Editorial

Measurement Does Not Take Place in a Legal Vacuum

Ideas Regarding Regulation (EU) 2017/745 of the European Parliament and of the Council on Medical Devices

Matthias Ziegler[1] and Dragos Iliescu[2]

[1] Psychological Institute, Humboldt-Universität zu Berlin, Germany
[2] Faculty of Psychology and Educational Sciences, University of Bucharest, Romania

A typical paper on psychological assessment features a test and provides details about reliability, validity, and other psychometric properties supporting the test score interpretation (Ziegler, 2014). The intended readership of such papers is often other scientists. For tests used outside of science such information is typically comprised within easier read test manuals, intended for practitioners. The set of rules to evaluate complex information is specified in several standards which have been developed over decades. For example, the European Federation of Psychologists' Associations' review model of psychological and educational tests (EFPA Board of Assessment, 2013) or the American Psychological Associations Standards for Educational and Psychological Testing (AERA, APA, & NMCE, 2014) contain guidelines to evaluate the information supporting a test score interpretation. To guarantee the quality of the actual diagnostic process tests are used in, specific norms, guidelines, or standards have been released nationally (e.g., in Germany the DIN 33430, 2016; or in the Netherlands as described in Evers et al., 2010; or by the British Psychological Society as described in Lindley & Bartram, 2012; or in Spain as described in Muñiz et al., 2011) and international (e.g., ISO, 2020) contexts. Herein, the accumulated scientific knowledge as well as practical considerations are summarized. All of these examples show the length the assessment community has gone to in order to ensure that measurement instruments and the actual diagnostic process involving those instruments are of high quality. In fact, these guidelines (or similar documents) probably belong to the success stories of our science when it comes to establishing standards and transferring them into good practice. However, a new EU regulation aimed at medical devices could threaten this success story and generate a whole new class of problems affecting the assessment, especially in the clinical area.

Regulation (EU) 2017/745 of the European Parliament and of the Council on Medical Devices

Regulation (EU) 2017/745[1] of the European Parliament and the Council on Medical Devices, also known as the Medical Devices Regulation (MDR), was developed in response to concerns about the safety and effectiveness of medical devices in the European Union (EU). It applies to all medical devices, including those used in the diagnosis, prevention, monitoring, treatment, or alleviation of diseases. The regulation aims to improve the oversight and regulatory conditions of medical devices in the EU and to ensure that only safe and effective devices are placed on the market and used with beneficiaries. The development of the MDR involved the European Parliament, the Council of the EU, and the European Commission (2009, 2020), as well as input from industry stakeholders and patient organizations. The regulation was adopted in 2017 and entered into force in 2021, with a 3-year transition period to allow manufacturers to meet the new requirements. Thus, in 2024, the MDR will be actionable.

We note here that members of the psychological assessment community were not among the entities listed as involved in the development of the MDR; they were, to the best of our knowledge, not identified among the stakeholders and thus not invited to participate (albeit, as we will

[1]The regulation in different languages can be found here: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02017R0745-20200424.

see, they may well be influenced by the MDR). Criticism against the MDR in the psychological assessment community is therefore not surprising. Nor is it surprising for *any* new regulation to be met by criticism: the MDR has been criticized by some medical industry stakeholders, who argue that the new requirements are overly burdensome and could lead to delays in the availability of new medical devices or even to companies leaving the European market (e.g., https://www.qualitydigest.com/inside/fda-compliance-article/eu-regulation-driving-medical-device-manufacturers-eu-market-112822). Some have also argued that the MDR could lead to higher costs for manufacturers and consumers (e.g., https://www.medtechdive.com/news/eu-mdr-costing-smaller-medtechs-5-of-their-annual-sales-survey/584399/. The MDR has been featured in several publications, including the *Journal of the American Medical Association* and the *European Journal of Health Law*. Scientific journals that have written about the MDR include the *Journal of Medical Devices* and the *Journal of Medical Devices, Data, and Systems*.

Despite the attention the MDR has received in the medical media and medical industry, little was it discussed in psychological assessment and in psychology in general. The German Board of Assessment and Testing (DTK)[2] was tasked by the German Federation of Psychological Associations to compose a statement dealing with the question of whether psychometric tests are medical devices. This statement was also accepted by the EFPA Board of Assessment and was discussed in different settings by other international associations, such as the International Test Commission and the European Association of Test Publishers. In the following, we will summarize the core ideas of both the MDR and the statement published by national and international assessment experts (DGPS, https://www.dgps.de/schwerpunkte/stellungnahmen-und-empfehlungen/stellungnahmen/details/messtheoretisch-fundierte-tests-sind-keine-medizinprodukte/; EFPA, https://www.efpa.eu/regulation-eu-2017745-medical-devices-efpa-response). We want to stress that this editorial is not written in an anti-EU spirit. On the contrary, the authors strongly support the European spirit in general and see the benefits EU-wide regulations can have. At the same time, it is part of any democratic process to openly discuss the effects of laws and regulations and their application in our specific professional community.

Core Ideas of the Medical Devices Regulation

To begin, it is important to clarify what a medical device is, according to the MDR. To this end, we will list passages from the regulation, which states "'medical device' means any instrument, apparatus, appliance, software, implant, reagent, material or other article intended by the manufacturer to be used, alone or in combination, for human beings for one or more of the following specific medical purposes:

- diagnosis, prevention, monitoring, prediction, prognosis, treatment or alleviation of disease,
- diagnosis, monitoring, treatment, alleviation of, or compensation for, an injury or disability,
- investigation, replacement, or modification of the anatomy or of a physiological or pathological process or state,
- providing information by means of in vitro examination of specimens derived from the human body, including organ, blood, and tissue donations, and which does not achieve its principal intended action by pharmacological, immunological, or metabolic means, in or on the human body, but which may be assisted in its function by such means." (European Commission, 2020, pp. 5–6)

Furthermore, medical devices are classified into several categories. Here, the following text passage is of potential importance:

> "Software intended to provide information which is used to take decisions with diagnosis or therapeutic purposes is classified as class IIa, … . Software intended to monitor physiological processes is classified as class IIa, except if it is intended for monitoring of vital physiological parameters, … ." (European Commission, 2020, p. 188)

When reading those passages, examples easily come to mind where psychological tests are used in a clinical setting with the purpose of diagnosing or treating mental illnesses. Furthermore, tests of cognitive ability are often used in neuropsychological settings for similar purposes. Such tests are often computer-based and thus, software in a broader sense. Thus, the first conclusion of those assessment experts this editorial is based on (i.e., German Board of Assessment and Testing) was that some of the tests used in clinical psychological assessment might be prone to be classified as medical devices when administered via a computer (though not when administered in paper-and-pencil format).

Implications of the Medical Devices Regulation for Quality Management

The actual regulation contains little guidance concerning quality management. More details can be found in the extensive annex (Annex I, 15.1):

[2]Matthias Ziegler, the first author of this editorial is a member of the German Board of Assessment and Testing and participated in drafting the statement featured here.

> "Diagnostic devices and devices with a measuring function, shall be designed and manufactured in such a way as to provide sufficient accuracy, precision and stability for their intended purpose, based on appropriate scientific and technical methods. The limits of accuracy shall be indicated by the manufacturer."

To provide evidence for accuracy, precision, and stability a complex process needs to be undertaken on a regular basis. This would not only invoke costs but also has further implications. The costs result from the large number of test takers which typically have to be assessed to have a data set sufficient in power to be utilized in typical psychometric quality tests.

It might be considered a positive fact that the terms accuracy, precision, and stability are featured in the MDR – all these are terms that are also typically used in psychological assessment when gauging a test score's reliability (Emons et al., 2007; Gignac, 2014; Hancock & Mueller, 2001; Nezlek, 2017; Revelle & Garner, 2022; Sijtsma, 2009; Zinbarg et al., 2005). Unfortunately, other quality criteria like validity, norming, or fairness to name but a few prominent examples, are not directly addressed in the MDR. This might seem like a blatant flaw at first but may be considered less stringent when corroborated with the following point.

The MDR makes an explicit statement about the units of measurement that are provided by medical devices (Annex 15.2): "The measurements made by devices with a measuring function shall be expressed in legal units conforming to the provisions of Council Directive 80/181/EEC." This directive (https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32009L0003&from=EN) contains a table (see Figure 1). This table lists the quantities medical devices measure and the units used for this purpose. Here we can find quantities and units that are typical in physics and other natural sciences, and which have a direct interpretation. Validity evidence is in such contexts not as important as it is for psychological tests where the typical unit is a sum of points or a factor score.

Thus, the assessment experts of the German Board of Assessment and Testing concluded in their statement that applying the MDR to psychological tests would run the risk of narrowing the quality focus to reliability aspects only. Moreover, the assessment experts stated the worry that some computer-based psychological tests might be placed into the medical device category IIa by legal authorities applying the MDR. Both worries seem reasonable and the effect on quality management of applying the MDR to clinical-psychological tests administered by computers seems indeed problematic. Before further weighing in on this, we will first outline what the literature defines as a psychometric test based on measurement theory. Again, we will closely follow the ideas in the statement published by the assessment experts of the German Board of Assessment and Testing.

What Is a Psychometric Test?

The European Federation of Psychologists' Associations' (EFPA) Board of Assessment defines a psychometric test as any evaluative device or procedure in which a sample of an examinee's behavior in a specific domain is obtained and subsequently evaluated and scored using a standardized process (EFPA Board of Assessment, 2013). This definition encompasses all instruments that fall under this category, regardless of the specific label they are given, such as tests, scales, questionnaires, inventory, or projective techniques. This definition is closely aligned with the definition provided by the APA Standards mentioned above.

Measurement according to a standardized process is a key aspect of psychometric testing. This is achieved by assigning numbers to observations in accordance with certain rules (Stevens, 1946). It is important to note that this assignment of numbers does not yield a directly interpretable quantity, as psychometric measurements do not measure or report in a natural unit (Michell, 2001). Instead, the measurement is interpreted by referencing a comparative value, such as by converting it to a normed value in norm-oriented testing or comparing it to a cut-off value derived based on a comparison sample in criterion-oriented testing. Such outputs from psychological assessments do not have a unit of measurement that is recognized in Directive 80/181/EEC (see Figure 1).

Psychometric tests based on measurement theory (e.g., classical or probabilistic test theory) do provide information that is taken into account in clinical-psychological diagnostics, but only after interpretation by diagnosticians and consideration of additional information (e.g., Witteman et al., 2018). As noted above, national and international guidelines exist on which to evaluate the quality of psychometric tests. These guidelines all define quality based on various criteria, including reliability (measurement accuracy, precision, stability), objectivity, validity, norming, fairness, and more. This assessment of quality is more comprehensive than that required by Regulation (EU) 2017/745 for medical devices and often includes detailed checklists of information to be considered in relation to the quality criteria.

Regarding the implementation of psychometric tests in software, it is important to note that there is generally no functional difference between a test administered using paper and pencil and a computer-based test with respect to Regulation (EU) 2017/745. In both cases, measurement is achieved by assigning numbers, but with computer-based testing, the assignment is done electronically rather than using evaluation templates. The conversion to standard

1.2.3. SI derived units with special names and symbols

Quantity	Unit		Expression	
	Name	Symbol	In terms of other SI units	In terms of SI base units
Plane angle	radian	rad		$m \cdot m^{-1}$
Solid angle	steradian	sr		$m^2 \cdot m^{-2}$
Frequency	hertz	Hz		s^{-1}
Force	newton	N		$m \cdot kg \cdot s^{-2}$
Pressure, stress	pascal	Pa	$N \cdot m^{-2}$	$m^{-1} \cdot kg \cdot s^{-2}$
Energy, work; quantity of heat	joule	J	$N \cdot m$	$m^2 \cdot kg \cdot s^{-2}$
Power ([1]), radiant flux	watt	W	$J \cdot s^{-1}$	$m^2 \cdot kg \cdot s^{-3}$
Quantity of electricity, electric charge	coulomb	C		$s \cdot A$
Electric potential, potential difference, electromotive force	volt	V	$W \cdot A^{-1}$	$m^2 \cdot kg \cdot s^{-3} \cdot A^{-1}$
Electric resistance	ohm	Ω	$V \cdot A^{-1}$	$m^2 \cdot kg \cdot s^{-3} \cdot A^{-2}$
Conductance	siemens	S	$A \cdot V^{-1}$	$m^{-2} \cdot kg^{-1} \cdot s^3 \cdot A^2$
Capacitance	farad	F	$C \cdot V^{-1}$	$m^{-2} \cdot kg^{-1} \cdot s^4 \cdot A^2$
Magnetic flux	weber	Wb	$V \cdot s$	$m^2 \cdot kg \cdot s^{-2} \cdot A^{-1}$
Magnetic flux density	tesla	T	$Wb \cdot m^{-2}$	$kg \cdot s^{-2} \cdot A^{-1}$
Inductance	henry	H	$Wb \cdot A^{-1}$	$m^2 \cdot kg \cdot s^{-2} \cdot A^{-2}$
Luminous flux	lumen	lm	$cd \cdot sr$	cd
Illuminance	lux	lx	$lm \cdot m^{-2}$	$m^{-2} \cdot cd$
Activity (of a radionuclide)	becquerel	Bq		s^{-1}
Absorbed dose, specific energy imparted, kerma, absorbed dose index	gray	Gy	$J \cdot kg^{-1}$	$m^2 \cdot s^{-2}$
Dose equivalent	sievert	Sv	$J \cdot kg^{-1}$	$m^2 \cdot s^{-2}$
Catalytic activity	katal	kat		$mol \cdot s^{-1}$

([1]) Special names for the unit of power: the name volt–ampere (symbol "VA") when it is used to express the apparent power of alternating electric current, and var (symbol "var") when it is used to express reactive electric power. The "var" is not included in GCPM resolutions.

Figure 1. List of units for results derived with medical devices according to the MDR. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32009L0003&from=EN. © Europäische Union, 1998–2023. Available under license CC BY 4.0 (https://creativecommons.org/licenses/by/4.0).

values is also computer-based, but it follows the same rules as those applied by human test administrators. As a result, the error rate can be minimized with software-based testing, but the resulting report and the resulting information are the same. Therefore, the use of the software does not have any particular diagnostic significance, and it would be inappropriate to classify the same instrument as a medical device in some instances and not in others, depending on the presentation mode, especially when technology is often nothing more than a page-turner for a test.

Conclusion by the German Board of Assessment and Testing

To summarize the statement by the assessment experts of the German Board of Assessment and Testing lays out the core ideas of the MDR and portrays how it could be used to construe an argument for a classification of software-based clinical-psychological tests as medical devices. They go on to lay out the risks for actual test quality by highlighting the MDR's sole emphasis on reliability-related quality aspects. They also identify a core feature the MDR ascribes to medical devices which is that the results capture a specific quantity with a specific unit. These quantities and units (Figure 1) are common in the natural sciences, particularly in physics, chemistry, and biology but not in psychology. This part of the MDR constitutes one of the two pillars on the experts from the German Board of Assessment and Testing base their conclusion. The other pillar is the fact that software-based clinical assessments in most cases are the computer versions of paper and pencil tests. Of course, there are psychological assessments that have no paper and pencil version. However, their use of a computer's capabilities is usually that they present animations, sounds, or other stimulation that would not be possible using paper and pencil. The actual scoring is still done by assigning numbers to observations, and the software serves a purpose in the presentation, not the scoring itself. Moreover, the actual scores have to be interpreted using norms or cutoffs. This general rule also applies to more complex scoring methods (e.g., factor scores from CFAs, person parameters from IRT models, machine learning-derived scores, etc.). These two arguments led the assessment experts to conclude (p. 9):

> "Based on the information provided previously, it is concluded here that psychometric tests are not medical devices. This is also true in the case that the procedures are used in a software-based manner.
>
> The decisive factor for this evaluation is that psychometric tests can be used in a software-based manner, but they do not constitute software in their own right. Otherwise, every digitally administered patient questionnaire, which can also request information that is taken into account in the diagnosis (e.g., age, gender, drug consumption), would be a medical device of category IIa with the corresponding necessary quality assurance measures. … In the interpretation of the regulation, a legally relevant contribution of software should therefore be independent as well as substantial and rather technical in nature.
>
> Furthermore, it is stated that the focus of quality assurance for medical devices is on measurement accuracy. This is justified since measurements according to Directive 80/181/EEC have natural units, which are directly interpretable. This is not the case for psychometric tests, and the scores used therefore do not appear in Directive 80/181/EEC. An interpretation is made by comparison with an empirically obtained reference sample. This also applies to neuropsychological procedures that use reaction times, for example. For this reason, demonstrating the validity of this test score interpretation is of paramount importance for psychometric tests. Accordingly, the above guidelines for assessing the quality of psychometric tests place much emphasis on testing validity: does the procedure measure what it claims to measure? This difference between measurement in the medical versus psychological sense is of considerable importance and supports the statement not to classify tests based on measurement theory as medical devices."

We concur with this evaluation and agree that psychological tests should not be classified as medical devices. It has to be noted here that the advent of artificial intelligence and deep learning in the psychological assessment will have to be considered in this context as well (Fokkema et al., 2022; Iliescu et al., 2022). Such methods are software-based and typically use a large variety of data, not only answers generated by test takers in response to items. However, the vast majority of such applications are consonant with a supervised learning process that tries to mimic the results of a psychometric test or a clinically derived diagnosis. Thus, none of the quantities and units characterizing medical devices are likely to be the outcome of such tools. In addition, while this seems to protect supervised learner algorithms imitating a psychometric test from classification as a medical device, supervised learners mimicking an actual diagnosis or treatment plan, being much more comprehensive and also final in nature, are most likely not covered by the arguments presented here. The statement should therefore not be perceived as a *carte blanche* for psychological assessment per se but rather as a very specific recommendation for software-based clinical-psychological tests.

How to Move on?

These ideas, arguments, and considerations emphasize the fact that psychological assessment does not occur in a legal vacuum, to paraphrase a prominent quote. It is also clear that psychological associations and assessment organizations need to take a more active stance in advocacy for our domain, helping to inform the legal bodies that are writing such directives and laws. A simple first answer to the

question stated above the paragraph could be to join such associations and help them to address these issues by becoming an active member.

These questions also need to be discussed in the psychological world by a broad range of experts. If psychologists are not part of the debate in those delicate moments when rules are developed that may apply to them, now or later, they will be relegated to apply them without question – and this is a sobering thought. Therefore, we invite all of you to start such a discussion regarding MDR. For this purpose, we created an email address where you can send your thoughts and comments (discussions.ejpa@gmail.com). Your input will be monitored by the EiC which will suggest suitable contexts for further debate.

References

AERA, APA, & NMCE. (2014). *Standards for educational & psychological testing*. APA.

DIN 33430. (2016). *DIN 33430:2016-07, Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen* [Job related proficiency assessment]. Beuth Verlag.

EFPA Board of Assessment. (2013). *EFPA Review Model for the Description and Evaluation of Psychological and Educational Tests*. http://www.efpa.eu/professional-development/assessment

Emons, W. H., Sijtsma, K., & Meijer, R. R. (2007). On the consistency of individual classification using short scales. *Psychological Methods, 12*, 105–120.

European Commission. (2009, March). *Regulation (EU) 2017/745 of the European Parliament and of the Council on Medical Devices*. EC. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32009L0003&from=EN

European Commission. (2020, April). *Medical Devices Regulation (MDR)*. EC. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02017R0745-20200424

Evers, A., Sijtsma, K., Lucassen, W., & Meijer, R. R. (2010). The Dutch review process for evaluating the quality of psychological tests: History, procedure, and results. *International Journal of Testing, 10*, 295–317.

Fokkema, M., Iliescu, D., Greiff, S., & Ziegler, M. (2022). Machine learning and prediction in psychological assessment. *European Journal of Psychological Assessment, 38*(3), 165–175. https://doi.org/10.1027/1015-5759/a000714

Gignac, G. E. (2014). On the inappropriateness of using items to calculate total scale score reliability via coefficient alpha for multidimensional scales. *European Journal of Psychological Assessment, 30*(2), 130–139. https://doi.org/10.1027/1015-5759/a000181

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future – Festschrift in honor of Karl Jöreskog* (pp. 195–216). Scientific Software International.

Iliescu, D., Greiff, S., Ziegler, M., & Fokkema, M. (2022). Artificial intelligence, machine learning, and other demons. *European Journal of Psychological Assessment, 38*(3), 163–164. https://doi.org/10.1027/1015-5759/a000713

ISO 10667-1:2. (2020). *Assessment service delivery – Procedures and methods to assess people in work and organizational settings*. Beuth Verlag. https://www.iso.org/obp/ui/#iso:std:iso:10667:-1:ed-2:v1:en

Lindley, P. A., & Bartram, D. (2012). Use of the EFPA test review model by the UK and issues relating to the internationalization of test standards. *International Journal of Testing, 12*, 108–121.

Michell, J. (2001). Teaching and misteaching measurement in psychology. *Australian Psychologist, 36*, 211–218.

Muñiz, J., Fernández-Hermida, J. R., Fonseca-Pedrero, E., Campillo-Álvarez, Á., & Peña-Suárez, E. (2011). Evaluación de tests editados en España [Evaluation of tests published in Spain]. *Papeles del Psicólogo, 32*, 113–128.

Nezlek, J. B. (2017). A practical guide to understanding reliability in studies of within-person variability. *Journal of Research in Personality, 69*, 149–155.

Revelle, W., & Garner, K. M. (2022). Measurement: Reliability, construct validation, and scale construction. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology*. Cambridge University Press.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107–120.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677–680.

Witteman, C. L. M., Van der Heijden, P., & Claes, C. (2018). *Clinical assessment: Psychodiagnostic decision making*. De Tijdstroom.

Ziegler, M. (2014). Stop and state your intentions! Let's not forget the ABC of test construction. *European Journal of Psychological Assessment, 30*(4), 239–242. https://doi.org/10.1027/1015-5759/a000228

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*, 123–133

Published online March 27, 2023

Matthias Ziegler
Psychological Institute
Humboldt-Universität zu Berlin
Rudower Chaussee 18
12489 Berlin
Germany
matthias.ziegler@hu-berlin.de

Dragos Iliescu
Faculty of Psychology and Educational Sciences
University of Bucharest
Sos. Panduri 90
050657 Bucharest
Romania
dragos.iliescu@fpse.unibuc.ro

