CHALLENGES OF MEASUREMENT IN FOREIGN LANGUAGE EDUCATION

İdil SAYIN

1. Introduction

Measurement can be defined as “the assignment of numerals to objects or events according to rules” (Stevens, 1946, p. 677). In a more well-defined way, measurement is “a process producing one or more property values that are attributed to the measurand with the aim of representing it” (Mari et al., 2015, p. 207). In addition to the general definition of measurement, it is noteworthy that conditions and assumptions of measurement may differ according to the field. In the physical sciences, measurement is expressing the observable attribute, events, objects, or their particular properties with numerical data. Whereas in social sciences, even though there is also quantification in the measurement, this is only possible on various assumptions. Accordingly, although the measured attribute in social sciences, unlike in physical sciences, cannot usually be directly observed, the measurement is made on the assumptions that the measured attribute exists in a certain amount in nature and can be quantified (Chadha, 2009). Another difference between the measurement in physical and social sciences is how the measurement takes place. In the physical sciences, measurement occurs through the direct interaction of objects with measurement tools (Pendrill & Fisher, 2013). Contrariwise, as what is intended to be measured in social sciences is in the human mind, the individual interacts with the measurement tool; that is, the measured phenomenon is indirectly measured (Salzberger, 2018). In summary, all sciences make use of measurement, they just use different methods and tools for this purpose. Although measurement methods and tools used in each science are similar, their features are different. Surely, this difference arises from the distinct nature of the hard data (physical properties, observable attributes) and soft data (non-physical properties, unobservable attributes) (Mari et al., 2015). It is noteworthy that challenges are encountered in all kinds of measurements, even where observable attributes are the subject of measurement (Furr & Bacharach, 2014). However, when unobservable attributes are at the center of measurement, challenges are far more complex and different. It is because the challenges found in the sciences where measurements are on unobservable attributes are either absent or largely eliminated in the physical sciences (Furr & Bacharach, 2014).

Challenges still present in social sciences flaw the interpretation of our measurement. These challenges have also led to frequent criticism of social sciences. Occasionally, social sciences has been criticized for not being a real science or even for being a pseudo-science per retaining measurement challenges (Feynman, 1981). Further, there are researchers who suggest that only observable attributes can be measured (for a comprehensive discussion see Mari et al., 2015). There are also those who call the measurement processes in social sciences in which observations are converted into numeric data a "black box" and referring this issue as the "Achilles' heel" of social measurements (Salzberger, 2018, p. 2). Although it may be impossible to completely eliminate these challenges, being careful when making the measurements and interpreting the results can minimize the adverse effects of these challenges.

One of the fields that requires the measurement of unobservable attributes is foreign language education. Measurements in foreign language education is conducted for various purposes such as diagnosis, placement and selection. However, these measurements may suffer from decreased understanding or flawed interpretation of the measured attribute due to several challenges. Therefore, raising consciousness regarding the issue can enable practitioners to be more careful in the planning and interpretation of measurements. Thus, this chapter aims at contributing to a better understanding of the complexity of such measurements and diverting attention to these by briefly describing the challenges of measurement in foreign language education. Even though categorization of these challenges may change, it is possible to gather these under three main categories. These categories are attribute/construct based, tool/method based, and human factors based. Attribute/construct based challenges arise from the complexity of identifying and describing the measured attribute. Tool/method based challenges concern the robustness of the tools and methods developed and used to measure these complex and unobservable attributes. Human factors based challenges originate from the individuals who are being measured or who make the measurement. The following sections will detail the related challenges.

2. Attribute / Construct Based Challenges

In measurement, measurand is not a person nor object, it is an attribute/construct of the object or person (Thorndike & Thorndike-Christ, 2014). Hence, attributes/constructs that are subject to the measurement hold great importance. The first challenge with attributes/constructs in foreign language education is that they can only be measured indirectly (Crocker & Algina, 1986). Measurement is conducted by assuming that psychological attributes/constructs manifest themselves through observable behaviors and that these observable behaviors are objective and countable (Chadha, 2009). In other words, we infer attributes/constructs by observing the ones that we assume are a manifestation of the attribute/construct in focus (Blanton & Jaccard, 2006). For example, a learner's foreign language proficiency cannot be observed, but this competence is supposedly measured by a set of observable behaviors (writing, listening, e.g.). This causes the measurement to be limited and vague (Chadha, 2009). Another challenge arises from the dynamic nature of attributes/constructs. Capturing and quantifying these non-static and constantly changing attributes/constructs at a point in time can also be considered a threat to the validity of the measurement (Chadha, 2009).

Another major challenge is conceptualizing. Before designing the measurement process, the attribute/construct to be measured must be determined and conceptualized (Doğan, 2020). Accordingly, a conceptual definition of the attribute/construct is required. The conceptual definition is the description of the attribute/construct in general terms. However, for a practitioner in social sciences, this is no easy task (Bulmer, 2001). It is because having numerous definitions for a single term is not exceptional in social sciences. For instance, a term that refers to the language “English” is hard to define. As the World Englishes view has started to be accepted more widely, this term has gotten harder to define (Hall, 2020). Who can

confidently define the English language? Whose English is the *real* English? Even if we claim that *real* English is the one that is spoken as a first language, we return to the previous question. Which one? While English is so difficult to define, even though it may seem like a fairly simple and perhaps superficial term, it is even more difficult to define various foreign language education-related terms. "Language" is another example of this polysemy. As Cook (2010) mentioned, the concepts of "language" differ across the different second language acquisition (SLA) theories. He even presented six different meanings of “language” used in SLA research (see Table 1).

Table 1. Different Meanings of “Language” (Cook, 2010, p. 7)

No	Meaning
1	a human representation system
2	an abstract external entity
3	a set of actual or potential sentences
4	the possession of a community
5	the knowledge in the mind of an individual
6	a form of action

According to him (Cook, 2010), all these various meanings ascribed to these terms affect the methods and tools that can be used in studies, especially because the meanings of these terms are often incompatible with each other. He argues that this polysemy causes practitioners to follow “separate paths on different maps” (p. 22).

After the conceptual definition, the operational definition must be determined. Operational definition refers to the procedures the phenomenon will be measured with. The operational definition of a measurement is determined after the conceptual definition. Because in cases where the variable to be measured cannot be directly observed or measured (latent variable), another variable that is assumed to represent this variable, namely the manifest variable, is measured. Since there is no consensus on conceptual definitions, it is normal for different people to have different definitions of the same structure and consequently different operational definitions. Later, a logical and numerical relationship of the structure with this manifest variable should be established (Lord & Novick, 2008). Therefore, measurements with different measurement procedures that claim to measure the same structure may have different results (Crocker & Algina, 1986). Another challenge stemming from the conceptualization issue is being unable to determine the scope of our measurement tools. It is not possible to measure the whole attribute/construct that we want to measure. As a result, only the "carefully chosen sample of behavioral dimensions" of that case is measured (Chadna, 2009, p. 18).

According to Bulmer (2001), the complexity of measurement in the social sciences stems from the existence of different positions that do not have much in common. A similar pluralism exists in SLA in terms of theories (Ellis, 2010). However, Ellis (2010) does not interpret this pluralism as negative; on the contrary, he argues that this is an indication of the intense interest towards

and richness of SLA. Conversely, considering this lack of unity as a limitation, Duncan (1984) argues that the absence of similar units similar to those in physical sciences (e.g. mass, length, etc.) in social sciences (except for economics) may be related to the fact that theories in social sciences are "fragmentary" and "undeveloped", and information is correlational rather than theoretical (p. 162).

Also relevant to the following section, not having a uniform conceptualization regarding any term incapables practitioners to adopt or develop an explicit measurement unit. This brings measurement tool development to a halt since it is expected that conceptualization instructs the development of measurement tools (Salzberger, 2018). Another similar criticism is resonated by American Physicist Richard Feynman (1981). Feynman (1981) argued that the social sciences are a pseudoscience that collects data without any rules (pseudoscience claims to be both scientific and factual, but consists of statements, beliefs, or practices that are incompatible with the scientific method (Curd & Cover, 1998)).

Another challenge in measurement is not having an absolute zero for a unit. Unlike a physical structure where an absolute zero can be easily appointed, it is not possible to determine an absolute zero in psychological structures (Chadha, 2009). In relation to this, absolute zero is required to interpret the ratio. Therefore, it would not be suitable to interpret the ratio in such measurements. For example, we cannot deduce that a student who scores 0 on a language ability exam does not have any language ability. It can only be said that the student could not answer any item correctly on the exam.

3. Tool / Method Based Challenges

Challenges encountered in the measurement also stem from tools and methods. The first related challenge arises from the lack of clear unit of measurement. As mentioned in the previous section, the absence of unified conceptualization fails to provide a clear unit of measurement (Salzberger, 2018). The absence of equal units may cause the same structure to give different results in different measurement procedures (Thorndike & Thorndike-Christ, 2014). In addition to the fact that these units are not unified, they are also arbitrary (Blanton & Jaccard, 2006). This arbitrariness prevents us from knowing how much a unit change in the observation transforms into the variable/construct in focus (Blanton & Jaccard, 2006). Consequently, having no absolute zero and working on an arbitrary one limit us. We cannot interpret a zero as an absence of a construct. Therefore, lack of an absolute zero requires us to work with measurement tools that are based on nominal, ordinal and at most interval scales of measurement (Doğan, 2020). This restricts the statistical operations that can be used when interpreting the latent variable that is assumed to be measured through the manifest variable with ordinal or nominal data. For example, a student who scored 100 on a language ability test cannot be interpreted as having twice the foreign language proficiency of a student who scored 50 on the same test. Likewise, it is not possible to interpret the real amount of an observed unit

of rising in measurement; it can only be said that there has been an increase (Blanton & Jaccard, 2006).

Another of the main challenges is to ensure that the measurement tool is suitable for measuring the assumed attribute/construct. In order to assure this, evidence should be collected on whether the measurement tool indeed measures the attribute/construct it claims to measure (Başokçu, 2020). It is also important to keep in mind that selecting the tool to be used is another complex issue. As there are tools that can better measure each attribute/construct, each tool also has its advantages and disadvantages. This affects the difficulty or ease of measurement and whether the measurement results are meaningful or not (Başokçu, 2020). One more challenge is that the measurement of an attribute/construct depends on the "composite scores" of measures of different attributes/constructs (Furr & Bacharach, 2014, p. 43). Although composite scores are convenient, it is difficult to determine which attributes/constructs represent the measured attribute/construct (Furr & Bacharach, 2014; Thorndike & Thorndike-Christ, 2014).

Relevant to the previous section, another challenge is being unable to measure the entire scope of an attribute/construct (Chadha, 2009). Capturing the entirety of an attribute/construct is not achievable because of the indefinite boundaries of the attribute/construct. Therefore, it is essential to note that inferences are made from only a limited sample of behavior that is believed to manifest the attribute/construct and this is not a complete presentation of it (Crocker & Algina, 1986). In this regard, Bachman (1995) points out a specific limitation that language testing faces, which is the requirement of teachers to measure their students' language abilities in controlled and limited conditions. However, it is unclear to what extent students' test performance is an accurate and effective indicator of their performance in real or non-test conditions. Furthermore, it is also quite difficult to determine the levels to be used to interpret the manifest variable (Chadha, 2009). Here, too, a threshold problem arises. What degree of performance can accurately represent students' language ability? Being able to answer this question brings us back to the conceptualization challenge (Bachman, 1995).

Score sensitivity is another challenge related to measurement tools. The precision of the scores in measuring the attribute/construct is important. Although this challenge exists in all areas of sciences, often poor sensitivity can easily be detected in physical sciences (Furr & Bacharach, 2014). On the other hand, determining such a problem is more complex in social sciences. That is because the presence of such a problem may not be noticed even after completing the measurement (Furr & Bacharach, 2014). Correspondingly, Başokçu (2020) states that test sensitivity is the main source of error.

Başokçu (2020) listed some features that challenge the accuracy and effectiveness of measurement tools. These are as follows;

- Unclear instructions (defects standardization in practice)
- Using sentence structures and words that are not suitable for the target audience,

- Not ordering the test items properly (easy questions should be presented at the beginning and at the end)
- Presence of response patterns (p. 64).

Also relevant to the following section, bias can be another challenging issue that can flaw the measurement process. Bias is systematically putting a group's performance at a disadvantage (Shepard et al., 1981). Bias can be in favor or against according to the group's gender, educational background, knowledge regarding a specific area, first language, or ethnicity (Elder, 2012). According to Kunnan (2007), there are three elements that contribute to bias in testing;

1. Using content or a language variety that is considered offensive or insulting to a group of people.
2. Causing a specific group of people to perform poorly on a test item.
3. Lacking standardization for measurement conditions.

A practitioner may unwittingly prefer a measurement tool or method that is biased. This is a great danger to the fairness of the test which also creates validity issues. Although there are various attempts to eliminate or minimize test bias, according to Elder (2012), the absence of bias in language assessment is unrealistic and unachievable.

Considering the above-listed challenges, we can say that there is a chicken-and-egg problem. While these challenges may result in utilizing measurement tools with flawed psychometric properties, the lack of understanding of psychometric properties may also cause poorly constructed tools to be used in measurement (Furr & Bacharach, 2014) that can also be mentioned as another challenge.

4. Human Factors Based Challenges

While procedures and tools are important for measurement, another important factor is the people involved in the measurement. This is because people make decisions in measurement, not procedures or tools (Thorndike & Thorndike-Christ, 2014). Therefore, human factors are considered one of the sources of the challenges experienced in measurements. These challenges can be caused by the person making the measurement or the person being measured. Bias, which was mentioned in the previous section, may also be dependent on human factors. Bias that the rater has about a particular group or person may affect the measurement results and cause a systematic error. Notions such as halo/horns effect (rater's tendency to evaluate a person positively or negatively according to a single feature being positive or negative), central tendency bias (rater's effort to gather measurement scores while scoring people at the center on a scale), leniency/strictness bias (rater's tendency to give overly high or overly low scores), similar-to-me effect (rater favoring people that is similar to him/her) can affect the type of bias. This remains a threat to the validity of the measurement as it is difficult to determine the amount and direction of bias. In some cases, people participating in the measurement may have the

impression that there is a bias in the measurement, even if there is none. For example, Kaplan and Saccuzzo (2018) note that the specific demographic information demanded in the measurement or the stated purpose of the measurement may cause minorities to experience "self-doubts" which cause poor performance due to stereotypes (stereotype threat) (p. 518). This may prevent the measurement from achieving its purpose (Chadna, 2009). It is also noteworthy that while there are researchers who argue that race may cause bias in measurements or cause different effects, it is still debated whether these differences are due to race or chance (Letukas, 2015).

Another issue is the preparation of measurement tools in accordance with the group that usually constitutes the majority of the population. Therefore, whether these groups have the same educational goals and a similar amount of motivation is one of the questions raised in this regard (Thorndike & Thorndike-Christ, 2014). For example, the essentiality of learning English as a foreign language may not have the same importance for a minority student as for others. Thus, the measurement of these students with different motivations, priorities, and educational goals by the same English exam may not yield correct results.

Challenges can also stem from the people being measured. People can react in ways that can flaw the measurement process. Participant reactivity is a well-known challenge in regard to this issue. Participant reactivity, also known as observer's paradox or Hawthorne effect (Furr & Bacharach, 2014), is the influence of measurement itself on the psychological state or process being measured. This influence may cause people being measured to behave differently due to various reasons. For example, people being measured may try to understand the purpose of the measurement and try to give answers that they think are desired (demand characteristic), people may try to change their behavior in order to influence the person conducting the measurement (social desirability), some individuals may consciously change their behavior in a bad/poor way (malingering) (Furr & Bacharach, 2014).

5. Conclusion

This chapter aims to briefly present the reflections of the challenges of the measurements in social sciences on the measurements in foreign language education. Challenges influence the measurement and prevent us from interpreting the results confidently. It becomes crucial, considering that measurement is at the heart of both education and sciences. Therefore, being aware of and understanding those challenges may help to minimize the effects of those challenges. To this end, challenges are sorted into and separately described under three categories. Although there are different challenges under these three categories, all of these categories are actually interrelated and contain overlapping challenges. These three categories are attribute/construct based, tool/method based, and human factors based.

It is obvious that a teacher must be careful when measuring, and interpreting results. That is because measurements in social sciences present special challenges and the measurement can only provide results that are convergent to reality and not the exact reality itself. Therefore, a

teacher being conscious of these challenges while taking action according to measurement results would be prudent and responsible. Moreover, it may be appropriate to develop the measurement processes very carefully with informed decisions and perhaps accompanied by an expert. In summary, these challenges should always inform our understanding and interpretation of our measurements. Additionally, we should always generate informed decisions to reduce the effects of these challenges.

Despite the impossibility of completely getting rid of those challenges, it may be possible to minimize those or reduce their effects. For example, even though many of these challenges arise from the nature of the attribute/construct, new studies can improve our operations for the better (Thorndike & Thorndike-Christ, 2014). Furthermore, we can raise the awareness of pre-service teachers through related courses during teacher training. Moreover, considering that measurements in foreign language education have unique challenges (authenticity, real-world performance, etc.), these challenges, and their influence should be one of the main topics to be covered in both pre-and in-service teacher training. Lastly, considering that the studies in the field of foreign language education shape language education and foreign language teacher education, it is very important to raise awareness among researchers about these challenges.

REFERENCES

Başokçu, T. O. (2020). Ölçme süreç ve sonuçlarının nitelikleri: Ölçme hatası, güvenirlik, geçerlik ve kullanışlılık. In N. Doğan (Ed.), *Eğitimde ölçme ve değerlendirme* (2nd ed., pp. 31-74). Pegem Akademi.

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist, 61*(1), 27–41. https://doi.org/10.1037/0003-066X.61.1.27

Bulmer, M. (2001). Social measurement: What stands in its way? *Social Research, 68*(2), 455-480. Retrieved from https://www.jstor.org/stable/40971466

Chadha, N. K. (2009). *Applied psychometry*. SAGE Publications India.

Cook, V. (2010). Prolegomena to second language learning. In P. Seedhouse, S. Walsh, & C. Jenks (Eds.), *Conceptualising 'learning' in applied linguistics* (pp. 6-22). Palgrave Macmillan London.

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory*. Wadsworth Group.

Curd, M., & Cover, J. A. (1998). *Philosophy of science: The central issues* (1st ed.).

Doğan, N. (2020). Temel kavramlar. In N. Doğan (Ed.), *Eğitimde ölçme ve değerlendirme* (2nd ed., pp. 1-30). Pegem Akademi.

Duncan, O. D. (1984). *Notes on social measurement: Historical and critical*. Russell Sage Foundation.

Elder, C. (2012). Bias in language assessment. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 406–412). Blackwell

Feynman, R. (1981). *Richard Feynman on the social sciences*. Interview by BBC. Retrieved April 7, 2021, from https://www.youtube.com/watch?v=tWr39Q9vBgo

Furr, R. M., & Bacharach, V. R. (2014). *Psychometrics: An introduction* (2nd ed.). SAGE Publications.

Hall, C. J. (2020). An ontological framework for English. In C. J. Hall & R. Wicaksono (Eds.), *Ontologies of English conceptualising the language for learning, teaching, and assessment* (pp. 13-36). Cambridge University Press.

Kaplan, R. M., & Saccuzzo, D. P. (2018). *Psychological testing: Principles, applications & issues* (9th ed.). Cengage Learning.

Kunnan, A. J. (2007). Test fairness, test bias, and DIF. *Language Assessment Quarterly, 4*(2), 109-112. https://doi.org/10.1080/15434300701375865

Letukas, L. (2015). *Nine facts about the SAT that might surprise you. Statistical report*. Retrieved from College Board Research website: https://files.eric.ed.gov/fulltext/ED562751.pdf

Lord, F. M., & Novick, M. R. (2008). *Statistical theories of mental test scores*. Information Age Publishing.

Mari, L., Carbone, P., & Petri, D. (2015). Fundamentals of hard and soft measurement. In A. Ferrero, D. Petri, P. Carbone, & M. Catelani (Eds.), *Modern measurements: Fundamentals and applications* (pp. 203-262). IEEE Press. https://doi.org/10.1002/9781119021315.ch7

Mayer, D. M., & Hanges, P. J. (2003). Understanding the stereotype threat effect with "culture-free" Tests: An examination of its mediators and measurement. *Human Performance*, *16*(3), 207-230. https://doi.org/10.1207/S15327043HUP1603_3

Pendrill, L. R., & Fisher, W. P. (2013). Quantifying human response: Linking metrological and psychometric characterisations of man as a measurement instrument. *Journal of Physics: Conference Series*, *459*, 012057. şpği10.1088/1742-6596/459/1/012057

Salzberger, T. (2018). Meeting Feynman: Bringing light into the black box of social measurement. *Journal of Physics: Conference Series*, *1065*, 072035. https://doi.org/10.1088/1742-6596/1065/7/072035

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics*, *6*(4), 317. https://doi.org/10.2307/1164616

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677–680. https://doi.org/10.1126/science.103.2684.677

Thorndike, R. M., & Thorndike-Christ, T. (2014). Fundamental issues in measurement. In *Measurement and evaluation in psychology and education* (8th ed., pp. 1-22). Pearson.

To Cite this Chapter:

Sayın, İ. (2022). Challanges of measurement in foreign language education. In A. Önal & K. Büyükkarcı (Eds.), *Essentials of foreign language teacher education,* (pp. 199-208). ISRES Publishing.

ABOUT THE AUTHOR

Res. Asst. İdil SAYIN

ORCID ID: 0000-0001-5546-2673

idilsayin@hacettepe.edu.tr

Hacettepe University

İdil Sayın currently works as a research assistant at Hacettepe University, Ankara, Turkey. She is also a Ph.D. student in the department of English Language Teaching (ELT) at her affiliated institution. Additionally, she is an MS student in the department of Measurement and Evaluation at the same university. She earned her Bachelor's degree in English Language Teaching (ELT) from Süleyman Demirel University. Later, she received her master's degree in ELT from the same university with her thesis about digital game-based language learning. She worked as a research assistant at Süleyman Demirel University for two years. Her research interests include measurement and evaluation in language education, instructional technologies, and quantitative research in language education