Title: The (In)Consistency of Teacher Survey Responses About Teacher Evaluation Implementation: Implications for Policymaking

Author:
Seth B Hunter
Assistant Professor of Education Leadership
George Mason University

Education Leadership
CEHD
4400 University Dr, MS2F1
Thompson Hall
Fairfax, VA, USA 22030
shunte@gmu.edu
Phone: 703 993 4485

Abstract

Several education agencies administer annual teacher surveys to learn about teacher evaluation implementation. Despite widespread use, we know relatively little about survey data quality. This brief describes the consistency of teacher reports regarding evaluation implementation using unique survey data collected multiple times within one semester. Teacher reports are significantly inconsistent across observations, raising concerns about the information gathered by annually-administered evaluation implementation surveys. The evidence implies that those using annually collected teacher evaluation implementation survey responses for high-stakes or costly decisions should do so cautiously or consider more frequent data collection.

Highlights

- Teacher reports are significantly inconsistent across observations, raising concerns about the information gathered by annually-administered teacher evaluation implementation surveys.
- The evidence implies that education agencies using annually collected teacher evaluation implementation survey responses for high-stakes or costly decisions should do so cautiously or consider more frequent data collection.

Keywords: educational policy, evaluation, school/ teacher effectiveness, validity/ reliability, correlational analysis, descriptive analysis, survey research

Introduction

In the wake of substantial teacher evaluation reforms over the last 15 years, several state and large local education agencies monitor teacher evaluation implementation via annually administered teacher surveys (See Appendix A). Indeed, these surveys appear to be the primary source informing education agency assessments of evaluation implementation.

Practical interest in teacher evaluation implementation is well-founded for several reasons. First, teacher evaluation implementation is costly (Stecher et al., 2016). Second, reformed evaluation systems can improve teacher effectiveness (Donaldson, 2021), a critical goal as students taught by more effective teachers experience better short- and long-term academic and non-academic outcomes (e.g., Chetty et al., 2014; Jackson, 2018). Finally, research suggests that evaluations' effects depend on implementation quality (Donaldson, 2021). However, despite the importance of teacher evaluation implementation and the widespread use of teacher surveys to monitor it, we know relatively little about the qualities of survey information collected.

This brief's purpose is to describe the consistency of teacher survey reports regarding formal classroom observation processes, a critical feature of evaluation reforms (Donaldson, 2021), using unique survey data that were collected multiple times within a single semester. If such teacher reports are inconsistent over a single semester, it casts doubt on the quality of information gathered by the related annual surveys collected by education agencies that inform evaluation-related decisions and policy.

Study Context

Tennessee policy assigns teachers one, two, or four classroom observations annually, though districts may assign additional observations. Soon after each observation, evaluators

score teachers using a standards-based rubric resembling Danielson's Framework for Teaching. State policy also stipulates that evaluators hold a post-observation conference within one week of each observation to share teacher performance feedback, identify improvement goals, and discuss strategies for improvement. Although the Tennessee Department of Education (TDOE) annually collected information about evaluation and other topics via the statewide-administered Tennessee Educator Survey, TDOE wanted to assess evaluation implementation over shorter periods to explore if it might revise its supports for evaluation implementation. Specifically, TDOE was interested in the implementation of post-observation conferences, a key feature of reformed evaluation systems (Donaldson, 2021). (See Appendix B for contextual details).

Data and Methodology

The analysis uses two panels of monthly-administered surveys from the 2018-19 Spring semester collected by a partnership among TDOE, the Tennessee Education Research Alliance, and five rural districts. Participating districts were recruited to represent the typical district in terms of 2017-18 annualized district-level measures of teacher effectiveness and evaluation implementation; all recruited districts agreed to participate and participating districts were representative of Tennessee districts (see Appendix C for details).

The study's partnership developed nine survey items to assess evaluation implementation. Items assessed the extent to which teachers agreed that: their observation scores reflected their performance (*ScoreRefl*); their evaluator referenced evidence during the conference (*Evidence*), possessed adequate expertise (*ObsExp*), encouraged teacher self-reflection (*Reflection*), and provided specific feedback (*Specific*); the conference was a two-way conversation (*2Way*); the conference feedback would improve their instruction (*ImpInst*); and, they could access relevant sources of expertise (*AccExp*) and professional development (*AccPD*) to reach evaluation-

informed improvement goals. The *Specific* item's responses ranged from 1=Generic to 5=Specific; otherwise, responses ranged from 1=Strongly Disagree to 6=Strongly Agree (see Appendix D). Teachers in participating districts received end-of-month TDOE emails from February through April asking if they received a formal observation within the last month. If so, teachers were invited to take the monthly survey, which remained open for one month after it was emailed. The analytic sample includes about 60 teachers who received two observations during the study period and completed two post-observation surveys. Seventy was the average number of days between responses and respondents resembled Tennessee non-participating teachers (see Appendix C for response rates).

The analysis estimates polychoric correlations[1] between teacher responses from their first and second within-semester survey submissions.

Key Findings

Teacher reports regarding observations and post-observation conference implementation vary significantly from their first to second survey submissions within the four-month study period (Figure 1). Although no correlations exceeded 0.68, the most consistent reports across both submissions concerned two-way post-observation conference conversations (*2Way*), access to relevant expertise (*AccExp*), and the extent to which: feedback would improve instruction (*ImpInst*), and evaluators encouraged teacher self-reflection (*Reflection*). As the remaining correlations are below 0.60, responses are weakly to moderately related from the first to second survey submissions. The least consistent reports applied to feedback specificity (*Specific*) and the extent to which: evaluators referenced evidence during the conference (*Evidence*), and observations scores reflected teacher performance (*ScoreRefl*).

[1] Ordinal data plausibly violates the assumptions underlying Pearson correlations; polychoric correlations are designed for ordinal data.

Discussion and Implications

Theoretically, sources of response inconsistency over the short, four-month study period fall into three broad categories: genuine variation in implementation across observations, bias, or random measurement error. Although the sources of inconsistency are worth further investigation, any combination of the three raises concerns about the information collected by traditional, annual teacher surveys concerning evaluation implementation. Measurement error or bias would infringe upon annual survey validity or reliability. Genuine variation in implementation across observations is also concerning as it implies that survey responses depend on the timing of survey administration. For example, suppose a teacher received an annual survey after her third annual observation. Although her responses may accurately represent implementation, her reports may differ significantly from those she would have reported had she received the annual survey after her fourth observation.

Ultimately, this brief suggests that policymakers and education agencies using annual surveys regarding teacher evaluation implementation to make significant decisions (e.g., deployment of costly implementation supports) should do so with caution. Alternatively, education agencies in systems where teachers receive multiple observations per year might administer multiple within-year surveys concerning evaluation implementation if implementation quality varies genuinely and substantially across observations.

References

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). The Long Term Impacts of Teachers: Teacher Value Added and Student Outcomes in Adulthood. *American Economic Review*, *104*(9), 2633–2679.

Donaldson, M. L. (2021). Teacher Evaluation Through the Lens of Psychology. In *Multidisciplinary Perspectives on Teacher Evaluation: Understanding the Research and Theory* (1st ed.). Routledge.

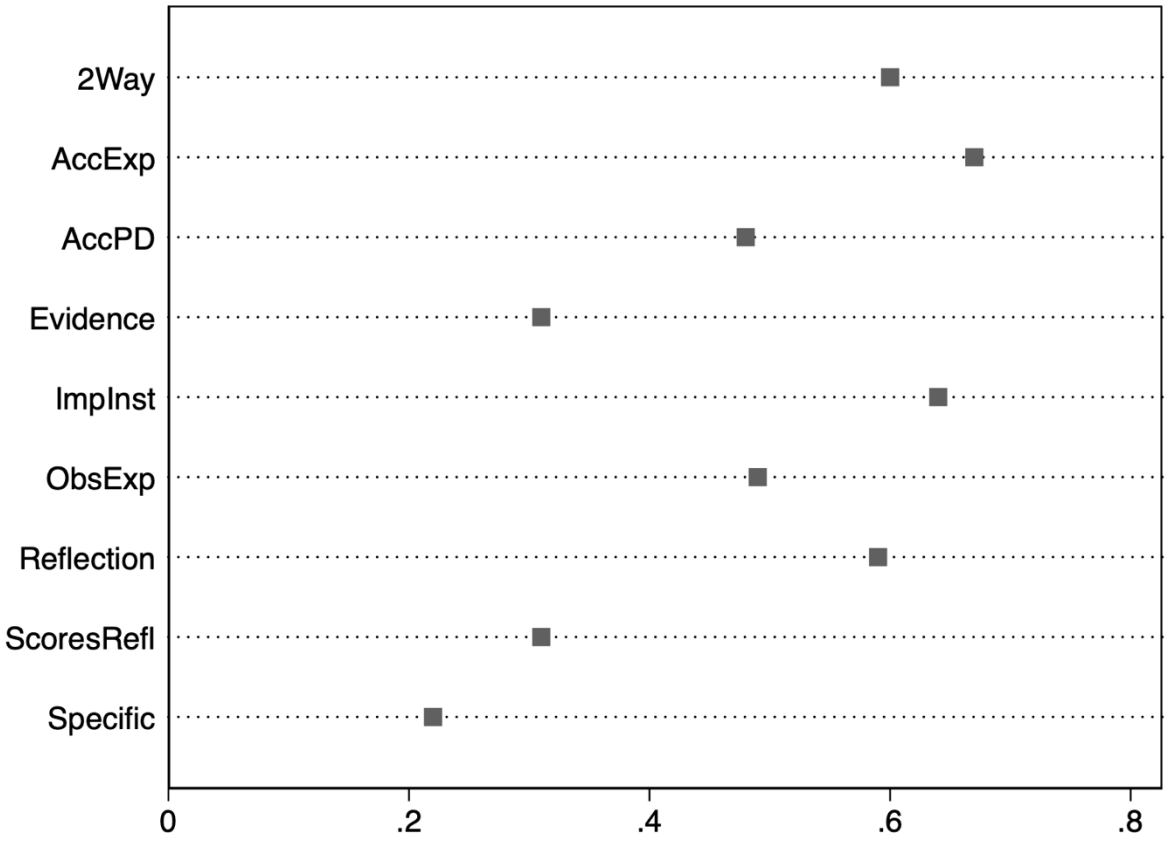
Hunter, S. B. (2020). The Unintended Effects of Policy-Assigned Teacher Observations: Examining the Validity of Observation Scores. *AERA Open*, *6*(2). https://doi.org/10.1177/2332858420929276

Jackson, C. K. (2018). What Do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes. *Journal of Political Economy*, *126*(5), 36.

Stecher, B. M., Garet, M. S., Hamilton, L. S., Steiner, E. D., Robyn, A., Poirier, J., Holtzman, D., Fulbeck, E. S., Chambers, J., & Brodziak de los Reyes, I. (2016). *Improving Teaching Effectiveness* (No. 9780833092212). http://proxy.library.vcu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,url,cookie,uid&db=ehh&AN=11254922&site=ehost-live&scope=site

Figure 1

Polychoric Correlations Between Within-Semester First and Second Survey Submissions



Notes: N(2Way) = N(Evidence) = N(ImpInst) = N(ObsExp) = N(Reflection) = 62. N(AccExp) = N(AccPD) = 61. N(ScoresRefl) = 67. N(Specific) = 55.

Appendix A. Use of Teacher Evaluation Implementation Surveys

Table A1 includes an incomplete list of educational agencies that administered surveys after 2011-12 to measure teacher perceptions of teacher evaluation implementation. Table A1 focuses on states that received Race to the Top grants and was created by searching Google and Google Scholar using terms such as "teacher evaluation implementation survey," "teacher evaluation perception survey," and "teacher perception survey."

The second column of Table A1 identifies each survey as "direct" or "indirect." The sole purpose of direct surveys was to gauge teachers' perceptions of evaluation implementation. For example, the Tennessee and Hawaii Departments of Education regularly sought direct feedback from teachers regarding evaluation implementation. Indirect surveys had a broader set of goals but included questions related to evaluation implementation. Several state agencies disseminated some version of the TELL (Teaching, Empowering, Leading, and Learning) survey, which does not solely focus on teacher evaluation implementation but includes items concerning evaluation implementation. Aside from TELL surveys, Brevard Public Schools (BPS) in Florida represents another example of an indirect survey, including questions related to the teacher evaluation implementation.

Table A1. Examples of States Administering Teacher Surveys Concerning Teacher Evaluation Implementaiton

Agency	Direct/Indirect	Relevant Links
Tennessee Department of Education	Direct	https://www.tn.gov/content/dam/tn/education/data/tchr_survey/2019/tchr_mod_evaluation.pdf
Hawaii State Department of Education	Direct	https://www.hawaiipublicschools.org/TeachingAndLearning/EducatorEffectiveness/EducatorEffectivenessSystem/Pages/home.aspx
The School District of Philadelphia (PA)	Direct	https://www.philasd.org/research/wp-content/uploads/sites/90/2020/03/Research-Brief_Teacher-Evaluation-Survey_FINAL_January-2015.pdf
Brevard Public Schools (FL)	Direct	BPS uses TNTP, Inc.'s Insight Survey. More about TNTP: https://tntp.org/teacher-talent-toolbox/insight-survey
Delaware Department of Education	Indirect	https://www.doe.k12.de.us/site/Default.aspx?PageType=3&DomainID=38&PageID=106&ViewID=6446ee88-d30c-497e-9316-3f8874b3e108&FlexDataID=20689
North Carolina Department of Public Instruction	Indirect	https://asqnc.com/
Ohio Department of Education	Indirect	https://education.ohio.gov/Topics/Teaching/Educator-Equity/TELL-Ohio
Massachusetts	Indirect	https://www.nysut.org/~/media/files/nysut/resources/2013/april/ted/mass_tlc_survey_finalreport.pdf?la=en
Kentucky Department of Education	Indirect	https://www.impactky.org/surveycontent
Colorado Department of Education	Indirect	https://www.cde.state.co.us/tlcc

Appendix B. Contextual Details

Observation and Conference Assignment

All Tennessee teachers receive an integer-based composite teacher Level of Effectiveness score ranging from one to five that is determined by observation scores and student outcomes (see Hunter, 2020 for details). Teachers with a prior-year score of five (one) are assigned one (four) observation(s). Teachers with composite scores ranging from two through four are assigned two observations if they have more than three years of experience; otherwise, teachers in this group are assigned four observations. State-assigned observations are minima; districts or schools may add to them. State policy also stipulates that teachers should receive half of their observations each semester and that post-observation conferences should occur within one week of every formal observation. If teachers do not receive a post-conference within one week, they may file a formal grievance. (Teacher and Principal Evaluation Policy, 2013).

The Tennessee Department of Education (TDOE) collects observation-level information, including observation dates; these data can be obtained via request through the Tennessee Education Research Alliance. The average teacher in this study's analytic sample received approximately 3.75 observations during 2018-19. As all teachers in the analytic sample took two surveys, all received at least two observations during the Spring 2018-19 semester.

Standards-Based Rubric

Tennessee provides districts the flexibility to choose their evaluation system as long as it meets state expectations (Teacher and Principal Evaluation Policy, 2013). More than 88% of districts use the Tennessee Educator Acceleration Model (TEAM), the state default system. The TEAM system includes a standards-based rubric that assesses classroom instruction according to three domains: Planning, Environment, and Instruction. Each domain includes several indicators

describing aspects of teaching (e.g., questioning, assessment) in terms of low (=1), middle (=3), or high performance (=5); scores range from 1 to 5.

References

Hunter, S. B. (2020). The Unintended Effects of Policy-Assigned Teacher Observations: Examining the Validity of Observation Scores. *AERA Open*, *6*(2). https://doi.org/10.1177/2332858420929276

Teacher and Principal Evaluation Policy, 5.201 7 (2013). http://www.tn.gov/sbe/Policies/5.201_Teacher_and_Principal_Evaluation_Policy_11-5-13.pdf

Appendix C. District Recruitment, Sample Descriptive Statistics and Response Rates

District Recruitment

The Tennessee Department of Education (TDOE) recruited districts to resemble the typical Tennessee district regarding four categories of prior-year district-level characteristics. The first characteristic concerned the school work environment and evaluation implementation. To assess this characteristic, TDOE identified a total of nine items from the 2017-18 Teacher and Administrator Tennessee Educator Surveys concerning these two constructs, then found the district-level mean responses to these items (see Table C1). Second, TDOE calculated district-level average teacher effectiveness using: TVAAS indices, Tennessee's *de facto* teacher value-added to student achievement measure, and its *de facto* teacher effectiveness composite score, Level of Effectiveness (LOE). LOE is a linear combination of teacher observations, student outcomes (e.g., achievement scores), and TVAAS scores for teachers of tested subjects (see Hunter, 2020 for details). Third, TDOE calculated a measure of 'misalignment,' defined as the absolute value between a teacher's observation and TVAAS scores, as it was concerned about the agreement between teacher observation and effectiveness scores. Misalignment ranged from 0 to 4 as TVAAS and observation scores ranged from 1 to 5. Finally, TDOE was also concerned with evaluators who did not differentiate their assessments of teacher performance, operationalized as the percentage of teacher observation scores that evaluators assigned to each of the observation score integer levels 1 through 5. TDOE effectively assumed that teachers vary in performance across the indicators assessed by the standards-based TEAM rubric; thus, it argued that there should be within-evaluator variation in performance scores. For recruitment, TDOE calculated district-level average teacher misalignment and average evaluator non-differentiation.

TDOE did not apply a formulaic recruitment process; it informally reviewed the four recruitment characteristics, then recruited five districts to represent the typical Tennessee district in terms of the four characteristics. To this end, TDOE was largely successful as the recruited districts closely resembled the typical Tennessee district (Tables C1, C2), and all five of the recruited districts agreed to participate. The five participating districts closely resembled non-participating districts regarding 2017-18 Tennessee Educator Survey responses (first characteristic); participating and non-participating district means in Table C1 were very similar and no differences were substantively significant. Participating and non-participating districts were also similar in terms of the second characteristic concerning TVAAS indices and teacher Levels of Effectiveness (Table C2).

The most significant practical differences concerned the two remaining characteristics: misalignment and non-differentiation (Table C2). The mean district-level teacher misalignment score in participating districts was 0.05 (0.45 SD) less than in non-participating districts; thus, participating district teachers receive more-aligned scores per TDOE's operationalization. District-level non-differentiation scores differed substantially; the district-level average evaluator in participating districts was ten percentage points less likely to differentiate than non-participating districts.

Response Rates and Sample Descriptive Statistics

The response rate among eligible teachers was between 31% and 38%. TDOE records describing the observation dates reveal that 176 teachers received at least two observations between the month before the first survey administration and month before the last administration. Of these 176 teachers, 55 (31%) was the fewest number of responses to a survey item, while 67 (38%) was the largest.

Teacher participants largely resemble non-participants regarding observable characteristics (Table C3). The years of experience, prior-year Level of Effectiveness, and prior-year observation scores between participants and non-participants are similar. However, participants' prior-year TVAAS scores are somewhat below non-participants (0.25 SD). In terms of education and demographics, the percentages of female participants and those holding more than a Bachelor's degree resemble non-participants. The greatest demographic difference concerns race: one percent of participants are nonwhite compared to 13 percent of non-participants.

References

Hunter, S. B. (2020). The Unintended Effects of Policy-Assigned Teacher Observations: Examining the Validity of Observation Scores. AERA Open, 6(2). https://doi.org/10.1177/2332858420929276

Table C1. Tennessee Educator Survey Items Used for Recruitment

Survey Item	Scale	District Level Statistics	
		Participants	Non-participants
School leadership provides useful feedback about my instructional practices.		3.17 (0.10)	3.18 (0.17)
Indicators from the teacher observation rubric are often referenced in informal discussions between teachers.		2.75 (0.13)	2.71 (0.15)
Indicators from the teacher observation rubric are often referenced in formal meetings where teaching is discussed.		2.95 (0.12)	2.93 (0.17)
In general, the teacher evaluation process used in my school has led to improvements in my teaching.		2.81 (0.11)	2.86 (0.19)
In general, the teacher evaluation process used in my school has led to improvements in student learning.	Strongly Disagree = 1; Strongly Agree = 4	2.71 (0.12)	2.79 (0.19)
I receive specific suggestions for professional learning that are tailored to my needs.		2.76 (0.10)	2.76 (0.18)
I gain information from statewide standardized exams that helps in refining my teaching practices.		2.16 (0.10)	2.12 (0.19)
Teacher item: Teachers hold all students to high academic standards.		3.31 (0.07)	3.27 (0.20)
School administrator item: Teachers hold students to high academic standards		3.49 (0.13)	3.29 (0.33)
During this school year (2017-2018), feedback that I received from my evaluator was (Please select the most appropriate answer)	Focused more on: helping me improve my teaching than making a judgment about my performance (=1); making a judgment about my performance than helping me improve my teaching (=2);	1.84 (0.13)	1.80 (0.14)

	Equally focused on helping me improve my teaching and making a judgment about my performance (=3).

Note: Means and standard deviations, in parentheses, listed. Districts are the unit of analysis. Non-participant sample size is 116 for the administrator survey item; otherwise, 121 districts contributed to the non-participant statistics. Participant data came from five districts.

Table C2. Non-Survey Information Used for Recruitment

Measure	District-Level Statistics	
	Participants	Non- participants
Misalignment between TVAAS and observation scores	0.15 (0.07)	0.20 (0.11)
Level of Effectiveness	4.18 (0.04)	4.23 (0.29)
TVAAS Index	3.14 (0.30)	3.15 (0.42)
Non-differentiation	0.09 (0.07)	0.19 (0.18)

Note: Means and standard deviations, in parentheses, listed. Districts are the unit of analysis. Non-participant sample size is 117 for the TVAAS, Level of Effectiveness, and Misalignment scores; 121 non-participant districts contributed to the Non-differentiation statistic. Participant data came from five districts.

Table C3. Sample and Population Teacher-Level Descriptive Statistics

Measure	Participants	Non-participants
Nonwhite	0.01 (.) [70]	0.13 (.) [71,244]
Female	0.84 (.) [70]	0.78 (.) [71,427]
More than BA	0.54 (.) [70]	0.60 (.) [71,395]
Experience	11.91 (9.73) [70]	11.90 (9.62) [70,781]
Prior-year Level of Effectiveness	4.08 (0.64) [62]	4.17 (0.79) [65, 964]
Prior-year TVAAS	2.71 (1.15) [21]	3.07 (1.43) [19,978]
Prior-year Observation Score	3.93 (0.45) [66]	4.04 (0.59) [70,074]

Note: Means, standard deviations in parentheses, sample size in brackets. Teachers are the unit of analysis.

Appendix D. Survey Items, Scales, and Descriptive Statistics

Table D1. Survey Items, Scales, and Descriptive Statistics

Item	Scale	Statistics	
		Wave 1	Wave 2
When my observer gave me feedback in the post-conference, s/ he used evidence and/ or data to support it. (*Evidence*)		*Agree* 3.70 (1.19)	*Agree* 3.70 (1.19)
The scores I received for this observation reflect my performance during the observation. (*ScoreRefl*)		*Agree* 3.76 (1.44)	*Agree* 3.91 (1.28)
My observer has the expertise necessary to evaluate my practice in the observed area. (*ObsExp*)		*Agree* 3.93 (1.19)	*Agree* 4.13 (0.94)
My observer prompted me to reflect on my current practice. (*Reflection*)	Strongly Disagree=1; Disagree=2; Slightly Disagree=3; Slightly Agree=4; Agree=5; Strongly Agree=6	*Agree* 3.91 (1.06)	*Agree* 4.01 (0.87)
The conference was a two-way conversation. (*2Way*)		*Strongly Agree* 4.32 (0.89)	*Strongly Agree* 4.47 (0.76)
I will be able to use the feedback I received in this conference to improve the quality of my instruction. (*ImpInst*)		*Agree* 3.99 (1.10)	*Agree* 4.19 (0.92)
I have the time needed to met with the person or people with relevant expertise in my school. (*AccExp*)		*Agree* 3.75 (1.20)	*Agree* 4.02 (0.91)
I have access to professional development that will help me implment suggestions based on the received feedback. (*AccPD*)		*Agree* 3.16 (1.26)	*Agree* 3.51 (1.16)
From general to specific, how would you characterize the feedback received during the post-conference? (*Specific*)	General Teaching Advice=1; Specific Strategies Tied to a Specific Lesson=5. Unlabeled middle integer responses 2 – 4.	*3* 2.72 (1.00)	*3* 2.72 (1.00)

Notes: Within each cell, the modal response is in italics, then the mean, and standard deviations are in parentheses. N(2Way) = N(Evidence) = N(ImpInst) = N(ObsExp) = N(Reflection) = 62. N(AccExp) = N(AccPD) = 61. N(ScoresRefl) = 67. N(Specific) = 55.