Ethics in Progress (ISSN 2084-9257). Vol. 12 (2021). No. 1,
Art. #3, pp. 15-38. DOI:10.14746/eip.2021.1.3

Making Human Traits Visible, Objectively and Validly, through Experimentally Designed Questionnaires

Georg Lind

(University of Konstanz; Germany; georg.lind@uni-konstanz.de)
ORCID: 0000-0002-2235-5465

Abstract: Researchers who need valid and objective data for evaluating their theories or the efficacy of educational methods and programs have to choose between two equally undesirable alternatives: either they can use "objective" methods which have a questionable validity, or they can turn to "subjective" assessment methods with better validity. In other words, while subjective approaches to the study of human traits may be, or really are, valid, they lack objectivity, that is, they may be biased toward the researcher's theory. On the other hand, objective approaches to the study of psychological traits often lack psychological underpinning but are solely designed to fit a certain statistical model. Thus, we cannot know what these instruments really measure.

Here I present a new approach to the study of human traits, which claims to be objective as well as psychologically valid, namely the concept Experimental Questionnaire (EQ). An EQ lets us make traits visible without relying on dubious statistical assumptions. Thus, it makes it possible for the researcher to test the psychological theory underlying its designs. The EQ methodology is not only an idea, but it has been applied for constructing the Moral Competence Test (MCT) and for testing the assumptions about the nature of moral competence which were used to design it. So far, all the studies have clearly confirmed their validity. This makes the MCT suitable for testing hypotheses regarding the relevance and teachability of moral competence, and, therefore, also for evaluating the efficacy and efficiency of educational methods of fostering this competence.

Experimentally designed questionnaires can also be used in other fields of educational and psychological research in which testable theories about the nature of its objects have been developed.

Keywords: Psychological measurement; standardized tests; theory; objectivity; validity.

Prologue

In 1979, I had a chance to meet Paul Meehl, the co-author of the famous *Minnesota Multiphasic Personality Inventory* (MMPI). I admired him for his writings on methodological issues in psychology, like "When to use your head instead of a formula" (Meehl 1957). At that time, I was not sure whether my new idea about psychological measurement

would hold water. After listening patiently to my critique of mainstream psychological testing and my idea that should replace it, he cautioned me. He said that psychologists and educators would hardly welcome this new idea because if they did, they would have to give up the tests with which they make their living. Today, 40 years later, I know he was right. Fortunately he had added "Go on!" That encouraged me to write this paper.

The Persisting Dilemma of Psychological and Educational Measurement

Millions of dollars are spent every year on tests of character, academic abilities, vocational skills, mental disorders and so on, in the hope that their findings help to improve therapy, education and the politics of mental health and education (Gregory 2018, 22). But anyone who seeks the service of psychology (which translates to "science of the mind") faces a persistent dilemma. One has to choose between two opposite approaches to the observation and measurement of psychological traits, both of which have their drawbacks:

"Subjective" (also called "qualitative") psychologists argue that the focus on studying the internal structure of the human mind will indeed provide badly needed insights on the human condition. The human mind, they insist, can be studied only by using *subjective* methods like clinical interview.

In contrast, "objective" ("quantitative") psychologists argue that if psychology wants to be recognized as a science, it must use only *objective* methods of measurement. Notably, both agree that the *internal* factors of the human mind and its *structure* are out of reach for objective measurement. Can psychology really become a science if it spares the direct, objective measurement of its very objects?

For many years eminent scholars have argued that this deficit has prevented psychological and educational research from developing into a real science (Travers 1951; Loevinger 1957; Miller 1962) and from playing a more constructive role in evaluating and improving education (Schoenfeld 1999; Ravitch 2013).

For centuries, psychology and education were part of philosophy and, therefore, the domain of subjective science. Philosophers who focused on the nature of the human mind mostly used subjective methods for studying it. Their methods tended to be *ideographic* (acknowledging the individuality of the person) and *holistic* (taking the whole structure of the individual personality into account).

This philosophical approach to the study of the human mind was challenged in the 19th century by objective psychologists who were at home in physics, biology and medicine. They argued that psychological research must be *nomothetic* (searching for general laws) and *objective,* studying people's behaviors instead of the structure of their mind: "The behaviorist recognizes no such things as mental traits, dispositions or tendencies," postulated Watson (1970/1924, 98), the founder of psychological *behaviorism,* which is

still very influential. He and his followers believe that psychological measurement should focus on behavior instead of on psychological traits: "A test is a standardized procedure for sampling behavior and describing it with categories or scores" (Gregory 2018, 23). Their object is not genuinely psychological but only somehow "related" to psychology: "We define psychological assessment as the gathering and integration of psychology-related data" (Cohen & Swerdlik 2018, 2).

This antagonism of the two approaches has caused a deep "crisis of psychology" (Bühler 1927) which hampers the progress of psychology as a science to this day. As the philosopher Wittgenstein (1953) noted: "In psychology there are experimental methods and conceptual confusion. The existence of experimental methods makes us think we have the means of solving the problems which trouble us; though problem and methods pass one another by." Eminent psychologists agree. Similarly, Graumann wrote: "Theoretical frameworks and methodological convictions are still too divergent, if not partially incommensurable" (Grauman 1960, 146, my transl.). Block (1977) also asserted that "perhaps 90% of the studies are methodologically inadequate, without conceptual implication, and even foolish" (Block 1977, 39). The educational researcher Travers observed "that the rather meager advances made in many areas of psychological measurement during the last 20 years are mainly a consequence of the fact that these areas are staffed mainly by technicians interested in producing useful instruments and not by scientists interested in expanding knowledge" (Travers 1951, 137). The statistician and psychologist Kempf wrote: "What usually is called psychological test theory is actually a statistical theory of item selection in order to produce a test with some desirable features" (Kempf 1981, 3, my transl.). Ten years later, Alan Schoenfeld (1999), former president of the *American Educational Research Association* (AERA) and an accomplished educational researcher and mathematician, complained that still "virtually none of the current assessments in wide use are grounded in well-developed theories of competence" (Schoenfeld 1999, 12). Therefore, he called for a moratorium on standardized testing until this basic issue has been solved. More and more educational researchers, teachers, parents and educational policy makers question the meaningfulness and validity of standardized testing (Amrein & Berliner 2002; Ravitch 2013; Sjoberg 2017; Koretz 2017).

Yet not much has changed. Textbooks on psychological tests and measurement do not respond to any of these complaints (Gregory 2018; Cohen & Swerdlik 2018) In psychological measurement, it seems, we have to choose between Scylla and Charybdis, that is, between a psychological object which cannot be measured through objective methods, on the one hand, and an objective method which rejects psychological objects, on the other.

How can we overcome this impasse? Is it really not possible to study the human mind objectively without giving up its genuine object?

The Critical Role of Theory in Measurement

In everyday life we measure all kind of things by reading a scale like a meter stick without much thinking. But we should remember that before we used the meter we had other means of measurement like our hands, feet or elbows. Usually, we do not give it any thought that measurement is something artificial, that is, something which is based on conventions and theoretical assumptions. But "there is no measurement without a theory and no operation which can be satisfactorily described in non-theoretical terms" (Popper 1968, 62). The theoretical assumptions concern, for example, the stability of the material of which the meter stick is made. A rubber band would not be suitable. If the stick is made out of metal, the surrounding temperature might cause the stick to constrict or expand and thus bias our measurement. Using a thermometer requires that the expansion of the fluid inside the instruments expands strictly proportionally to the surrounding temperature. Research shows that this is true only within a certain range of temperature. Outside this range, the thermometer gives incorrect numbers.

The same is true for psychological measurement. When we "read" people's intelligence, morality, political attitudes etc. from their visible behavior, this reading is also based on theoretical assumptions, namely assumptions about the relationship of observable behavior to the things we are interested in. As I will discuss below, even when the relationship between a certain behavior and a certain trait looks simple as in the case of classical attitudes scales, attitude tests produce ambiguous data. For example, a score in the middle range of a conservatism scale can mean that the participant has a "middle attitude toward conservatism." But it can also mean that he or she has no attitude toward conservatism at all, or that he or she has a high differentiated attitude (Scott 1968). The relationship between overt behavior and underlying traits can be even more complex when we look at the relationship between answers to a clinical interview and, let's say, a participant's stage of moral development (Lind 1989).

Measurement theories are the link between reality and our theories of reality. If measurement does not provide us with valid data about reality, our thinking and our decisions will be misled by false images of the world. Hence it is essential that measurement theories are testable and that we actually do test them. Only if our measurement provides valid data, can we trust them and use them for examining the empirical truth of hypotheses about the relevance, determinants and teachability of psychological traits.

will now discuss the two main approaches to measurement in current psychology: the "subjective" and the "objective" approaches. As I mentioned, both have severe shortcomings.

The "Subjective" Approach

Subjective psychologists base their measurement on the assumption that our

behavior is determined mainly by unconscious affects and cognitions. In other words, they believe that only through the study of the unconscious level of our mind can we really understand human behavior and make education, therapy, and politics more effective. They also believe that unconscious affects and cognitions cannot be assessed directly but only indirectly, namely through interpreting people's visible performances in certain situations or their answers to the psychologist's questions. Interpretation means that measurement requires researchers to clearly define their object clearly and concisely, and to make assumptions about its nature in behavioral terms, so that these assumptions can be objectively tested. These assumptions or hypotheses should be based on coherent theories which have been tested by different researchers.

A scientific definition of a psychological object should allow us to examine the truth of the assertion that a test is a valid measure of that object. Unfortunately, in psychology the object of measurement is rarely defined in a clear and concise way. Rather definitions are tautological, fuzzy, or evasive definitions and, therefore, do not allow us to judge a test's validity. For example, "intelligence" is often tautologically defined as "what is measured by intelligence tests" (Bailey 1994, 57); or its definition is vague and ambiguous, like this: "One of Sternberg's very succinct definitions of intelligence states: Intelligent behavior involves adapting to your environment, changing your environment, or selecting a better environment"[1]. A psychological definition is not "succinct" if it states only what is "involved." Moreover, if several "definitions" (plural!) are available, confusion is inevitable. A definition is evasive if it encompasses everything a person does: "the aggregate or global capacity of the individual to act purposefully, to think rationally, and to deal effectively with his environment" (David Wechsler). No test can measure such a "global capacity", but can only assess a small section of it. The claim that such a section is representative for the whole is not testable.

However, there are a few exceptions. Take for example Lawrence Kohlberg's research on the nature of moral judgment competence. He defines it "as the capacity to make decisions and judgments which are moral (i.e., based on *internal* principles) and to act in accordance with such judgments" (Kohlberg 1964, 425; emphasis added). "Morality (...) defined as 'having moral principles' includes (…) inner conformity to them in difficult situations rather than outer conformity in routine situations" (Kohlberg 1958, 3).

Kohlberg's definition is short and clear: (1) It defines the affective aspect of morality in terms of the individual's *inner* moral principles or orientations, (2) It defines the cognitive aspect of morality as the *structure* of his or her overt moral judgments, and (3) It defines both as *aspects of visible action* or *behavior*. His definition of a psychological trait is distinct from most studies of morality (and other human traits): "The trouble with such [studies] is that they describe the person externally in terms of his impact on and relation to his culture and to the observer's values. They will not tell us how the individual thinks, what values he *actually* holds" (Kohlberg 1958, 82; emphasis added).

1 https://thesecondprinciple.com/optimal-learning/sternbergs-views-intelligence/

But how can we make an "inner" trait visible? When Kohlberg started his research, there were only subjective methods available. Kohlberg followed Piaget's lead and developed his clinical interview method, the *Moral Judgment Interview* (MJI). In this, the interviewer confronts participants with several dilemma stories in which the protagonists have to make a presumably difficult decision: Whatever they decide, they violate a moral principle. The participants should say whether they agree or disagree with the protagonist's decision and why. The interviewer follows up their answers to get a rich picture of their reasons, and also probes into counter-arguments: Which reasons could justify the opposite opinion? The answers of the interviewees are recorded, transcribed, and then categorized by a trained scorer into one of the six "cognitive-developmental stages" which Kohlberg (1984) had defined. In the early version of the MJI, the scorer was to read through the complete answers of a participant and then decide which "stage" of moral reasoning would best fit them.

Kohlberg based his method on two assumptions or postulates, namely that people's moral cognitions are organized as a structural whole and that they develop in a pre-determined invariant sequence. He considered the scoring to be valid only if it agrees with these two postulates. Since the interview data did not agree well enough with these two postulates, he and his students revised the scoring system several times in order make it better fit with the data (Kohlberg 1984).

Inevitably, Kohlberg's measurement model came under attack from objective psychologists. These questioned the MJI's "objectivity" and "reliability" (Kurtines & Greif 1974). They argued that morality must be assessed (1) Through judging their behavior by external standards of morality (instead of talking to them about their behavior and other people's behavior), and (2) By scoring small pieces of behavior instead of looking at it holistically.

In response to these attacks, he and his collaborators also changed the method of scoring in order to make it more agreeable with the requirements of objective psychology (Colby *et al.* 1987). But they not only changed the method but actually jettisoned Kohlberg's original concept of moral competence:

• While Kohlberg (1958, 1963) originally defined moral behavior as behavior which is based on *internal* moral principals, the MJI uses *external* standards for scoring the individuals' responses: "I include in my approach a normative component. (...) That is, I assumed the need to define philosophically the entity we study, moral judgment, and to give a philosophic rationale for why a higher stage is a better stage" (Kohlberg 1984, 400; see also Lind & Nowak 2015).

• Originally Kohlberg based the Stage scores on a holistic analysis of people's total response pattern. He considered *structure* as the hallmark of his cognitive-structural approach: "The structure to which I refer is (...) a construct to subsume a variety of different manifest responses. The responses of subjects to the dilemmas and their subsequent responses to clinical probing are taken to reflect, exhibit, or manifest

the structure" (Kohlberg 1984, 407). Later he and his colleagues fragmented the interviewees' response into "items." They instructed the scorer to score each item individually instead: "Each item must have face validity in representing the stage as defined by the theory" (Kohlberg 1984, 410). However, through this itemizing of the scoring method, the cognitive-structural properties of an individual's response pattern become invisible. To reclaim some of their original structural idea, they instructed the scorer to put an answer into a higher stage, if it was "included in a higher stage idea." They argue that "ideas are often expressed within a higher stage context, which when taken literally or out of context would be scored at a lower stage" (Colby *et al.* 1987, 177). This attempt to save their original cognitive-structural feature of the method came under attack by objective psychologists again (Rest 1979, 60).

• Whereas Kohlberg formerly defined moral competence as an *ability* and tested this ability by confronting the respondents with tough probing questions and counterarguments, later he and his students eliminated these tasks in the revised MJI, presumably for the same reasons as for itemizing the scoring; namely to increase the statistical "reliability" of the MJI: "Test reliability and test construct validity are one and the same thing" (Kohlberg 1984, 424).

Similarly, objective tests of moral development like the *Defining Issues Test* (DIT) by Rest (1979), the *Socio-Moral Reflection Measure* (SRM) by Gibbs *et al.* (1992), and Haidt's (2007) test of moral values even take this accommodation of the definition of moral competence to standardized testing a step further: They score their respondents' answers in regard to *external* standards. Moreover, while some claim that they assess the *structural* properties of people's moral judgment, their scoring methods chop up the person's structure of judgments into atomized items, thus making structural information invisible. Some even claim to measure moral *competence*, but their tests lack any moral task.

You can't have your cake and eat it. Kohlberg agreed to improve the "reliability" of his clinical interview method at the expense of its theoretical validity (Lind 1989). This means that he actually gave up his original concept of moral competence in order to comply with the doubtful psychological assumptions underlying the so-called objective approach.

The "Objective" Approach

"Objective" psychologists believe that psychological theories bias measurement and that we better do without them. Measurement, they demand, should be based only on visible acts or behavior, but not on theory (Watson 1923).

However, they cannot in fact avoid theoretical assumptions. Instead of psychological theory, they base their measurement on *statistical theory*. This theory determines which

items and which scoring methods are regarded as valid. In other words, their statistical models define their object of measurement. Statistical models, they seem to believe, are more objective than psychological models. But they are not, as we will see.

The famous *Studies in the Nature of Character by* Hartshorne and May (1928) are a good example of objective behaviorists' approach to psychological measurement. Funded by a church organization, they wanted to test experimentally the hypothesis that character exists and that it is fostered through religious instruction. They confronted participants with situations in which they were tempted to cheat and observed how they reacted. They recorded the agreement or disagreement of these reactions with their standard of honesty. They explicitly discarded any psychological and philosophical interpretations of their subjects' behavior, because "no progress [of psychological science] can be made, however, unless the overt act be observed and, if possible, measured without any reference, for the moment, to its motives and its rightness or wrongness" (Hartshorne & May 1928, 11).

Obviously, the authors believed that we are able to read the character strength of a person directly from his or her reactions to temptations to deceive, like reading a temperature scale: the current temperature is simply the reading on the display plus/minus some error of reading or of malfunctioning of the scale. Similarly, they believe that we can reduce measurement error just by reading those reactions several times and calculating the average score.

However, objective measurement is also based on a theory, not on psychological theory but on a statistical theory such as "Classical Test Theory" or the *Theory* of *Mental Tests* (Gulliksen 1950), and its variants like "Item Response *Theory*," IRT. Notably, their theory is not about psychological objects but about statistical constructs, for example, about "latent variables," "latent classes," or statistical "factors." Through this theory, they create their own object of measurement, which may best be described as a "*homo statisticus*." Although the textbooks on CTT and ITR are usually voluminous (e.g., Cohen & Swerdlik 2012, has 612 pages), this homo statisticus is described by a very simple statistical formula: Y = T + e. This formula means that the reading of the scale ("Y") is simply the addition of the subjects' "true" behavior ("T") and some random error ("e"). The formulas of more sophisticated statistical test theories are more complex but are essentially based on similar statistical assumptions (Wilson 2005).

Can this *homo statisticus* be used to understand, predict and enhance human behavior? Can we use this statistical construct, for example, for examining the empirical validity of psychological theories of intelligence and morality, or for judging the efficacy of therapeutic and educational programs, or for evaluating students' achievement? The answer is no. This becomes obvious when we translate the statistical formulas underlying this construct into plain language. They allege that:

• *Observation is simple*. Objective psychologists believe that we can directly read the participants' behavior without any psychological interpretation. As we have seen above, they believe that, for example, the participants' behavior in an honesty experiment

enables us to directly read his or her character. By definition this behavior is not affected by any other factor like the type of temptation, the participants understanding of the test, or by their moral competence. To use another example, if a person gets 105 points on an IQ test, by definition it means nothing other than that this person really has an intelligence of 105 plus/minus some random error of the test.[2] No other interpretation is considered.

• *Error is random*. Any aberration of this statistical model from the real data is believed to be caused only by a *random* error of "reading the scale," meaning that no systematic factor of the participant or of the testing circumstances affects our reading of human behavior.

• *Repeated observation of identical behavior is possible*. Because objective psychologists believe that any error is purely random, it averages to zero. Therefore, they assert, measurement error can be simply reduced to any smallness simply by repeating the reading as often as needed (the so called "law of large numbers"). But this requires us to believe that people respond to replications of the questions or task always in the same way, and that they are willing to do so. But not even objective psychologists seem to believe this. They hardly ever confront participants with identical questions or identical tasks. They reason that people would remember their answers, or refuse repetitions. So even behaviorists admit that there are internal factors (like remembering, thinking, vigilance) interfering with observation and that, therefore, variation of behavior should not be regarded just as random error.

• *Similarity of behavior can be determined purely statistically*. Because objective psychologists avoid psychological concepts, they use statistical means for defining the "similarity" of tasks and questions. They define two behaviors as similar if the participants show them together. So, for example, if people answer two different questions in the same way, they are considered similar, or, if they solve task A and also task B in a math test, these two tasks are considered similar. If the items do not show statistical similarity, they are excluded from the test even though they may be considered as highly valid by experts on the subject matter. Note that if all test items which threaten the reliability of the test are excluded from the test as "dissimilar," the measurement model becomes immune to refutation through data. This immunization violates a basic standard of good science, namely refutability (Popper 1967). It also calls the objectivity of objective psychology into question and creates an illusionary reality. For example, Burton (1963) argued that the studies by Hartshorne and May (1928) would have actually proven the existence of a uniform character if the researchers had eliminated all experiments from their analysis that were "unreliable." In other words, Burton reasons that there are two groups of people: those who are always honest and those who are always dishonest, and never in between, in all thinkable situations – *except* in all those situations in which they

2 Variants of the CTT like Items-Response-Theory are more complex but rest basically on the same idea, namely that the behavior results from a random process (Allerup 2007).

behave differently.

• *Error and reliability are an attribute of the measurement instrument*. If they were an attribute of the measurement, they would not change from one application to another. But they do. Item selection does NOT lead to a stable estimate of a test's reliability, but it varies from one test sample to another and from one test administration to another. For example, even though PISA tests are carefully trimmed on the basis of many prior studies and the replacement of "unreliable" tasks, the final tests still deviate substantially from the statistical model on which their construction was based (Wuttke 2006; Jablonka 2007). If data change, it is not because of the tests. Tests are mostly, if not always, perfectly stable. Just observe a printed copy of the PISA-test for some months: you will find no change!

Objective psychologists like to compare their tests to the measurement by craftsmen and astronomers. Carpenters usually read their meter stick twice. This is enough to make sure that they do not accidently saw the beams for a house construction too long or too short. The one-time repetition has the advantage that it hardly affects its object (although their yard stick may leave some marks behind) and that the interval between the two readings is so short that the object does not change during the repetition. Observing human behavior is much trickier. Do we really always read the same thing when we repeat our observations like a carpenter does? In certain contexts, it may suffice to repeat a test question only once to make sure that it is correctly recorded. But in contrast to the carpenter's wood, people try to make sense out of test questions. So people may feel annoyed when being asked the same questions twice without a cause. For example, if we ask a person twice how she feels, she will answer the second question only if we explain that we did not understand her first answer, or that we wanted to observe change. But in the latter case, a different answer does not indicate an error but a change of feeling. In these cases, the repetition of the test question does not produce random error, but rather a systematic change of behavior.

Astronomers repeat their measurements more often. They do this because they want greater precision than a carpenter. Since many observations require a longer period, their targeted star (or the Earth) may move in the meantime, and their data reflect not only random measurement error but also a change of location. This will bias their measurement and the repetitions do not average to zero. Astronomers can differentiate such systematic influences from random reading error by looking at the distribution of their data. Only in as far as their data are distributed like a bell do they consider them to be caused only by reading error.

In contrast, objective psychologists usually avoid testing the hypothesis of random error and thus they overlook any systematic bias and ambiguity of their measurements (Wuttke 2007; Jablonka 2007). They may overlook, for example, as Scott (1968) showed, that scores in the middle range of an attitudes scale can have three very different meanings:

They can mean, as researchers mostly assume, (1) that the respondents have a medium attitude toward the declared object of the scale (like "conservatism"). But these scores can also mean (2) that they do not have such an attitude at all, but instead rate the items in regard to other criteria. Or these scores could mean (3) that the respondents have a differentiated attitude which involves more than the one attitude.

In order to clarify this ambiguity, I re-analyzed the findings from a longitudinal study on university students' political attitudes (Lind 1985a). The authors of this study reported that at in the first semester, students' attitudes become more liberal, and after graduation they become more conservative again. They interpreted these changes of students' attitudes as a consequence of their adaptation to different environments, which presumably changed from conservative to liberal (university) and than back again to conservative (workplace).

However, my secondary analysis of their statistics for "measurement error" over the span of university study, actually revealed a *structural* transformation of students' political attitudes: first the error was large, then decreased and then increased again. This supports the hypothesis that students had hardly any "conservative" attitude when they entered university. Then they developed a consistent (liberal) attitude, and finally their attitude became more differentiated, so that "measurement error" increases again and the scores moved back again to the middle of the conservatism scale. In other words, the students did not just adapt to their environment but also developed a higher competence for political reasoning. It was the researcher's statistical model which made students' structural development look like a pure "to and fro" of attitudes.

The blindness of objective psychology to structural aspects of human behavior explains Hartshorne and May's (1928) failure to produce evidence for the existence of character. Only after the completion of their study did Hartshorne and his colleagues admit that excluding internal traits from their observations was a mistake: "The essence of the act is its pretense. Hence [character] can be described and understood only in terms of the human elements in the situation. It is not the act that constitutes the deception, nor the particular intention of the actor, but the relation of this act to his intentions and to the intentions of his associates" (Hartshorne & May 1928, 377) The authors also admitted the blindness of their measurement model to the competence aspect of character: "A trait such as honesty or dishonesty is an achievement like ability in arithmetic, depending of course on native capacities of various kinds" (Hartshorne & May 1928, 379). Already some years earlier, the psychiatrist Levy-Suhl (1912) was surprised to find in his study of juvenile delinquents that they upheld the same moral values as non-delinquent youth. Therefore he hypothesized that they actually must differ in respect to their moral maturity, which psychologists were not able to measure at that time.

Another example for the discrepancy between the statistical measurement model and psychological reality is the OECD's *Programme for International Student Assessment* (PISA). The physicist Joachim Wuttke (2007) found much "evidence for multidimensiona-

lity," that is, for the existence for several internal factors. This contradicts the measurement model on which the PISA tests are based. The evidence, he notes,

> is even more striking on the background that the cognitive items actually used in PISA have been preselected for unidimensionality: Submissions from participating countries were streamlined by 'professional item writers,' reviewed by national 'subject matter experts,', tested with students in think aloud interviews, tested in a pre-pilot study in a few countries, tested in a field trial in most participant countries, rated by expert groups, and selected by the consortium (…). Only one-third of the items that had reached the field trial were finally used in the main test. Items that did not fit into the idea that competence can be measured in a culturally neutral way on a one-dimensional scale were simply eliminated. Field test results remain unpublished, although one could imagine an open-ended analysis providing valuable insight into the diversity of education outcomes. This adds to Olsen's (…) observation that in PISA-like studies the major portion of information is thrown away (Wuttke 2007, 249–250).

If objective psychologists used this thrown-away information they could interpret respondents' test scores more adequately. They would discover, for example, that the same task which is designed to challenge the respondents' *math competence* might actually challenge quite different dispositions, namely their ability to guess the "right" answer, their ability to copy it from other test takers, their knowledge of how to handle tests (test skill), their ability to stay awake on long testing cycles, and their ability to master their test anxiety, just to name a few of the factors which can influence a testee score. Or they might discover that "wrong" answers do not indicate a lack of math competence but that the testee made only a small error, or was not able to read the often wordy instructions quickly enough or was blocked by test anxiety. (Wuttke 2007) Similarly, behaviorist psychologists, who operationally define participants' moral character as an "honest" reaction to a situation of temptation, give them a high score regardless of whether or not these actually have high moral standards, or only incidentally acted "honestly" in this situation, or succeeded without the need to cheat because they knew all the answers (in fact cheating correlated negatively with IQ), or wanted to help a friend by letting her copy their test answers. So these scores have highly ambiguous psychological meaning.

Calling all these possible causes of test scores "random error" prevents any improvement of these tests and any progress of psychology as a science (Rosenthal & Rosnow 1997; Loevinger 1957). Moreover, it also undermines the trust in the validity of these tests. How can we expect consumers to trust tests, when even the chosen "test format or method of assessment can cause large differences in student scores?" (Walberg *et al.* 1994, 232) How can we rely on expensive studies like PISA for educational policy-making if it "is dominated and driven by psychometric [i.e., statistical] concerns, and much less by educational," writes the nuclear physicist Sjoberg (2007, 212).

How can we call these tests "psychometric" if there is no "psycho" in its metric? While the physical units of measurement are physically defined and standardized, the units of "standardized psychometric tests" are not defined psychologically and are not standardized objectively but only statistically. Their metric changes with the data of each

study, like rubber bands which stretch and bend as needed but are not *reliable* in the true sense of this word.

In spite of their blindness in regard to psychological theories, objective "psychologists" claim that their statistical models can be used to evaluate psychological theories, therapeutic methods, educational policies and competencies of people. They underpin this claim with a naming trick: (a) They give their statistical constructs psychological names like intelligence, character, or conservatism, and (b) They equate pattern of correlations across groups of people with an individual mind's "structure." But like family names, these names do not actually establish a real relationship between statistics and psychology. Or would Mrs. Miller allow an unrelated Mr. Miller to share her bedroom, just because he bears the same family name?

Anyway, this trick seems to work. World-wide, millions of dollars are spent every year on "objective" tests of academic abilities, vocational skills, character, mental disorders, and so on, in the hope that they can help to improve therapy, education and the politics of mental health and education. These tests have severe consequences for millions of students, job applicants, career seekers, mentally ill people, teachers, educational policy makers and many more who are tested many times throughout their lives, and also for decision-makers who base their policies on reported test scores. Because these tests measure something different from what they pretend to measure, they can cause a lot of damage. If these tests are bad, they will mislead us when we use them to evaluate methods and policies of therapy and education. If, for example, bad teaching practice produces higher scores on these tests than good teaching practice, they will defeat our educational system (Sjoberg 2017).

The dilemma of objective psychologists, it seems, is rooted in the ambiguous meaning of the word "objective." This word can take on quite different meanings:

• *Transparency*: This is an essential requirement of real science and good psychological practice. Only if data collection and scoring are fully transparent and uniform can they be critically examined by third parties. The questions and tasks of objective tests are usually transparent but often not available for the independent experts. The scoring of the answers is obscure for the customers. Instead of reporting the numbers of solved tasks, the scores are multiplied to make differences look large, and are transformed to make them look like a bell-shape. Ironically, the bell-shape indicates that the scores are pure error scores. Natural traits are hardly ever distributed like that: "An investigation of the distributional characteristics of 440 large-sample achievement and psychometric measures found all to he significantly non-normal at the alpha .01 significance level" (Micceri 1989, 156; see also Walberg *et al.* 1984). Finally, test scores are often obscure because important information like item selection and participants' attrition rates is held back.

• *Freedom from theory*: To be objective we need an object. Theories are an essential

basis of any measurement. If we want to measure a psychological object like an orientation or a competence, we need a psychological theory to define its nature. If we ban any psychological theory from our measurement method, we deprive it literally of its object. The test's reliability and precision become meaningless and its results useless.

• *Statistical standards*. The term "standardized" in the word "standardized tests" is actually a misnomer, because it does not mean that an individuals' test scores depend on a fixed standard but it means that an individual's scores depends on other people's test scores, namely how they compare to the scores of some sample of people. Such relative "standards" suggest wrong interpretations. For example, if a student solves ten more tasks on a test than he did last time, and if at the same time the members of the standardization sample also solve ten more tasks, his score (e.g., "percentile") will not increase and thus make him look as if he did not learn anything.

• *External or internal standards*. Objectivity is often used to mean external standards for scoring the participants' responses. However, objectivity can also require that we score a test in regard to the individuals' own standards, for example, if we want to measure how much progress they have made in regard to their *own* learning aims, or if we want to measure moral competence, which is defined as behaving in accordance with one's *own* moral principles.

How to Make Psychological Traits Visible

As I have shown, both mainstream approaches to measuring psychological attitudes and competencies are questionable. It has been often suggested to ease these problems by combining them in educational research and evaluation. But combining two bad meals does not make a good dinner. We should rather seek to find a better way of measurement which can replace the currently used ones.

As we have seen above, objective psychologists assume that *internal* psychological traits are not directly observable, and that structure is irrelevant, that is, that individual responses to test questions are unambiguously revealing the human trait under investigation and no other ones. In contrast, subjective psychologists target internal psychological traits can be made visible only by subjective methods on subjective ratings instead of on direct observations. But, as Jean Piaget (1965) admits, this is not a solution: "The point, then, that we have to settle is whether the things that children say to us constitute, as compared to the real conduct, a conscious realization (...), reflection (...) or psittacism (...). We do not claim to have solved the problem completely. Only direct observation can settle it" (Piaget 1965, 115).

Already a hundred and fifty years ago the Dutch psychologist Franciscus Donders (1969/1868) showed how we can *directly* observe psychological traits. He was probably the first who discovered that we can test measurement hypotheses in the same way as we test hypotheses about the impact of external factors on human behavior. He hypothesized

that humans are not merely machines who always react to stimuli like an automaton, but that they also *think* when it is needed. To test this hypothesis he designed a simple experiment for which he constructed an ingenious time recorder for measuring very short reaction times. When he gave his participants clearly distinct stimuli, they reacted as quickly as an automaton. But when he gave them similar stimuli, they presented them with a "dilemma", so that their reactions took much longer. Obviously under the second condition they had to *think* before reacting.

The problem of the ambiguity of human behavior has been solved in principle by the Hungarian-American psychologist Egon Brunswik (1955). He has shown how we can disentangle multiple factors of behavior with what he called the "diacritical method." His idea was that we must design our observation as a multivariate experiment, in which the traits that are believed to determine particular responses are used as "design factors." Only such a structural experimental design of our observation can make the determining traits.

On the basis of Donders and Brunswik's ideas, I have developed the concept of *Experimental Questionnaire*, EQ (Lind 1982; 2019). EQs make human traits directly visible without involving dubious assumptions. They do not require statistical expertise in order to see the trait under investigation. The translation of the visible results into numerical scores is only done in order to facilitate the statistical analysis of mass data.

EQs confront the participants with a carefully designed *pattern* of stimuli, tasks, questions, or situations. The pattern is designed as an individual *multivariate experiment*. The design-factors of this multivariate experiment are chosen to directly correspond with the dispositional factors that are hypothetically involved in the participant's response to those tasks, questions or situations. That is, the construction of the design-factors of an EQ requires a psychological theory about the measurement object. When the design-factors of EQs are chosen to be independent of each other, we can literally see the impact of each hypothesized factor on a participant's responses in the pattern of an individual's responses, in a similar way as we can read brain activities from the monitor of a brain scanner.

On the base of this new methodology, I constructed the first objective *Moral Competence Test*, MCT (formerly called *Moral Judgment Test)* (Lind 1978, 2019). After reviewing a vast amount of research (Lind 1985b; 2002: Lind & Nowak 2015) and considering modern ethical theories (especially, Habermas 1990), I have defined moral competence *as the ability to solve problems and conflicts on the basis of moral principles through thinking and discussion instead of though violence, deceit, or complying with others* (Lind 2019). Like Kohlberg (1963) I consider as criterion for moral competence inner moral orientations instead of external standards. Moral competence, as the MCT defines it, is the ability to behave in accordance with one's own moral principles instead of with conformity to other people's judgments.

More specifically, moral competence becomes visible when a person judges the

arguments concerning a controversial decision in regard to the arguments' *perceived moral quality* instead of their *opinion agreement*. As Keasey (1974) observed in a series of experiments, this ability seems to be low in most people. So we decided to use this task for measuring moral competence. If people have not developed such a moral sense, they cannot solve problems und conflicts through moral thinking and moral discussion but must use violence, deceit or submission to others.

In order to make visible people's ability to *rate arguments in regard to their moral quality instead of to their opinion-agreement or to the particular context*, we used three hypothesized traits as design-factors for the MCT:

(1) *Dilemma context*: The participants are confronted with two stories in which a protagonist has to make a difficult decision. They are asked to take sides: Was the protagonist's decision right or wrong?

(2) *Opinion-agreement*: After each story the participants are to rate several arguments supporting and opposing their own opinion. They should say how much they reject or accept them on a scale from -4 to +4).

(3) *Moral quality*: All the arguments have all been painstakingly written so as to represent a clearly distinct *moral quality*, namely one of the *six types of moral reasoning* described by Kohlberg (1984). In order to secure their *theoretical validity*, the arguments were reviewed by several experts of Kohlberg's stage-typology and then revised accordingly (Lind 1978; Lind & Wakenhut 2010).

Thus each item of the MCT represents a specific manifestation of the three dispositional factors which may determine people's judgment. The items of the MCT have a 6 x 2 x 2 multivariate experimental design. Due to this experimental design, we can literally see the respondent's degree of moral competence directly by looking at their *pattern* of responses. Only the whole pattern of a respondent's behavior contains the structural information which defines moral competence. If we looked only at isolated responses, we would not be able to see their structure. Isolated items are bare of any structure.

In order to facilitate further statistical analysis, one can translate this visible pattern into the numerical C-score (C for competence). The C-score is the proportion of individual judgment variation caused by the moral quality of the arguments as compared to the respondent's total judgment variation. The C-score ranges from 0 to 100, the higher scores indicating higher moral competence. Mean C-scores are rarely higher than 30, indicating that for most people it is indeed rather difficult to engage in a moral discourse. Most people judge arguments mostly, or even solely, on the basis of their agreement with their opinion.

The measurement theory on which the MCT is based can be rigorously tested without saving circularity. As already mentioned, the content validity of its items (arguments) has been examined through ratings by several experts of Kohlberg's typology

of moral orientations. The structural validity of the MCT can be experimentally tested in regard to four prominent hypotheses of cognitive-developmental theory:

• *Competence nature of morality*: In contrast to many other psychologists, Piaget (1965) and Kohlberg (1958, 1964) hypothesized that moral behavior is not only affective in nature but also cognitive, that is, it is not only determined by people's moral orientations (values, attitudes, principle, and so on) but also by their moral competence. This competence hypothesis has been clearly supported by experiments. While participants can be instructed to fake their moral orientations upward (Emler *et al.* 1983), the same kind of instruction fails to make participants fake their MCT's C-scores upward (Lind 2002). Experiments also showed that the ability to estimate other people's moral competence is positively correlated with their own moral competence (Wasel cited in Lind 2002).

• *Moral competence is a unique skill*. The ability to solve moral dilemmas is not just a linguistic skill but is a unique competence. This has been shown by the research team of Kristin Prehn (2013) of the Charité Hospital in Berlin. Moral competence as measured with the MCT correlates highly with neural activities in the right dorso-lateral prefrontal cortex (DLPC) when subjects' brain activities are studied in a brain scanner while they are confronted with moral problems: the lower their C-score is the longer their right DLPC is busy. This phenomenon does not show when the subjects are confronted with linguistic problems.

• *Hierarchical preference order*: Kohlberg (1958, 1984) and Rest (1969) hypothesized that the six types of moral orientations – which were identified on the basis of philosophical analysis – form a universal order of moral adequacy. This hypothesis lets us predict that people will prefer these types according to their order. This, as Karl Popper (1968) would say, is a very informative, because daring, hypothesis. Since the six types of orientation can be ordered in 720 (= 6!) different ways, the risk of a coincidental confirmation of the hypothesis is very small ($p = 1/720 = 0.0014$). Note that this risk is much smaller than the risk of accidently confirming a conventional statistical hypothesis ($p < 0.05$). The risk of accidental confirmation becomes *extremely* small if, for example, we test this hypothesis with ten people ($p = 0.0014^{10}$). It is even more astonishing that his hypothesis has been almost unanimously confirmed in many empirical studies (Lind 1986; 2002).

• *Simplex structure of moral orientations*: Kohlberg (1958) hypothesized that the correlations between the six types of moral orientation show a "simplex structure," which means that neighboring orientations correlate more highly with each other than with more distant orientations. The many MCT studies support this hypothesis with very few exceptions (Lind 1978, 2002).

• *Affective-cognitive parallelism*: Piaget (1976) hypothesized that affective and cognitive aspects of behavior are "parallel." This hypothesis has two important implications. First, Piaget saw orientations and competences not as separable components, but as two *distinguishable aspects of behavior*. This means that he rejected the prevailing notion

that human traits are components which can be separated from each other and from behavior, and can be measured separately. So all attempts to assess them separately are in vain. Second, his parallelism hypothesis lets us predict that the higher people's moral competence is, the more clearly they will prefer higher types of moral orientations, and reject low, inadequate types. Only with the MCT does it become possible to test Piaget's hypothesis, because only this test allows us to measure affect and cognition as distinct but inseparable aspects. So far, all MCT studies have very clearly supported Piaget's parallelism hypothesis (Lind 2002, 2013).

The exceptionally clear confirmation of these four core hypotheses shows that moral competence is something real and than it can be made visible in an objective and valid way. This gives us the opportunity to test hypotheses about the relevance, development and teachability of moral competence in an objective and unbiased way.

- *Relevance:* Already studies using Kohlberg's *Moral Judgment Interview* have found that moral competence determines our social behavior more than any other psychological trait. Experimental and correlation studies using the MCT confirm and extend these findings. Moral competence seems to be highly instrumental for such important behaviors like helping people in distress, engaging in democracy, obeying the law, respecting a contract, blowing the whistle, fulfilling academic achievement requirements, and making quick decisions (Lind 2019). A C-score above 20 seems to be critical. Only when people have a moral competence higher than 20 does their behavior in experiments show some determination by inner moral orientations. People who lack any moral competence either conform to the perceived opinion of the majority of people or to the orders of an authority, like in the Milgram experiment (Kohlberg 1984).

- *Development:* MCT research has refuted the cognitive-developmental postulate of invariant sequence of development: People's moral competence can regress if they do not have an opportunity to use it for a longer period of time (Lind 2000; Schillinger 2006; Lupu 2009; Saeidi 2011). Yet it supports the findings of Kohlberg and his associates that moral competence can be effectively fostered through certain methods of dilemma discussions (Lind 2002).

- *Education:* Finally, the MCT lets us objectively and economically measure the efficacy and efficiency of methods and programs of moral education, like the *Konstanz Method of Dilemma Discussion* (KMDD) (Lind 2002, 2019; Hemmerling 2014).

Conclusion

Experimentally designed tests let us make psychological traits visible, validly and objectively. Experimentally designed tests accomplish what subjective psychologists always wanted to achieve, and what objective psychologists could not deliver: They make it possible to measure the properties of humans' internal traits through direct observation of their behavior. Moreover, experimentally designed tests allow us to examine the

truth of the assumptions on which they are based. Experimentally designed tests can be used not only in moral psychology but in any field of psychology in which testable theories and clear definitions of their objects are available.

A final caveat: experimental questionnaires should be used only for research and evaluation of methods and programs, not for evaluating people. That is, the MCT must not be used for high-stakes testing of students, teachers, or named institutions. There is no evidence that test-based sanctions lead to better learning and better behavior. Moreover, sanctions undermine the validity of psychological tests and, therewith, impede their usefulness for improving therapy and education (Amrein & Berliner 2002; Ravitch 2013; Koretz 2017). When used for high-stakes testing, tests wear out within a few years and must be substituted by new content. Thus their findings can be compared only through daring statistical constructions (Linn 2010). In contrast, the MCT celebrates its 44th anniversary and still has not had to be changed (beside a few minor editorial corrections). Thus it has provided us with a great wealth of data that stretch over a long period of time and across many cultures. This in turn allows us to test many hypotheses on the nature, relevance, and teachability of moral competence.

Acknowledgements

I would like to thank the following persons for encouraging me to write this article and for commenting on at least parts of earlier drafts of it: Małgorzata Steć, Hans Brügelmann, Jim Fearns, Wilhelm Kempf, Martina Reinicke, and Aiden Sisler.

References

Allerup P. 2007. "Identification of Group Differences Using PISA Scales – Considering Effects of Inhomogeneous Items," in S. T. Hopmann, G. Brinek, & M. Retzl (Eds.), *PISA zufolge PISA* (pp. 175–202). Berlin: LIT-Verlag.

Amrein A. L. & Berliner D. 2002. "High-stakes Testing, Uncertainty, and Student Learning," *Education Policy Analysis Archives* 10 (18):1–74.

Ausch R. 2016. *Methodological Problems with the Academic Sources of Popular Psychology*. Lanham: Lexington Books.

Bailey K. 1994. *Method of Social Research*. New York: The Free Press.

Block J. 1967. "Advancing the Psychology of Personality," in D. Magnusson & N. Endler (Eds.), *Personality at the Crossroads* (pp. 37–63). Hillsdale, NJ: Lawrence Erlbaum.

Bühler K. 1927. *Die Krise der Psychologie* [The crisis of psychology]. Jena: Gustav Fischer.

Cohen R. J. & Swerdlik M. E. 2018. *Psychological Testing and Assessment. An Introduction to Tests and Measurement.* 9th edition. New York: McGraw Hill.

Colby A., Kohlberg L., Colby A., Speicher B., Hewer A., Candee O., Gibbs J., & Power C. 1987. *The Measurement of Moral Judgment: Standard Issue Scoring Manual*, Vol. I & II. New York: Cambridge University Press

Donders F. C. 1969 (Orig. 1868). "On the Speed of Mental Processes," *Acta Psychologica. Attention and Performance II*, 30:412–431. Retrieved from http://www2.psychology.uiowa.edu/faculty/mordkoff/InfoProc/pdfs/Donders 01868.pdf (on February 25, 2021). DOI: 10.1016/0001-6918(69)90065-1

Emler N., Renwick S., & Malone B. 1983. "The Relationship between Moral Reasoning and Political Orientation," *Journal of Personality and Social Psychology* 45:1073–80.

Graumann C. F. 1960. "Eigenschaften als Problem der Persönlichkeitsforschung," in P. Lersch & H. Thomae (Eds.), *Handbuch der Psychologie, Bd. IV, Persönlichkeitsforschung und Persönlichkeitstheorie* (pp. 87–154). Göttingen: Hogrefe.

Gregory R. J. 2018. *Psychological Testing: History, Principles, and Applications*. London, England: Pearson Publisher.

Gulliksen H. 1950. *Theory of Mental Tests*. New York: Wiley.

Habermas J. 1990. *Moral Consciousness and Communicative Action*. Cambridge: MIT Press.

Haidt J. 2007. "The New Synthesis in Moral Psychology," *Science* 316:998–1002.

Hartshorne H. & May M. A. 1928. *Studies in the Nature of Character. Vol. I: Studies in Deceit, Book I & II*. New York: Macmillan.

Hemmerling K. 2014. *Morality behind Bars – An Intervention Study on Fostering Moral Competence of Prisoners As a New Approach to Social Rehabilitation*. Frankfurt/Main: Peter Lang.

Jablonka E. 2007. "Mathematical Literacy: Die Verflüchtigung eines ambitionierten Testkonstrukts in bedeutungslosen PISA-Punkten," in T. Jahnke & W. Meyerhöfer (Eds.), *Pisa & Co. Kritik eines Programms* (pp. 247–280). Hildesheim: Franzbecker.

Keasey Ch. B. 1974. "The Influence of Opinion-agreement and Qualitative Supportive Reasoning in the Evaluation of Moral Judgments," *Journal of Personality and Social Psychology* 30:477–482.

Keller M. 1990. "Zur Entwicklung moralischer Reflexion: Eine Kritik und Rekonzeptualisierung der Stufen des präkonventionellen moralischen Urteils in der Theorie von Kohlberg," in M. Knopf & W. Schneider (Eds.), *Entwicklung. Allgemeine Verläufe – Individuelle Unterschiede* (pp. 19–44). Göttingen: Verlag für Psychologie.

Kempf W. 1981. *Testtheorie.* University of Konstanz, Dept. of Economy and Statistics, a mimeographed paper.

Kohlberg L. 1958. *The Development of Modes of Moral Thinking and Choice in the Years 10 to 16*. University of Chicago. Unpublished doctoral dissertation.

Kohlberg L. 1964. "Development of Moral Character and Moral Ideology," in M. L. Hoffman & L. W. Hoffman (Eds.), *Review of Child Development Research*, Vol. I (pp. 381–431). New York: Russel Sage Foundation.

Kohlberg L. 1976. "Moral Stages and Moralization: The Cognitive-developmental Approach," in T. Lickona (Ed.), *Moral Development and Behavior: Theory, Research and Social Issues* (pp. 31–53). New York: Holt, Rinehart & Winston.

Kohlberg L. 1979. *The Meaning and Measurement of Moral Development. The Heinz Werner Lecture Series,* Vol. 13. Worcester, MA: Clark University Press.

Kohlberg L. 1984. *The Psychology of Moral Development.* Vol. II*: Essays on Moral Development*. San Francisco: Harper & Row.

Koretz D. 2017. *The Testing Charade. Pretending to Make Schools Better*. Chicago: University of Chicago Press.

Kurtines W. M. & Greif E. B. 1974. „The Development of Moral Thought: Review and Evaluation of Kohlberg's Approach," *Psychological Bulletin* 81:453–470.

Levy-Suhl M. 1912. „Die Prüfung der sittlichen Reife jugendlicher Angeklagter und die Reformvorschläge zum § 56 des deutschen Strafgesetzbuches," *Zeitschrift für Psychotherapie und Medizinische Psychologie* 4:232–254.

Lind G. 1978. "Wie misst man moralisches Urteil? Probleme und alternative Möglichkeiten der Messung eines komplexen Konstrukts," in G. Portele (Ed.), *Sozialisation und Moral* (pp. 171–201). Weinheim: Beltz.

Lind G. 1982. "Experimental Questionnaires: A New Approach to Personality Research," in A. Kossakowski & K. Obuchowski (Eds.), *Progress in Psychology of Personality* (pp. 132–144). Amsterdam, NL: North-Holland.

Lind G. 1985a. "Attitude Change or Cognitive-moral Development? How to Conceive of Socialization at the University," in G. Lind, H. A. Hartmann, & R. Wakenhut (Eds.), *Moral Development and the Social Environment. Studies in the Psychology and Philosophy of Moral Judgment and Education* (pp. 173–192). Chicago: Precedent Publishing.

Lind G. 1985b. "The Theory of Moral-cognitive Development: A Socio-psychological Assessment," in G. Lind, H. A. Hartmann, & R. Wakenhut (Eds.), *Moral Development and the Social Environment* (pp. 21–53). Chicago: Precedent Publishing Inc.

Lind G. 1986. "Cultural Differences in Moral Judgment Competence? A Study of West and East European University Students," *Behavior Science Research 20:*208–225.

Lind G. 1989. "Essay Review: 'The Measurement of Moral Judgment' by Anne Colby, Lawrence Kohlberg *et al*.," *Human Development 32:*388–397.

Lind G. 2000. "Moral Regression in Medical Students and Their Learning Environment," *Revista Brasileira de Educacao Médica* 24(3):24–33.

Lind G. 2002. *Ist Moral lehrbar? Ergebnisse der modernen moralpsychologischen Forschung*. 2nd Edition. Berlin: Logos-Verlag.

Lind G. 2005. *The Cross-cultural Validity of the Moral Judgment Test:* Findings from 28 Cross-cultural Studies*. Presentation at the conference of the American Psychological Association, August 18–21, 2005, Washington, D.C.

Lind G. 2013. "Thirty Years of the Moral Judgment Test:* Support for the Dual-aspect Theory of Moral Development," in C. S. Hutz & L. K. de Souza (Eds.), *Estudos e pesquisas em psicologia do desenvolvimento e da personalidade: uma homenagem a Angela Biaggio* (pp. 143–170). São Paulo: Casa do Psicólogo.

Lind G. 2017a. "From Donder's Dilemma to Objective Internal Assessment: How Experimental Developmental Psychology Can Contribute to Moral Education," *Psychologia Rozwojowa* 22(3):15–24.

Lind G. 2017b. *Wie kann man moralische Orientierung im Verhalten erkennen? Von Brunswiks diakritischer Methode zum Moralische Kompetenz-Test (MKT)*. Beitrag zur Moralforschertagung, January 26–28, 2017, Universität Leipzig.

Lind G. 2019. *How to Teach Morality.* Berlin: Logos.

Lind G. & Nowak E. 2015. "Kohlberg's Unnoticed Dilemma – The External Assessment of Internal Moral Competence?," in B. Zizek, D. Garz, & E. Nowak (Eds.), *Kohlberg Revisited* (pp. 139–154). Rotterdam – Taipei – Boston: Sense Publishers.

Lind G. & Wakenhut R. 2010. "Testing for Moral Judgment Competence," in G. Lind, H. A. Hartmann, & R. Wakenhut (Eds.), *Moral Judgment and Social Education* (pp. 79–105). Rutgers, NJ: Transaction Books, 2nd edition.

Linn R. L. 2000. "Assessments and Accountability," *Educational Researcher 29(*2):4–16.

Loevinger J. 1957. "Objective Tests as Instruments of Psychological Theory," *Psychological Reports* 3:635–694.

Lupu I. 2009. „Moral, Lernumwelt und Religiosität. Die Entwicklung moralischer Urteilsfähigkeit bei Studierenden in Rumänien in Abhängigkeit von Verantwortungsübernahme und Religiosität." Doctoral dissertation, University of Konstanz, Germany.

Meehl P. E. 1958. "When to Use Your Head Instead of the Formula?," in H. Feigl, M. Scriven, & G. Maxwell (Eds.), *Minnesota Studies in the Philosophy of Science* (pp. 498–506). Minneapolis, MI: University of Minnesota Press.

Micceri T. 198. "The Unicorn, the Normal Curve, and Other Improbable Creatures," *Psychological Bulletin* 105(1):156–166.

Piaget J. 1965 (Orig. 1932). The *Moral Judgment of the Child*. New York: The Free Press.

Piaget J. 1976. "The Affective Unconscious and the Cognitive Unconscious," in B. Inhelder & H. H. Chipman (Eds.), *Piaget and His School* (pp. 63–71). New York: Springer.

Popper K. R. 1968. *Conjectures and Refutations*. New York: Harper & Row.

Prehn K. 2013. "Moral Judgment Competence: A Re-evaluation of the Dual-Aspect Theory

Based on Recent Neuroscientific Research," in E. Nowak, D. Schrader, & B. Zizek (Eds.), *Educating Competencies for Democracy* (pp. 9–22). Frankfurt am Main – Bern – Brussells – New York: Peter Lang Editions.

Ravitch D. 2013. *The Death and Life of the Great American School System: How Testing and Choice Are Undermining Education*. New York: Basic Book.

Rest J. R. 1969. "Level of Moral Development As a Determinant of Preference and Comprehension of Moral Judgments Made by Others," *Journal of Personality* 37(1):220–228.

Rest J. R. 1979. *Development in Judging Moral Issues.* Minneapolis, MI: University of Minnesota Press.

Rosnow R. L. & Rosenthal R. 1997. *People Studying People. Artifacts and Effects on Behavioral Research*. New York: Freeman & Co.

Ryle G. 1949. *The Concept of Mind*, Chicago: University of Chicago Press.

Scott W. A. 1968. "Attitude Measurement," in G. Lindsey & E. Aronson (Eds.), *Handbook of Social Psychology*. Vol. II (pp. 204–273). Reading, MA: Addison-Wesley.

Saeidi-Parvaneh S. 2011. *Moral, Bildung und Religion im Iran – Zur Bedeutung universitärer Bildung für die Entwicklung moralischer Urteils- und Diskursfähigkeit in einem religiös geprägten Land.* Doctoral Dissertation, University of Konstanz.

Schillinger M. 2006. *Learning Environments and Moral Development: How University Education Fosters Moral Judgment Competence in Brazil and two German-speaking Countries*. Aachen: Shaker Verlag.

Schoenfeld A. H. 1999. "Looking Toward the 21th Century: Challenges of Educational Theory and Practice," *Educational Researcher* 28:4–14.

Sjoberg S. 2007. "PISA and 'Real Life Challenge': Mission Impossible?," in S. T. Hopfmann *et al.* (Eds.), *PISA zufolge PISA* (pp. 201–224). Berlin: LIT-Verlag.

Sjoberg S. 2017. "PISA Testing – A Global Educational Race?," *Europhysics News.* https://www.europhysicsnews.org/articles/epn/pdf/2017/04/epn2017484p17.pdf

Travers R. M. 1951. "Rational Hypotheses in the Construction of Tests," *Educational and Psychological Measurement* 11:128–137.

Walberg H. J., Strykowski B. F., Rovai E., & Hung S. 1984. „Exceptional Performance," *Review of Educational Research* 54(1):87–112.

Walberg H. J. & Haertel G. D. 1994. „The Implications of Cognitive Psychology for Measuring Student Achievement," in OECD (Ed.), *Making Education Count* (pp. 219–236). Paris: OECD.

Watson J. B. 1970 (Orig. 1924). *Behaviorism*. New York: Norton.

Wilson M. 2005. *Constructing Measures. An Item Response Modeling Approach*. Mahwah, NJ: Erlbaum Associates Publishers.

Wittgenstein L. 1953. *Philosophical Investigations*. Trans. by G.E.M. Anscombe. New York: The Macmillan Company.

Wuttke J. 2005. “Fehler, Verzerrungen, Unsicherheiten in der PISA-Auswertung” [Errors, biases, uncertainties in PISA interpretation], in T. Jahnke & W. Meyerhöfer (Eds.), *Pisa & Co. Kritik eines Programms* (pp. 101–154). Hildesheim: Franzbecker.

Wuttke J. 2007. “Uncertainties and Bias in PISA,” in S. T. Hopmann, G. Brinek, & M. Retzl (Eds.), *PISA zufolge PISA* (pp. 241–264). Berlin: LIT Verlag.

(for more references on EQ and MCT see: http://moralcompetence.net)