות (within). שונות בין קבוצות נחשבת לשונות שיטתית ונובעת מהמניפולציה שעשה החוקר במשתנה בלתי תלוי אחד כדי להבחין בין קבוצות המחקר. השונות בתוך קבוצות היא שונות מקרית ונחשבת לטעות משום שהיא מערפלת את ההבדלים בין הקבוצות שהן במוקד המחקר. בניתוח שונות רב-גורמי מוסיפים משתנים בלתי תלויים (גורמים) ומפרקים את השונות הכללית לשונות הנובעת מכל אחד מהמשתנים הבלתי תלויים ומהאינטראקציה ביניהם (כל אלה מהווים את השונות השיטתית) ולשונות הנובעת מהשארית שנחשבת טעות מקרית. ההרחבה מניתוח שונות חד-גורמי לרב-גורמי תורמת להגדלה של מרכיבי השונות השיטתית ולהקטנה של מרכיבי שונות הטעות שאין יודעים להסבירה.

ומה קורה בהרחבה של CTT ל-GT? בדומה לניתוח שונות חד-גורמי, ב-CTT מפרקים את השונות הכללית לשני מרכיבים בלבד: לשונות השיטתית שהיא שונות הציון האמיתית, שנובעת מהבדלים בין נשואי המדידה (בדרך כלל בין אנשים) בגין התכונה הנחקרת, ולשונות הטעות שהיא שונות מקרית. ב-GT איננו מדברים על משתנים כמו ב-ANOVA, אלא על המקורות לטעות המדידה. אנו מפרקים את הטעות לכמה מרכיבים, לאינטראקציות ביניהם ולאינטראקציות בינם לבין נשואי המדידה. ההרחבה מ-CTT ל-GT תורמת להקטנת השונות המקרית שאיננו יודעים להסבירה.

אמנם בדרך כלל נשואי המדידה הם אנשים, אך Cardinet, Touneur, & Allal (1976, 1981) וכן Cardinet, Johnson and Gianreto (2010) חושבים שכל אחד מהגורמים המשתתפים במחקר ההכללה (מחקר שמתבסס על תורת ההכללה) יכול לשמש נשוא המדידה.

4. שלבים בביצוע מחקר הכללה

במחקר הכללה מבקשים לבדוק את דיוק המדידה תוך שימוש בתורת ההכללה. קרדינה וחבריו (2010) מונים את הצעדים האלה במחקר הכללה:

1. החלטה על מערך התצפיות
2. דגימת שטחות (מערך האמידה)
3. הגדרה של מטרת המדידה
4. אמידה של טעות התקן של המדידה (דיוק המדידה)
5. חישוב של מקדמי הכללה (הערכה של מערך המחקר)
6. ביצוע של מחקרי אופטימיזציה

4.1 החלטה על מערך התצפיות

שטחות והיחסים ביניהן

הנתונים שאנו מנתחים הם בדרך כלל ציונים (למשל, ציוני יכולת, מיומנות, עמדות, וכדומה), והם מהווים את המשתנה התלוי ב-ANOVA. מקורות הטעות (למשל, שאלות, מועדי בחינה, מעריכים) הם המשתנים הבלתי תלויים, הגורמים. בתורת ההכללה המונח האנלוגי לגורם (משתנה בלתי תלוי) בניתוח שונות הוא שטחה (Facet), והרמות השונות של השטחה נקראות תנאי המדידה. לדוגמה, כל אחת מהשאלות במבחן במתמטיקה היא אחד התנאים של השטחה "שאלות", וכל אחד מהמורים שמעריכים את החיבור הוא אחד התנאים של השטחה "מעריכים" (במונחים של ANOVA היינו אומרים ש"השאלות" הן משתנה בלתי תלוי, וכל אחת מהשאלות היא אחד מערכיו, וגם ה"מעריכים" הם משתנה בלתי תלוי וכל אחד מהם הוא אחד מערכיו). כאשר המחקר כולל יותר מאשר שטחה אחת, יש להתחשב גם באינטראקציות בין השטחות בניתוח הנתונים. לדוגמה, במחקר שבו בודקים שונים מעריכים חיבורים בנושאים שונים, נרצה לבדוק את האינטראקציה בין שטחת המעריכים לבין שטחת החיבורים, ואז נוכל לראות אם מעריכים נבדלים ביניהם במידת החומרה שבה הם מעריכים חיבורים בנושאים שונים.

בכל תהליך מדידה יש להגדיר תחילה מהו מערך השטחות, כלומר, מהן השטחות המעורבות בתהליך, מה היחסים ביניהן ומיהם נושאי המדידה. בפרק זה נתייחס לשני סוגי מערכים: מערך צולב (crossed design) ומערך מקונן (nested design). במערך צולב כל תנאי בשטחה אחת (לרבות כל אחד מנושאי המדידה) מופיע עם כל אחד מהתנאים בשטחה האחרת. לדוגמה, מבחן רב-בכר שכולל 20 פריטים ($I = \text{items}$) מועבר לקבוצת סטודנטים ($p = \text{pupils}$), וכל סטודנט עונה על כל 20 הפריטים. הסטודנטים נושאי המדידה ושטחת הפריטים הם גורמים צולבים. מערך התצפיות הוא, אם כן, מערך צולב המסומן: $p \times i$ או π . בדוגמה זו יש רק שטחה אחת המהווה את מקור הטעות – הפריטים. דוגמה נוספת: 50 סטודנטים כותבים 3 חיבורים ($t = \text{tasks}$) כל אחד. כל חיבור מוערך על ידי אותו זוג

מעריכים ($r = \text{raters}$). נשואי המדידה הם הסטודנטים ומקורות הטעות הם שטחת המעריכים ושטחת החיבורים. גם בדוגמה זו המערך צולב והוא מסומן: $p \times r \times t$.

ואולם מה קורה אם כל חיבור מוערך על ידי זוג אחר של מעריכים? כלומר, 50 סטודנטים כותבים 3 חיבורים ($t = \text{tasks}$) כל אחד, אך כל חיבור מוערך על ידי זוג אחר של מעריכים ($r = \text{raters}$). במקרה זה שטחת המעריכים מקוננת בשטחת החיבורים ומסומנת $r:t$. הסטודנטים הם נשואי המדידה, והשטחות הן המעריכים והחיבורים. שטחת המעריכים מקוננת בתוך שטחת החיבורים, ולכן מערך התצפיות כולו נקרא **מערך מקונן** ומסומן: $p \times (r:t)$. במערך מקונן כל תנאי של שטחה אחת מופיע אך ורק עם אחד התנאים של השטחה האחרת.

ניתוח הנתונים נעשה באמצעות פירוק השונות הכללית לרכיביה.

דוגמה שתלווה אותנו לאורך כל הפרק ומבוססת על ברנן (Brennan, 1992) – מבחן כתיבה

נניח שרשות ארצית מבקשת לפתח מבחן להערכה של כושר כתיבה. בשלב ראשון הרשות תאפיין את סוג המטלות שייכללו במבחן; למשל, מטלות עיוניות שבהן התלמיד נדרש לדון בנושא מסוים. אפיון זה מגדיר את מרחב התצפיות הקבילות של שטחת המטלות. כמו כן הרשות תאפיין את המעריכים שיעריכו את רמת הביצוע של התלמידים; למשל, מעריכים בעלי ותק של שלוש שנות הוראה של שפת אם. אפיון זה מגדיר את מרחב התצפיות הקבילות של שטחת המעריכים. התלמידים הם נשואי המדידה. כמו כן נניח שהתקבלה החלטה ברשות, ולפיה בבדיקה הנוכחית כל מעריך (r) יעריך כל אחת ממטלות המבחן (t); כלומר, מדובר במערך שטחות צולב $t \times r$, ונוסף על כך כל אחד מהתלמידים (p) יבצע כל אחת מהמטלות. מערך התצפיות הוא, אם כן, צולב ויסומן כך: $p \times t \times r$.

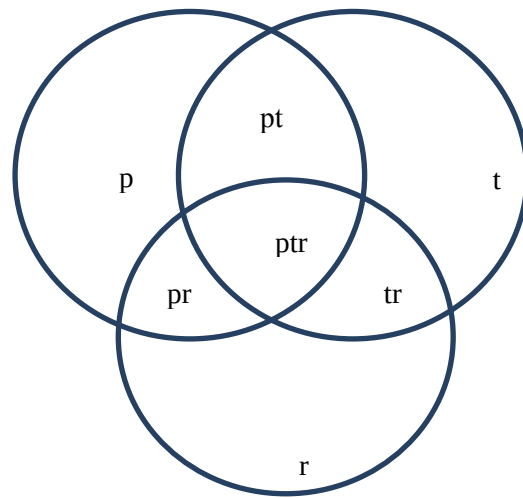
המודל המציג את הציון הנצפה של תלמיד מסוים במטלה בודדת שמוערכת על ידי מעריך בודד X_{ptr} הוא:

$$(1) \quad X_{ptr} = \mu + v_p + v_t + v_r + v_{pt} + v_{pr} + v_{tr} + v_{ptr}$$

μ הוא ממוצע הציונים מעבר לתלמידים, למטלות ולמעריכים (ממוצע כללי), ובמילים אחרות μ הוא ממוצע ציוני המרחב, ו- v מסמן את כל אחד מהאפקטים הבלתי תלויים. כל אפקט מייצג שטחה או אינטראקציה בין שטחות, והוא מוצג כהפרש בין הממוצע שלו לבין הממוצע הכללי μ . v_p הוא אפקט התלמיד – הפער בין ממוצע הציונים של התלמיד מעבר למטלות ולמעריכים לבין הממוצע הכללי (הפער בין ציון המרחב של התלמיד לבין μ), v_t הוא אפקט המטלה (הפער בין ממוצע הציונים על המטלה לבין μ), v_r הוא אפקט המעריך (הפער בין ממוצע הציונים שנתן המעריך לבין μ), v_{pt} מייצג את אפקט האינטראקציה בין התלמיד לבין המטלה (המידה שבה הציון של תלמיד במטלה מסוימת נבדל מציון המרחב שלו), v_{pr} מייצג את אפקט האינטראקציה שבין התלמיד לבין המעריך (המידה שבה מעריך חורג מרמת ההערכה הממוצעת שלו לטובת נבחן מסוים או לרעתו), v_{tr} הוא אפקט האינטראקציה בין המטלות למעריכים, ו- v_{ptr} הוא אפקט האינטראקציה בין שלוש השטחות במערך שלא ניתן להפרידו מגורמים שלא הוצגו במפורש ומהווים מקורות נוספים לטעות.

אפשר להציג את מערך התצפיות בעזרת דיאגרמת וואן שבה כל מעגל מייצג תלמידים או מטלות כתיבה או מעריכים. שבעת השטחים מייצגים את שבעת האפקטים שמרכיבים את הציון הנצפה, ואזורי החפיפה מייצגים את האינטראקציות בין האפקטים.

תרשים 1. דיאגרמת וון עבור המערך $p \times t \times r$



משוואה (1) מציגה את הציון הנצפה של התלמיד באמצעות אפקטים בלתי תלויים, ולכן שונות הציון היא סכום השונות של האפקטים כפי שהוא מוצג במשוואה (2).

$$(2) \quad \sigma^2(X_{ptr}) = \sigma^2(p) + \sigma^2(t) + \sigma^2(r) + \sigma^2(pt) + \sigma^2(pr) + \sigma^2(tr) + \sigma^2(ptr)$$

שימו לב $\sigma^2(\mu) = 0$.

$\sigma^2(X_{ptr})$ מסמן את השונות (ההיפותטית) של ציון תלמיד מסוים במטלה בודדת שהוערכה על ידי מעריך יחיד. אפשר לפרש שונות זו כך: נניח שתלמיד נבחן אין סוף פעמים, ובכל פעם נדגמת עבורו מטלה יחידה שמוערכת על ידי מעריך יחיד. המטלה והמעריך נדגמים אקראית מכל תנאי המדידה האפשריים שהוגדרו על ידי החוקר. בכל פעם התלמיד מקבל ציון. שונות ציונים אלה היא $\sigma^2(X_{ptr})$.

על מנת שאפשר יהיה לאמוד את רכיבי השונות יש צורך במדגם שכולל כמה תלמידים, כמה מטלות וכמה מעריכים.

4.2 דגימת שטחות (מעריך האמידה)

שטחות מקריות ושטחות קבועות (Fixed and random facets)

זיהוי השטחות ומעריך התצפיות אינו מספיק על מנת לאמוד את רכיבי השונות ולהגדיר את המודל המתאים ביותר לניתוח הנתונים. לשם כך יש לדעת איך נדגמו תנאי השטחות לצורך המחקר, או במילים אחרות, מהו סטטוס הדגימה של השטחות. שטחה יכולה להיות "מקרית", "סופית מקרית" או "קבועה". שטחה היא מקרית, אם תנאי המדידה שלה נדגמו באופן מקרי מתוך מרחב תצפיות קבילות גדול מאוד – תאורטית מתוך מרחב אין-סופי (למשל, 50 תלמידים שנדגמו מתוך אוכלוסייה של תלמידי כיתה ד, 20 פריטים שנבחרו מתוך כלל הפריטים האפשריים שבדקים יכולת מסוימת). אם מרחב התצפיות הקבילות אינו גדול, ובמחקר יידגמו רק כמה תנאי מדידה, סטטוס הדגימה של השטחה ייחשב מקרי סופי (למשל, במחקר יידגמו שניים מתוך עשרה מחוזות חינוך). שטחה מוגדרת כקבועה, אם כל תנאי השטחה שמעניינים את החוקר כלולים במחקר; במילים אחרות, אם כל מרחב התצפיות הקבילות שלה כלול במחקר (למשל, מגדר, מחוז גיאוגרפי, שיטות הוראה מסוימות).

אין הכרח שלכל השטחות במחקר יהיה אותו סטטוס דגימה. מעריך תצפיות מסוים יכול לכלול הן שטחות מקריות והן שטחות קבועות. זאת ועוד, סטטוס הדגימה של שטחה יכול להשתנות בהתאם למטרת המחקר. סטטוס הדגימה של השטחות קובע מהו המודל שישמש לאמידה של רכיבי השונות ואת הדרך שבה יחושבו טעויות המדידה ומקדמי ההכללה, והוא כמובן משפיע על הפרשנות שניתנת לתוצאות. כאשר השטחה מוגדרת כקבועה, נוכל להסיק מסקנות רק בנוגע לתנאים המסוימים שנכללו במחקר. כאשר השטחה מוגדרת כמקרית, נצטרך לדעת באיזו מידה אפשר להכליל את תוצאות המחקר על תנאים אחרים של השטחה. מכל האמור לעיל יוצא אפוא שהגדרה של סטטוס הדגימה של השטחות הוא שלב חשוב במעריך האמידה.

לאחר שהוגדרו שטחות המחקר ונקבע סטטוס הדגימה שלהן, נאספים הנתונים המתאימים ומחושבים אומדני השונות של מקורות הטעות באמצעות מודל ANOVA מתאים. האמידה מתבססת על ה-EMS (Expected mean square – EMS), והיא נעשית בעזרת תוכנות מתאימות. שלב זה במחקר ההכללה נקרא G-study. זה השלב שבו אומדים את רכיבי השונות של הציון שהוא יחידת המדידה הקטנה ביותר. בדוגמה שלנו – ציון התלמיד במטלה יחידה שהוערכה על ידי מעריך יחיד.

אפשר למצוא תיאור מפורט של המשוואות המתאימות וטכניקת האמידה בספרו של ברנן (Brennan, 2001, p. 6).

המשך דוגמה (מבחן כתיבה)

במחקר שהציג ברנן הוחלט על $n_r=2$ מעריכים, שכל אחד מהם מעריך $n_t=3$ מטלות שנכתבו על ידי $n_p=50$ תלמידים. מעריך התצפיות הוא, אם כן, מעריך צולב. במחקר זה מניחים כי מרחב התצפיות הקבילות הן של התלמידים והן של שתי השטחות (מטלות ומעריכים) הוא אין-סופי. אי לכך סטטוס הדגימה של השטחות הוא מקרי, ואפשר להשתמש במשוואה (2) כדי לקבל את האומדנים לרכיבי השונות.

האומדנים לשבעת רכיבי השונות שהתקבלו הם אלה :

$$\begin{aligned}\hat{\sigma}^2(p) &= .25 & \hat{\sigma}^2(t) &= .06 & \hat{\sigma}^2(r) &= .02 \\ \hat{\sigma}^2(pt) &= .15 & \hat{\sigma}^2(pr) &= .04 & \hat{\sigma}^2(tr) &= .00 \\ \hat{\sigma}^2(ptr) &= .12\end{aligned}$$

האומדן לשונות ציון של תלמיד מסוים במטלה יחידה שמוערכת על ידי מעריך יחיד הוא :

$$\hat{\sigma}^2(X_{ptr}) = 0.25 + 0.06 + 0.02 + 0.15 + 0.04 + 0.00 + 0.12 = 0.64$$

$\hat{\sigma}^2$ מסמן אומדן לשונות. נזכיר שוב שמשמעות השונות הזו היא זאת: נניח שתלמיד נבחן אין סוף פעמים. בכל פעם נדגמת עבורו מטלה יחידה שמוערכת על ידי מעריך יחיד. המטלה והמעריך נדגמים אקראית מכל תנאי המדידה האפשריים שהוגדרו על ידי החוקר. בכל פעם התלמיד מקבל ציון . האומדן לשונות ציונים אלה הוא $\hat{\sigma}^2(X_{ptr})$ ובמקרה שלנו 0.64.

נדגים את משמעות האפקטים באמצעות ההסבר על $\hat{\sigma}^2(p)$. בדוגמה של ברנ (2001) יפורש רכיב זה כך: לכל תלמיד יש ציון מרחב שהוא ממוצע הציונים שלו מעבר לכל המטלות והמעריכים במרחב התצפיות הקבילות. השונות של ציוני המרחב של כל התלמידים היא $\sigma^2(p)$ – זוהי בעצם שונות הציונים האמיתיים. מאחר שציוני המרחב הם ציונים תאורטיים (שכן המבחן לא כלל באמת את כל המטלות והמעריכים האפשריים במרחב), אין באפשרותנו לחשב את שונותם, ולכן אנו אומדים אותה באמצעות $\hat{\sigma}^2(p)$. את אפקט המטלה ואת אפקט המעריך אפשר לפרש באופן דומה.

המטלות שנבדקו נבדלות זו מזו במידת הקושי שלהן, והמעריכים נבדלים זה מזה במידת החומרה של ההערכה שלהם. מהאומדנים שהתקבלו אפשר לראות שמרכיב המטלה גדול פי שלושה ממרכיב המעריך, ולכן אפשר להסיק שמידת הקושי של המטלה משפיעה על ציונו של התלמיד הרבה יותר ממידת החומרה של המעריך.

נסדר את התלמידים לפי ציוניהם בכל מטלה. $\hat{\sigma}^2(pt)$ הוא אומדן למידה שבה סדר זה משתנה ממטלה למטלה, ו- $\hat{\sigma}^2(pr)$ הוא אומדן למידה שבה סדר התלמידים משתנה ממעריך למעריך. התוצאות מראות שמרכיב האינטראקציה בין התלמידים למטלות גדול כמעט פי ארבעה ממרכיב האינטראקציה בין התלמידים למעריכים. מהממצא הזה ומהממצא שהתקבל עבור אפקט המטלה ואפקט המעריכים אפשר להסיק שהמטלות משפיעות הרבה יותר מהמעריכים על השונות בציוני התלמידים.

כפי שנאמר לעיל, אומדני השונות שהתקבלו מתייחסים לציונו של תלמיד מסוים במטלה יחידה שהוערכה על ידי מעריך יחיד. אלה הם האומדנים המחושבים בשלב הראשון במחקר ההכללה – שלב ה-G-study. אומדנים אלה משמשים לנו בסיס לחישוב אומדנים של רכיבי שונות במצבים שלא נבדקו עדיין. המצב הפשוט ביותר שיעניין אותנו בדרך כלל הוא אמידת השונות של הציון הממוצע של התלמיד מעבר למספר מטלות ומספר מעריכים. אין צורך להרץ שוב מודל של ANOVA, אלא אפשר לקבל את

האומדנים הדרושים מהאומדנים שחושבו קודם לכן על ידי פעולות חילוק פשוטות – האומדנים של שונות הרכיבים חלקי מספר דרגות החופש המתאים. כדי להבחין אומדנים אלה מהאומדנים הקודמים נסמן את השטחות באותיות גדולות.

המודל המציג את הציון הממוצע של תלמיד מסוים מעבר למספר מטלות ומספר מעריכים X_{pTR} הוא:

$$(3) \quad X_{pTR} = \mu + v_p + v_r + v_R + v_{pT} + v_{pR} + v_{TR} + v_{TR}$$

האומדנים לרכיבי השונות של הציון הממוצע בדוגמה שלנו שכוללת שלוש מטלות ושני מעריכים הם אלה:

$$\begin{aligned} \hat{\sigma}^2(p) &= .25 & \hat{\sigma}^2(T) &= .02 & \hat{\sigma}^2(R) &= .01 \\ \hat{\sigma}^2(pT) &= .05 & \hat{\sigma}^2(pR) &= .02 & \hat{\sigma}^2(TR) &= .00 \\ & & \hat{\sigma}^2(pTR) &= .02 & & \end{aligned}$$

האומדן לשונות ציון ממוצע של תלמיד מסוים מעבר לשלוש המטלות ושני המעריכים הוא:

$$\hat{\sigma}^2(X_{pTR}) = 0.25 + 0.02 + 0.01 + 0.05 + 0.02 + 0.00 + 0.02 = 0.37$$

הערה: יש חוקרים שיראו בחישוב של שונות הציון הממוצע חלק מהשלב הראשון (שלב ה-G-study), ויש אחרים שיראו בו חלק מהשלב השני במחקר ההכללה, ובו נדון בהמשך הפרק.

4.3 הגדרה של מטרת המדידה

לאחר השלבים שכללו הגדרת שטחות, היחסים ביניהן, הגדרת סטטוס הדגימה שלהן, הגדרה של נשואי המדידה ואמידה של רכיבי השונות בעזרת המודלים השונים לניתוח שונות, מתחיל השלב של ניתוח ההכללה שמטרתו לספק מידע על המהימנות של תהליך המדידה ועל דיוק האמידה. ניתוח זה מתבסס על אומדני השונות שהתקבלו בשלב ה-G study.

בשלב של ניתוח ההכללה יש להגדיר אם מטרת המדידה היא יחסית (relative) או מוחלטת (absolute). נאמר שמדידה היא יחסית, כאשר משמעות הציון נגזרת ממקומו היחסי בהתפלגות הציונים (למשל, במבחני קבלה לאוניברסיטה). נאמר שמדידה היא מוחלטת, כאשר המטרה היא למקם את האדם על סולם המדידה ללא קשר לציונים אחרים (למשל, במבחן נהיגה). שני סוגי המדידה מושפעים מטעות המדידה, אך מדידה מוחלטת מושפעת ממנה, בדרך כלל, במידה רבה יותר. טעות התקן של המדידה מחושבת עבור הציונים של נשואי המדידה.

כשם שמטרת המדידה קובעת את סוג הציון (מוחלט או יחסי), כך היא קובעת גם את סוג טעות התקן של המדידה – טעות מוחלטת או טעות יחסית.

4.4 אמידה של טעות התקן של המדידה (דיוק המדידה)

בדוגמה של מבחן הכתיבה שבה מרחב ההכללה הוא אין-סופי ונשואי המדידה הם התלמידים, $\sigma^2(p)$ היא השונות של ציוני המרחב שלהם (ה"שונות האמיתית" ביכולת כתיבה). שאר השונות הן רכיבים של שונות הטעות.

שונות טעות מוחלטת (אבסולוטית) – $\sigma^2(\Delta)$

הטעות המוחלטת היא ההפרש בין הציון הנצפה של אדם לבין ציון המרחב שלו (אנלוגי לפער בין הציון הנצפה לבין הציון האמיתי ב-CTT). במונחים של המודל שהצגנו במשוואה 3, ציון המרחב של אדם הוא $\mu_p = \mu + v_p$, ולכן הטעות המוחלטת היא:

$$(4) \quad \Delta_p = X_{pTR} - \mu_p$$

$$(5) \quad \Delta_p = v_T + v_R + v_{pT} + v_{pR} + v_{TR} + v_{pTR}$$

הטעות המוחלטת כוללת את כל מקורות הטעות שמשפיעים על הציון של התלמיד, את המקורות שמשפיעים על גובה הציון שלו (השטחות והאינטראקציות ביניהן) ואת המקורות שמשפיעים גם על מיקומו ביחס לתלמידים האחרים (האינטראקציות בין נשואי המדידה והשטחות). שונות הטעות המוחלטת היא סכום השונות של כל מקורות הטעות האלה, והיא מחושבת בעזרת האומדנים לרכיבי השונות:

$$(6) \quad \hat{\sigma}^2(\Delta) = \hat{\sigma}^2(T) + \hat{\sigma}^2(R) + \hat{\sigma}^2(pT) + \hat{\sigma}^2(pR) + \hat{\sigma}^2(TR) + \hat{\sigma}^2(pTR)$$

בדוגמה שלנו:

$$(6) \quad \hat{\sigma}^2(\Delta) = .02 + .01 + .05 + .02 + .00 + .02 = 0.12$$

שורש שונות הטעות $\hat{\sigma}^2(\Delta) = +\sqrt{0.12} = 0.35$ הוא אומדן לטעות התקן המוחלטת של המדידה. לפיכך רווח סמך לציון המרחב של תלמיד ברמת סמך 68% הוא $X_{pTR} \pm 0.35$.

שונות טעות יחסית – $\sigma^2(\delta)$

טעות יחסית מוגדרת כפער שבין שני ציוני הפרש: (1) סטיית הציון הנצפה של תלמיד מתוחלת הציונים הנצפים של כלל התלמידים; (2) סטיית ציון המרחב של התלמיד מתוחלת ציוני המרחב של כלל התלמידים. כל אחד מציוני ההפרש האלה מייצג את מיקומו של התלמיד ביחס לממוצע של כלל התלמידים. במונחים של המודל שהצגנו במשוואה (3) הטעות היחסית היא:

$$(7) \quad \delta_p = (X_{pTR} - \mu_{TR}) - (\mu_p - \mu)$$

כאשר μ_{TR} הוא תוחלת הציונים הנצפים X_{pTR} .

כלומר,

$$(8) \quad \delta_p = v_{pT} + v_{pR} + v_{pTR}$$

כפי שאפשר לראות, הטעות היחסית מורכבת מכל הרכיבים שמייצגים אינטראקציות של השטחות עם נשואי המדידה, והשונות של הטעות היחסית היא סכום השונות של רכיבים אלה. שונות הטעות היחסית אנלוגית לשונות הטעות בתורת המבחנים הקלאסית.

בדוגמה שלנו האומדן לשונות הטעות היחסית הוא :

$$\hat{\sigma}^2(\delta) = .05 + .02 + .02 = .09$$

שורש השונות $\hat{\sigma}^2(\delta) = +\sqrt{0.09} = 0.30$ הוא אומדן לטעות התקן היחסית של המדידה. אנו רואים כי טעות התקן היחסית קטנה מטעות התקן המוחלטת $\hat{\sigma}(\Delta) = .35$. טעות התקן היחסית כוללת רק את האפקטים שיכולים להשפיע על סדר הציונים של התלמידים, ואילו הטעות המוחלטת כוללת נוסף על כך גם את האפקטים האחרים המשפיעים על הציון. אם כן, הטעות היחסית כוללת פחות רכיבי טעות, ולפיכך שונותה קטנה יותר משונות הטעות המוחלטת. עקב כך מתן פירוש יחסי לציון הוא מדויק יותר ומושפע פחות מטעויות המדידה מממתן פירוש מוחלט לציון.

4.5 חישוב של מקדמי הכללה (הערכה של מערך המחקר)

מקדמי הכללה (מהימנות)

מקדם ההכללה אנלוגי למקדם המהימנות ב-CTT השווה ליחס בין שונות הציונים האמיתיים לבין שונות הציונים הנצפים. בתורת ההכללה הציון האמיתי הוא ציון המרחב, ושונות הציונים הנצפים היא סכום השונות של ציוני המרחב ושל הטעות. כאשר הטעות יחסית, מקדם ההכללה נקרא Generalizability Coefficient, ונהוג לסמנו ρ^2 .

$$(8) \quad \rho^2 = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(\delta)}$$

ρ^2 הוא היחס בין שונות של ציוני המרחב של נשואי המדידה לבין סכום השונות של ציוני המרחב ושל הטעות היחסית. במערך תצפיות צולב שכולל את נשואי המדידה ושטחה אחת בלבד, מקדם ההכללה ρ^2 זהה למקדם המהימנות אלפא של קרונברך.

מקדם זה מהווה אינדיקציה למידה שבה ברפליקציה של תהליך מדידה שכולל שלוש מטלות ושני מעריכים נקבל תוצאות דומות לתוצאות המחקר הנוכחי. במילים אחרות, המקדם מהווה אינדיקציה למידה שבה ברפליקציה של תהליך מדידה יישמר אותו מדרג ציונים של נשואי המדידה.

בדוגמה שלנו האומדן למקדם ההכללה ρ^2 הוא :

$$\hat{\rho}^2 = \frac{.25}{.25 + .09} = .74$$

כאשר הטעות היא טעות מוחלטת, מקדם ההכללה נקרא index of dependability, ונהוג לסמנו Φ :

$$(9) \quad \Phi = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(\Delta)}$$

Φ הוא היחס בין שונות של ציוני המרחב של נשואי המדידה לבין סכום השונות של ציוני המרחב ושל הטעות המוחלטת.

מקדם Φ מתחשב בשונות הטעות המוחלטת, שהיא בדרך כלל גדולה יותר משונות הטעות היחסית, ולכן Φ הוא בדרך כלל נמוך מ- ρ^2 .

בדוגמה שלנו האומדן למקדם Φ הוא :

$$\hat{\Phi} = \frac{.25}{.25 + .12} = 0.68$$

ואכן $0.74 > 0.68$.

משמעות של מקדם ההכללה

מקדם ההכללה הוא מדד המסכם את התוצאות של מחקר הכללה : הוא מהווה אינדיקציה למידה שבה כלי המדידה או תהליך המדידה מבחינים באופן מהימן בין נשואי המדידה (למשל, תלמידים, כיתות, בתי ספר, שיטות לימוד) ללא קשר למאפיינים הספציפיים של הכלים או התהליך (למשל, איזו קבוצת פריטים ואילו מעריכים שימשו במדידה). מקדם ההכללה מאפשר לדעת באיזו מידה תוצאות המחקר ניתנות להכללה. ערכו של מקדם ההכללה נע בין 0.00 לבין 1.00. סוג ההחלטות שיש לקבל מכתוב את סוג מקדם ההכללה שבו נשתמש. למשל, בתהליך מיון, שבו חשוב מקומו היחסי של תלמיד, נהוג להשתמש במקדם ההכללה ρ^2 . מקדם של 0.80 לפחות הוא מקדם סביר שמאפשר להשתמש בתהליך המדידה לצורך קבלת החלטות חשובות. במבחני נהיגה שבהם יש חשיבות לציון המוחלט, נהוג להשתמש במקדם ההכללה Φ .

4.6 ביצוע של מחקרי אופטימיזציה

בשלב הבא במחקר ההכללה בודקים באיזו מידה וכיצד אפשר לשפר את תהליך המדידה, וכן בודקים איך ישפיעו שינויים במערך התצפיות על דיוק המדידה. הבדיקות שנעשה הן מסוג "מה יקרה אם". למשל, מה יקרה אם נשנה (נקטין או נגדיל) את מספר התנאים בשטחות הנבדקות, מה יקרה אם נגדיר בצורה שונה את סטטוס הדגימה של שטחה אחת או יותר (נניח, נגדיר שטחה מקרית כקבועה), מה יקרה אם נשנה את מערך השטחות (נניח נשתמש במערך מקוון במקום במערך צולב). כל בדיקה כזו היא בעצם מחקר. איננו צריכים לבצע בפועל את המחקר מכיוון שאת תוצאותיו אפשר לגזור מהאומדנים של רכיבי שונות הטעות שהתקבלו ב-G-study.

כדי שנוכל לעשות זאת, מומלץ שמערך התצפיות ב-G-study יהיה צולב. במערך כזה אומדים את מרב האומדנים שמהם אפשר לחשב מקדמי הכללה ולגזור אומדנים לרכיבי שונות במגוון מערכים אחרים (שלא נבדקו) בפעולות חשבון פשוטות ללא צורך להריץ מחדש מודל של ANOVA.

אם נבצע את הבדיקות מסוג "מה יהיה אם", נקבל אוסף של תוצאות שיעזרו לנו להחליט מהו תהליך המדידה הטוב ביותר (המיטבי) עבורנו. מהו מערך התצפיות שייתן לנו את התוצאות הטובות ביותר, שיאפשר לנו להכליל את הממצאים על מרחב ההכללה שמעניין אותנו במידת הדיוק הרבה ביותר. נוכל, למשל, לענות על השאלה אם כדאי לנו להשתמש ביותר מטלות ו/או ביותר מעריכים, אם ברצוננו להגדיל את דיוק המדידה.

תהליך הבדיקה שתיארנו בפסקה האחרונה (מה יהיה אם...) נקרא שלב ה-D-study. התהליך נקרא גם שלב קבלת ההחלטה. ההחלטה שעלינו לקבל היא מהו התהליך המיטבי עבורנו. מהו התהליך שבעזרתו נוכל לצמצם במשאבים מבלי לשלם בדיוק המדידה. האם נוכל להשתמש במערכים בזבזניים פחות ועדיין לקבל תוצאות טובות? למשל, להשתמש במערך מקוון במקום במערך צולב, לצמצם את מספר התנאים בשטחות, לקבוע שטחה כקבועה וכן הלאה. במצב של משאבים מוגבלים חיסכון בבדיקת מספר קטן יותר של תנאים בשטחות יכול לאפשר לנו להגדיל את מספר נשואי המדידה וכך להגדיל את מהימנות הבדיקה.

דוגמה

נשתמש באותם נשואי המדידה ובאותן השטחות שבדוגמה של מבחן הכתיבה. נבדוק איך משפיע שינוי במספר התנאים של השטחות על טעויות התקן של המדידה: מה יקרה לטעות התקן, אם נשנה את מספר המטלות ו/או את מספר המעריכים. בתרשים 2 כל גרף מתאר את הטעות המוחלטת כפונקציה של מספר המטלות עבור מספר מסוים של מעריכים. מהתרשים עולה בבירור כי טעות התקן קטנה, כשמספר המטלות גדל או כשמספר המעריכים גדל. עם זאת, ההבדלים בטעות עקב הגדלת מספר המטלות גדולים יותר מההבדלים שנובעים מתוספת מעריכים. טעות התקן של המדידה כמעט שווה, כאשר יש שלושה או ארבעה מעריכים, והיא קרובה מאוד לטעות התקן כשמדובר בשני מעריכים. לכן אפשר להסיק שאפשר להסתפק במחקר זה בשני מעריכים בלבד. הממצאים אינם מפתיעים. הם נובעים מהעובדה ששונות הטעות של המטלות (0.06) גדולה משונות הטעות של המעריכים (0.02), ושונות האינטראקציה בין הנבחן למטלה (0.15) גדולה משונות האינטראקציה בין הנבחן למעריך (0.04). לסיכום, ההחלטה על מספר התנאים (למשל, מספר המטלות ו/או מספר המעריכים) היא פונקציה של מידת הדיוק הנדרשת. גודל טעות התקן שהחוקר מוכן להשלים אותה תלוי בשימוש שיעשה בציון המבחן ובשיקולים של עלות, משך הזמן הזמין למבחן וכן הלאה.

תרשים 2



כפי שכבר צוין, המערך האידיאלי ב-G-study הוא מערך תצפיות צולב שבו השטחות מקריות. לפעמים נרצה להשתמש במערכים אחרים משום שמרחב ההכללה שמעניין אותנו מצומצם או שניאלץ לעשות זאת מפאת המשאבים המוגבלים שברשותנו. מערכים שאינם צולבים ישמשו אותנו בדרך כלל בשלב ה-D-study. להלן נציג שתי דוגמאות של מערכים אחרים: באחת סטטוס הדגימה של אחת השטחות הוא קבוע, ובאחרת מערך השטחות הוא מערך מקוון. בשתי הדוגמאות אנו מתבססים על האומדנים לרכיבי השונות שנאמדו בשלב ה-G-study וממצעים אותם בהתאם למערך התצפיות ולמערך האמידה המתאימים. לכן השטחות יסומנו באותיות גדולות.

5. דוגמאות נוספות

5.1 דוגמה למערך תצפיות עם שטחה קבועה

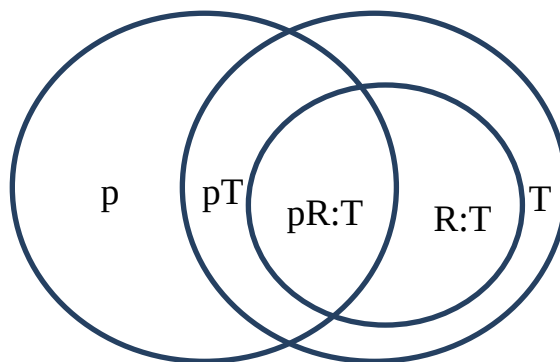
נשתמש בדוגמת מבחן הכתיבה של ברנן (1992b) ונניח שהחוקר מעוניין להכליל את ממצאי המחקר על מעריכים נוספים, אך לא על מטלות נוספות. במקרה זה שטחת המטלות תוגדר כשטחה קבועה, ושטחת המעריכים תישאר מקרית. האומדנים לרכיבי השונות יחושבו כמו במודל מעורב של ANOVA. כאשר סטטוס הדגימה של השטחות מעורב, מרחב ההכללה מצומצם יותר ממרחב ההכללה במערך שבו סטטוס הדגימה של כל השטחות הוא מקרי. מערך התצפיות בשני המקרים זהה: $p \times T \times R$, אך חישוב השונות של טעויות התקן שונה. בדוגמה שבה שטחת המטלות קבועה ושטחת

המעריכים מקרית, שונות הטעות המוחלטת היא 0.05, ושונות הטעות היחסית היא 0.04. האומדן למקדם ההכללה $\hat{\rho}^2$ הוא 0.88. מקדם הכללה זה מהווה אינדיקציה למידה שבה התוצאות שיתקבלו ברפליקציה של תהליך המדידה, שכולל את אותן שלוש מטלות אך מדגם שונה של מעריכים, יהיו דומות לתוצאות הנוכחיות. זהו מקדם גבוה יותר מהמקדם שחישבנו עבור המערך שבו סטטוס הדגימה של שתי השטחות היה מקרי (0.74). מקדם ההכללה הגבוה יותר אין פירושו שיכולת ההכללה בהכרח גבוהה יותר. בדוגמה זו מרחב ההכללה מצומצם יותר, ואפשר להכליל את הממצאים רק על שטחת המעריכים ולא על שטחת המטלות.

5.2 דוגמה למערך תצפיות עם מערך שטחות מקונן

במאמר של ברנן (1992b) מוצגת דוגמה שבה כל מטלה מוערכת על ידי קבוצת מעריכים שונים, אבל התלמידים נשואי המדידה מבצעים את כל המטלות. מערך השטחות בדוגמה זו הוא מערך מקונן – המעריכים מקוננים בתוך המטלות, ומערך התצפיות מסומן כך: $p \times (R:T)$. תרשים 3 מתאר דיאגרמת וואן של מערך זה.

תרשים 3: דיאגרמת וואן עבור המערך: $p \times (R:T)$



בתרשים 3 קינון השטחות מתבטא בכך שהעיגול המייצג את המעריכים R נמצא כולו בתוך העיגול המייצג את המטלות T . מהדיאגרמה עולה שאפשר להציג את ציון הנבחן כסכום של חמישה אפקטים:

$$(10) \quad X_{pTR} = \mu + v_p + v_T + v_{R:T} + v_{pT} + v_{pR:T}$$

שונות הציון הממוצע היא:

$$(11) \quad \sigma^2(X_{pT:R}) = \sigma^2(p) + \sigma^2(T) + \sigma^2(R:T) + \sigma^2(pT) + \sigma^2(pR:T)$$

אומדני השונות של התלמידים, המטלות והאינטראקציה ביניהם זהים לאומדנים שהתקבלו במערך הצולב, אך האומדנים לשונות הרכיבים שמכילים את המעריכים שונים כמובן (משום שאי-אפשר להפריד בין אפקט המעריך לאפקט המטלה). אומדני השונות שהתקבלו בדוגמה הם אלה:

$$\begin{aligned}\hat{\sigma}^2(p) &= .25 & \hat{\sigma}^2(T) &= .02 \\ \hat{\sigma}^2(pT) &= .05 & \hat{\sigma}^2(R : T) &= .003 \\ & & \hat{\sigma}^2(pR : T) &= .027\end{aligned}$$

האומדן לשונות הטעות המוחלטת $\hat{\sigma}^2(\Delta)$ הוא 0.10, האומדן לשונות הטעות היחסית $\hat{\sigma}^2(\delta)$ הוא 0.077 והאומדן למקדם ההכללה $\hat{\rho}^2$ הוא 0.76. שלושה מדדים אלה שונים במקצת מהמדדים שהתקבלו במערך הצולב. מערך השטחות משפיע אפוא על גודל טעויות התקן של המדידה ועל מקדמי ההכללה.

המסקנה העיקרית הנובעת מהדוגמאות שהוצגו בסעיפים האחרונים היא שאי-אפשר להעריך כראוי את ה"מהימנות" ואת "טעות התקן של המדידה" מבלי לדעת את כל המאפיינים של תהליך המדידה (כגון: מערך השטחות, מספר התנאים בכל שטחה ומרחב ההכללה שנקבע למעשה על ידי סטטוס הדגימה של השטחות).

6. סיכום

תורת ההכללה מספקת מסגרת שימושית לבדיקת מהימנות (דיוק מדידה) של התנהגויות. היא מרחיבה את תורת המבחנים הקלאסית בכך שהיא מספקת כלים:

- א. לאמידת השונות של מספר מקורות טעות;
- ב. לבניית מודל מתאים של מדידה לצורך קבלת החלטות יחסיות (norm relative) ומוחלטות (domain referenced);
- ג. לחישוב מקדמי הכללה שמתאימים למטרות המדידה: יחסיות או מוחלטות;
- ד. לביצוע מחקר אופטימיזציה לשם תכנון ויישום מערך מדידה יעיל.

במחקר הכללה שני שלבים:

שלב ראשון – G-study – שעיקרו פיתוח תהליך מדידה שיתאים למגוון מטרות ושימושים. בדרך כלל מערך התצפיות יהיה צולב והשטחות מקריות. בשלב זה אומדים את מרב האומדנים האפשריים של השונות של מקורות הטעות.

שלב שני – D-study – שעיקרו יישום של תהליך המדידה שפותח בשלב הראשון לסיטואציה מסוימת. בשלב זה ננצל את המידע מהשלב הראשון כדי לערוך בדיקות מסוג "מה יהיה אם" וכדי לתכנן את תהליך המדידה המיטבי למטרה מסוימת.

המשאבים המוגבלים לא תמיד מאפשרים לנו לבצע את שני השלבים בנפרד זה מזה. נשתמש באותו מערך תצפיות כדי לאמוד את השונות של מקורות הטעות, את השונות של הציון הממוצע וגם את מקדמי ההכללה. התוכנות הקיימות אכן עושות זאת.

יתרונותיה של תורת ההכללה באים לידי ביטוי במיוחד במחקרי הכללה שבהם בשלב ה-G-study משתמשים במערך צולב שבו השטחות מקריות, ולכל שטחה תנאים רבים. במערך תצפיות כזה מושגת יציבות רבה יותר של האומדנים של רכיבי השונות (כלומר, מושגת טעות אמידה קטנה יותר של רכיבי השונות). מהאומדנים שמקבלים אפשר לחשב אומדנים נוספים עבור מספר גדול של מערכים אחרים, שייבדקו בשלב ה-D study.

בשל ריבוי הפרטים הטכניים הנדרשים להצגת התאוריה של תורת ההכללה והשימוש בה, נדרשו שנים רבות להטמעתה. לשימוש בתורת ההכללה ולכלים שהיא מספקת יש כיום מעמד ברור וחד-משמעי בתחום של הערכת הדיוק והאיכות של מדידת התנהגות. אנו מוצאים את מקדמי ההכללה כמעט בכל מחקר שעוסק במדידה.

7. תוכנות זמינות לניתוח הכללה

Mushquash ו-O'connor (2006) מתארים תכניות SPSS, SAS ו-MATLAB שנכתבו לצורך ניתוח הכללה. הם מציעים שתי גרסאות של תוכנות. האחת פשוטה יותר אך גמישה פחות (G1) ואחרת הדורשת התערבות גדולה יותר של המשתמש, אך מאפשרת יותר גמישות בהגדרת המערכים (G2). התוכנות ניתנות להורדה מהרשת.

תכנית בשפת Fortran נכתבה על ידי Crick ו-Brennan (1983), ואפשר להזמינה מברנן באתר:

www.education.uiowa.edu/casma/GenovaPrograms.htm

תוכנה זו מספקת גם טעויות תקן עבור האומדנים וכן מקדמי הכללה. היא מתאימה למערכים מרובי שטחות וגם לניתוחי הכללה רב-משתנים.

ביבליוגרפיה

- Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer.
- Brennan, R. L. (1992). Generalizability Theory. *Educational Measurement: Issues and Practice*, 41(2), 27-34.
- Cardinet, J., Johnson, S., & Gianreto, P. (2010). *Applying generalizability theory using EduG*. New York: Routledge, Taylor & Francis Group.
- Cardinet, J., & Tourneur, Y., & Allal, L. (1981). Extension of generalizability theory and its applications in educational measurement. *Journal of Educational Measurement*, 13, 119-135.
- Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. *Journal of Educational Measurement*, 13(2), 119-135.
- Crick, J. E., & Brennan, R. L. (1983). *Manual for GENOVA: A generalized analysis of variance system* (American College Testing Technical Bulletin 43). Iowa City: ACT, Inc.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.
- Mushquash, C. & O'Connor, B. P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods*, 38, 542-547.
- Shavelson, R. J. & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.
- Shavelson, R. J., Webb, N. M., & Rowley G. L. (1989). Generalizability Theory. *American Psychologist*, 44, 922-932.

שאלות תרגול

1. הסבירו מהי הכללה לפי תורת ההכללה.
2. כיצד מגדירים חוקרים את מרחב ההכללה?
3. תלמידי נהיגה נבחנים בסימולטור שבו הם מתאמנים בחנייה. כל אחד מהם מקבל עשר מטלות חנייה. מדריך אחד מעריך את כל המטלות. נקבעה נקודת חתך שציון גבוה ממנה משמעותו שתלמיד הנהיגה עבר את המבחן.
 - א. מהו מערך התצפיות במחקר זה (נשואי המדידה והשטחות)?
 - ב. מהו סוג מערך התצפיות?
 - ג. מנו את המקורות לטעות מדידה במערך זה.
 - ד. מהו מערך האמידה המתאים ביותר, לדעתכם?
 - ה. איזה מקדם הכללה כדאי לחשב במקרה זה?
4. כמו כן הוחלט להציע קורס מתקדם ל-10% הנבחנים הטובים ביותר מקרב תלמידי הנהיגה. איזה מקדם הכללה כדאי לחשב במקרה זה? מה יהיה היחס בינו לבין מקדם המהימנות אלפא של קרונברך?
5. בקבוצה אחרת של תלמידי נהיגה היו שני מדריכים, וכל אחד מהם העריך את הביצוע של התלמידים בשמונה מטלות חנייה שהוצגו בסימולטור.
 - א. מהו מערך התצפיות במקרה זה?
 - ב. מהם הרכיבים של שונות הטעות שאפשר לאמוד במקרה זה?
6. קבוצה שלישית הורכבה מארבעה נהגים מנוסים שהיו אמורים לעבור רענון נהיגה. לכל אחד מהנהגים בקבוצה הוצגו בסימולטור שלוש מטלות חנייה מורכבות. כל מטלת חנייה הוערכה על ידי זוג אחר של מעריכים. מהו מערך התצפיות במקרה זה?
7. במה שונה החלטה מוחלטת מהחלטה יחסית?
8. מהו G-study ומהו D-study? מה הקשר ביניהם?
9. במחקר הכללה שונות הטעות המוחלטת שהתקבלה היא $\sigma^2(\Delta) = 0.16$. חשבו רווח סמך לציון המרחב ברמת סמך של 0.90.

תשובות לשאלות תרגול

1. ההכללה בתורת ההכללה מתייחסת למידת האמון שאפשר לתת להכללה ממדגם מסוים של פריטים (הזדמנויות, מעריכים ועוד) על כלל הפריטים (הזדמנויות, מעריכים ועוד) הקבילים.
2. מרחב ההכללה הוא אוסף המצבים (רפליקציות) שעליהם אפשר להכליל את תוצאות המחקר.
3. א. נשואי המדידה הם תלמידי הנהיגה. מטלות החנייה הן שטחה עם עשרה תנאים. מאחר שמדריך יחיד מעריך את כל המטלות, הוא לא ייחשב שטחה.
 ב. מערך התצפיות הוא מערך צולב שבו תלמידי הנהיגה (p) צולבים עם המטלות (t), והוא מסומן: $p \times t$.
 ג. במערך זה מקורות הטעות הם: מטלות חנייה (t) והאינטראקציה בין לבין נשואי המדידה ($p \times t$). נוסף כי רכיבי השונות שאפשר לאמוד במערך זה הם שלושה: שונות הנובעת מהבדלים בין התלמידים, שונות שמקורה בהבדלים בין המטלות ושונות שנובעת מהאינטראקציה בין התלמידים למטלות (שאי-אפשר להפרידה מיתר מקורות הטעות האפשריים).
 ד. נראה שבמקרה זה החוקר רוצה לדעת מהי יכולת החנייה של התלמידים, כלומר, להכליל מביצוע התלמידים בעשר מטלות חנייה אלה על כלל מטלות החנייה האפשריות. לפיכך סטטוס הדגימה של שטחת המטלות הוא מקרי.
 ה. מכיוון שההחלטה שיש לקבל על בסיס של ציוני המבחן היא מוחלטת, מקדם ההכללה המתאים הוא מקדם Φ .
4. במקרה זה ההחלטה על בסיס של ציוני המבחן היא החלטה יחסית, ולכן מקדם ההכללה המתאים הוא ρ^2 . מאחר שבמערך התצפיות יש שטחה אחת בלבד – שטחת המטלות, מקדם ההכללה ρ^2 זהה למקדם המהימנות אלפא של קרונברך.
5. א. מערך התצפיות בשאלה הוא מערך צולב: נשואי המדידה הם תלמידי הנהיגה. מקורות הטעות הם: שטחת המדריכים (עם שני תנאים) ושטחת מטלות החנייה (עם שמונה תנאים).
 ב. הרכיבים של שונות הטעות הם: מטלות, מדריכים, אינטראקציה בין המטלות למדריכים, אינטראקציה בין המדריכים לנשואי המדידה, אינטראקציה בין המטלות לנשואי המדידה ואינטראקציה משולשת בין מטלות, מדריכים ונשואי המדידה (שהיא בלתי ניתנת להפרדה משארית הטעות).
6. מערך התצפיות הוא מערך מקונן. המעריכים מקוננים בתוך מטלות החנייה. נשואי המדידה הם תלמידי הנהיגה. השטחות הן שטחת המטלות ושטחת המעריכים.
7. החלטה מוחלטת נוגעת לרמת הידע של נבחן, לרמת המיומנות או לעמדתו בנושא כלשהו ללא קשר לביצוע של אחרים במבחן. החלטה יחסית נוגעת למקומו היחסי של ציון הנבחן בקבוצת נבחנים מסוימת.
8. G-study הוא שלב במחקר הכללה שעיקרו מציאת אומדנים ראשוניים למקורות הטעות ולדיוק המדידה. בשלב זה אומדים את רכיבי השונות עבור נשוא מדידה יחיד ועבור כל אחד מתנאי המדידה בכל אחת מהשטחות. D-study הוא שלב במחקר הכללה שעיקרו קבלת החלטה בנוגע

לתהליך המדידה הטוב ביותר למטרה מסוימת. כדי לקבל את ההחלטה הנכונה משתמשים בשלב זה באומדנים שהתקבלו ב-G-study.

9. ראשית יש לחשב את טעות התקן המוחלטת של המדידה. טעות זו היא $0.4 = \sqrt{0.16}$. רווח הסמך ברמת סמך 90% הוא:
 הציון של נשוא המדידה $\pm (0.4 \times 1.65)$, שווה לציון ± 0.66 .