Buchbestellung: Logos, Berlin

What is moral competence? Can it be measured? Can it be taught effectively? If so, how? This book explores these questions from three perspectives: experimental psychology, curriculum development, and instructor training. Part one discusses the research from which, like a jig-saw puzzle, a comprehensive picture of the nature, development, and teachability of morality emerges. The picture focuses on moral competence, the ability to solve problems and conflicts on the basis of moral principles through deliberation and discussion rather than violence and deceit. Part two explains how moral competence can be taught effectively with the Konstanz Method of Dilemma Discussion (also known as Discussion Theater), which has been used with great success to foster moral development in schools and universities, military installations, prisons, and retirement communities in many countries. The book describes the Method, gives vivid illustrations of its use, and provides psychologists, teachers, and professional trainers with resources and guidance in its application.

"The definitive, research-based book on morality teaching with highly useful applications to educational practice. Highly recommended."

Dr. Herbert Walberg, Emeritus Professor of Education and Psychology, University of Illinois at Chicago.

"We all want to be 'good' Lind contends—it's part of our human inheritance. But being morally competent, he shows, is enhanced and nourished when educators develop propulsive learning opportunities for students to practice and develop."

Dr. William Ayers, Distinguished Professor Emeritus of Education, University of Illinois at Chicago.

"Lind's mastery of the history and philosophy of morality and moral education is quite apparent. He writes of the complex issues bound up in morality in a beautifully clear and persuasive manner."

Dr. Richard M. Felder, Hoechst Celanese Professor Emeritus, North Carolina State University.

"Dr. Lind's experimental and educational approach to morality is unique worldwide."

Dr. Ewa Nowak, Professor of Philosophy and Ethics, Adam Mickiewicz University, Poznan, Poland.

"I really didn't think that one can discuss highly controversial issues in such a peaceful way. I learned a lot."

A forty year old participant of a KMDD/DT session.

Logos Verlag Berlin ISBN 978-3-8325-5005-9

Georg Lind How to Teach Moral Competence

Chapter 4:
How to make moral competence visible

How to Teach Moral Competence

GEORG LIND

NEW: Discussion Theater

λογος

Socrates: But if this be affirmed, then the desire of good is common to all, and one man is no better than another in that respect?

Meno: True

Socrates: And if one man is not better than another in desiring good, he must be better in the power of attaining it?

Meno: Exactly.

Socrates: Then according to your definition virtue would appear to be the power of attaining good.

Socrates (469 – 399 BC)
Quoted from Plato: *Meno*

"Now as virtue is an end, and is desirable on its own account, without fee and reward, merely for the immediate satisfaction which it conveys."

David Hume (1751)
The enquiry concerning the principles of morals

"I know no safe depository of the ultimate powers of the society but the people themselves; and if we think them not enlightened enough to exercise their control with a wholesome discretion, the remedy is not to take it from them, but to inform their discretion by education. This is the true corrective of abuses of constitutional power."

Thomas Jefferson (1820)
Letter to William C. Jarvis. ME 15:278

The author has written this book as a source of information and *not* as a replacement for actually taking part in the Discussion Theater / KMDD® training program described in this book.

He and the publisher have made every effort to ensure the accuracy of the information at the time of publication. They are not responsible for any adverse effects of the use or application of this information. In particular, we have designed dilemma discussions, which trigger moral emotions. Only through feeling these emotions participants can learn to deal effectively with them. Consequently, only well-trained and certified KMDD/DT teachers, who have learned how to deal with moral emotions, should carry out dilemma discussions. Misapplication could render the method ineffective or could even have unintended negative effects on participants.

KMDD is an internationally registered trademark. Only persons who possess a valid certificate as a *KMDD Trainee/Teacher/Trainer* are permitted to use it for advertisement.

This is a new version of my book "How to teach morality" (2016). I have slightly changed the title and made several editorial changes. New are parts of Chapter 4 ("Making moral competence visible") and the chapter on *Discussion Theater*.
The 2016 edition has been translated into Chinese and Korean. The first German edition (Lind 2003) is also available in Greek and Spanish.

The *Konstanzer Methode der Dilemma-Diskussion* (*KMDD*) ® is a registered trade mark in China, the European Union, Switzerland and Turkey. Registration in more countries is pending.

Table of Contents

Chapter 4

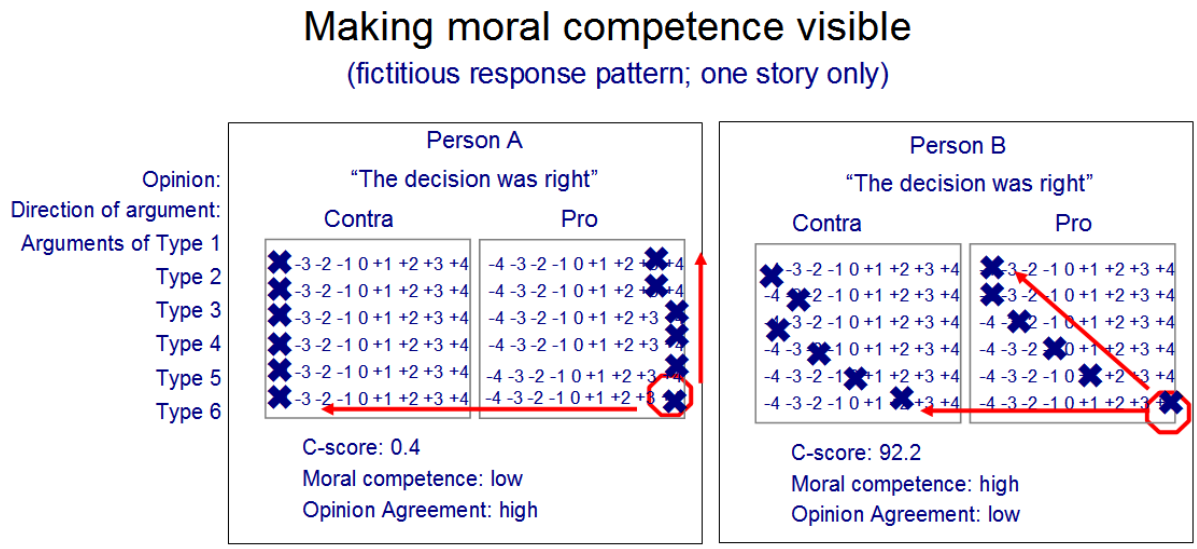
Making moral competence visible

> "The surest beacon for psychometric theory
> is the real involvement in the solution of
> psychological problems."
> (Loevinger 1957, p. 693)

In the previous chapters, I outlined the state of the art of moral competence theory from my point of view. How well does it agree with reality? Obviously, we cannot answer this question unless we can measure moral competence, that is, to make moral competence visible. Only then, we can submit our hypotheses to empirical and experimental tests.

4.1 From theory to measurement

If moral competence really exists, it should be visible in people's behavior as if our bones are visible in an X-ray shot. With the *Moral Competence Test* (MCT), it is indeed possible to display graphically the moral competence of an individual person. When you look at the graph below, you can directly see the two (fictitious) persons' levels of moral competence. You do not need any statistical knowledge for this. They had to rate pro- and contra-arguments which represented six different type of moral orientation on a scale from -4 (reject) to +4 (accept). The information contained in these graphs has been translated into a number, the so-called C-score, in order to be able to do further analyses. The C-score can range from 0 (no moral competence) to 100 (perfect moral competence as defined in this book). Person A on the left side got a very low C score, Person B on the right a very high one. Do you have any idea why? (If you are not sure, you can consult the footnote on this page.)[7]



[7] Person A is obviously no able to judge the arguments by their moral quality. He or she judges them only by their opinion-agreement. In contrast, Person B judges pro and contra arguments by their moral quality and hardly by their opinion-agreement.

Making moral competence visible was not easy. We had to solve five problems. The first problem was to find a clear definition of moral competence or to define it ourselves. Without a clear definition of what we want to measure, we cannot interpret our observations. The second problem was to find or create a moral task that would require moral competence. If we want to measure a competence, we always need an appropriate task. Without a difficult task, a test is not a test of competence but may be a test of moral attitude. The third problem was to define an adequate standard for moral competence. Without a standard, we cannot say what is high or low on our scale. *External* standards that we use with achievement tests cannot be used in a test, which is to measure an individual's ability to follow his or her *own* moral standards. The forth problem was to find a proper design for our test that could disentangle the different factors, which may determine a participant's responses. The final problem was to translate the visualization of a people's moral competence into an adequate number so that we can analyze our hypotheses.

Defining moral competence

Kohlberg's remarkable definition of *moral judgment competence* provided us with a good starting basis for defining moral competence. He defined 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, p. 425) This definition is remarkable in several ways. It is remarkable because it exists. Very few, if any, psychologists provide such clear definition of the object of study. It states that moral competence become visible only when the participants are confronted with the task to make decisions and judgments. It also implies that moral competence must be scored in regard to the participants' own moral standards ("based on internal principles"), not in regard to the researcher's norms. Finally, Kohlberg's definition makes clear that moral competence is neither an unobservable construct nor a latent trait, but is a real trait which should be visible in people's behavior if it exists ("and to act in accordance with such judgment").

I felt that we needed to modify and extend this definition. Kohlberg and Piaget used the term judgment often to describe overt, verbal moral reasoning while moral competence is an unconscious processes which people might not be aware of and do not reveal in their ethical reflections. In order to avoid confusion I dropped the term "judgment." I use now only the term *moral competence*.

We cannot study of morality adequately, if we confine it to individual thinking. We must also look at collective behavior and communication, as Habermas' (1990) has rightly argued. For him the highest moral principle is to solve conflicts through discourse instead using violence. Hence, I have included in my definition of moral competence also the ability to solve conflicts between opposing moral concerns by discussion with opponents.

Finding a moral task

A test of moral competence is only valid if it contains a difficult moral task. Kohlberg's definition does not specify which tasks would qualify as moral tasks. In his own clinical interview method, he confronts participants with the task to reason about the decisions of fictional protagonists in dilemma stories: Was their decision right or wrong? Why do you think so? In other words, the "task" for the interviewee is to judge another person's decision-making and to defend his or her judgment with arguments. Is this really a difficult task for the

interviewees? There is much empirical support of this claim. (Broughton 1978a, b; Colby et al. 1987; Lind 2002) However, there is also doubt. (Kurtines & Greif 1974; Haidt 2001) Apparently, respondents can fake their answers in the Kohlberg interview upward when instructed accordingly. (Haste 1985) This would not be possible if the task of the test was difficult. This would be the case if the task for the interviewees would be to cope with a real conflict instead of to reason only about the conflict of another person.

However, in many countries there are tight ethical limits for inducing real conflicts for research purposes. For example, it would be unethical if we would seduce research participants to cheat on a real achievement test or harm somebody else, or to deceive them about the real purpose of an experiment.

Two authors helped me to find a difficult task that would raise no ethical concerns, Habermas and Keasey. Habermas' (1990) theory of communicative ethics suggests that discussing hot issues could be such a task. His discourse principle requires us to solve conflicts through a moral discourse instead of through violence. This requires the participants of the discourse to defend their views with moral arguments and also to understand and accept the moral arguments of their opponents. Keasey (1974) found in his experimental study that many people are not able to make a differentiated judgment when dealing with counter-arguments. Everyday experience confirms this finding. When confronted with arguments in a debate, many people judge them merely because of their opinion-agreement regardless of their moral quality. They welcome any argument, which supports their own side, even if it disagrees with their own moral orientations, but reject any argument opposing their sided, even if it agrees with their moral orientation. Research shows that this applies to most people. For many people it is a difficult or even insoluble difficult task to rate arguments in regard to their moral quality instead of to their opinion-agreement.

Therefore, the task to judge arguments of different moral quality and different opinion-agreement seemed to be a perfect task to measure the core of people's moral competence, namely their ability to *judge arguments in regard to their moral quality instead of in regard to their opinion-agreement (or other non-moral criteria).* After more than forty years of using the MCI, we can say, indeed, it is.

Selecting a standard for scoring moral competence

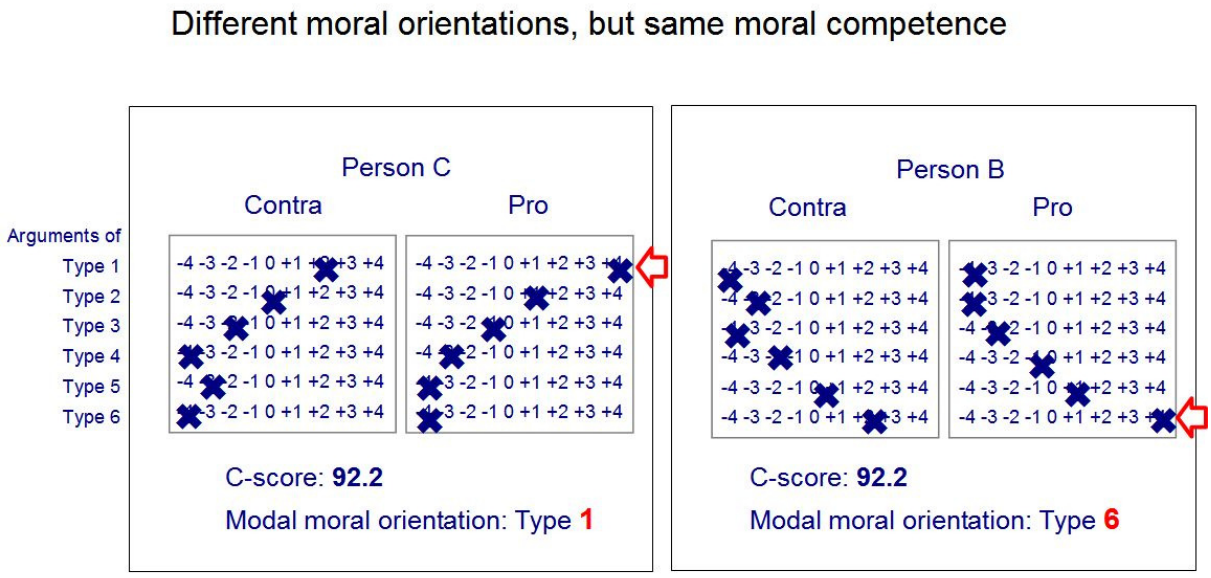
By which standards should we score this ability? This question was tricky. On the one side we cannot, like we do in achievement tests, score the participants' responses according to some *external* criterion. If we agree that behaving morally means to behave according to one's inner moral principles or standards, we must score people's responses in regard to their *own* moral orientations. They should not get low scores because they prefer to discuss a certain dilemma on a different level of moral discourse than the researcher expects them to do.

In spite of his internal definition of moral competence, Kohlberg used an external standard for scoring the answers of the participants of the Moral Judgment Interview. (Lind & Nowak 2015) He postulated that his Six Stage model of moral reasoning represents an objective standard of moral adequacy of moral reasoning about moral dilemmas: "I include in my approach a normative component. [...] This normative concern has led me to rely upon philosophic as well as psychological theory in defining what I study. 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; p. 400)

Kohlberg and his team assigned the highest moral competence score only to interviewees

whose reasoning exhibited a Type-Six moral orientation. Yet not all people may share his external standard, even if they have high moral competence. Not each solution of a conflict requires Type-Six moral orientations. Many problems and conflicts can be adequately solved invoking lower Types of reasoning. Invoking higher Types could even be inadequate. For example, if I have a conflict between drinking wine or water, Type 2 reasoning seems an adequate level of moral deliberation: I prefer to drink what I like better. Other people may see it differently, for example people whose religion forbids drinking alcohol. They may rightly argue that we should reflect and discuss this decision on a higher level. In both cases, the preference for a certain Type of moral orientation does not give us any cue about the protagonists' degree of moral competence. If we would require the participants to prefer certain Types of moral orientations, we would not measure their moral competence but their conformity to an external norm.

Therefore, we decided to score the responses of our participants without reference to the moral orientation, which they prefer. This means that we assign the highest competence score (of 100) to a person even when he or she prefers Type 1 moral reasoning to all other types. (See graph below.)



Note: The "Types" correspond to the six Kohlbergian Stage-Orientations. The C-score can range from 0 to 100.

In contrast to Kohlberg, I also decided to measure the two aspects distinctly. Only this way we could study their empirical relationship. In fact, studies found a very high correlation between the two aspects of moral judgment: the higher a person's moral competence the more clearly he or she prefers high type moral reasoning, and rejects low type moral reasoning. (Lind 1986b; see also the section bellows about the MCT's empirical validity.) This high correlation has been confirmed by all studies without an exception. This finding supports Piaget's (1976) theory of affective-cognitive parallelism. It also supports indirectly the usefulness of the data produced with the Kohlbergian interview method and moral preference tests, which contain no moral task and use external standards for scoring. That is, even though these methods are not valid, we can use their data as if they were valid. However, we should be cautious. The correlation between moral competence and moral orientation is not perfect and may break down under some circumstances.

You may object and say there is still some external norm involved when we compare people's moral competence. This is true. However, I feel more justified to use the pure moral competence score of the MCT for comparing and judging people's moral competence than to use

certain moral principles as standards. We can suppose that everyone agrees that high moral competence is better than low moral competence and its consequences: violence, deceit and submission under authorities. If this is true, internal standards and external standards become identical.

Making moral competence visible through test design

The graph above shows why we need to look at the *whole pattern* of an individual's responses to a carefully crafted, *complex design of stimuli* in order to assess a human trait. We cannot infer it just by looking at isolated responses. They are ambiguous. If respondents rate an argument of Type 6 very high this rating is ambiguous. It can mean that this individual appreciates the moral quality of Type-Six arguments, or that this individual likes it because of its opinion-agreement. Only when we see this single response in the context of a person's total pattern of response, we can be sure what it really means. (See the red-circled mark in the above graph.)

Already Scott (1968) had shown how the ambiguity of isolated responses calls the validity of any psychological test into question, even simple attitude tests. This ambiguity of isolated acts also explains why Hartshorne and May (1928) failed to assess moral character through observing isolated acts like cheating on tests. They gave their research participants hard academic tests. They left them unattended but observed them through one-way mirrors: Would they cheat or would they stay honest? Based on their observations they assigned an honesty scores to each participant: The less people cheated the stronger, the researchers believed, were their moral character. There assessment method was based on the conviction that a trait like moral character can be assessed only by observing an isolated act (cheating or not cheating), without considering its context and its motivation: "Moral behavior must be observed, and measured, without any reference ... to its motives or its rightness or wrongness. The first question to ask is what did the subject do? Unless this question is answered in quantitative terms so that what he did is clearly known, there is little use in going on to ask why he did it, and still less use in speculating whether he is to be blamed or praised." (Hartshorne & May 1928, p. 11) At the end of their book, they insightfully admitted: "The essence of the act is its pretense. Hence, it 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." (p. 377)

Already in 1955, Egon Brunswik offered a brilliant solution for this problem that has gone unnoticed by most researchers. He showed that we could not understand the intentions of an act if we look at isolated acts but only if we design our observation as if we design an experiment, using the factors of interest as so-called design factors. His proposal also implies repeated measurements, but not in order to estimate the preciseness of observation of isolated acts, as researcher usually do in psychological and educational measurement. We rather need them in order to disentangle the factors that may determine these acts. In everyday life, we apply Brunswik's idea all the time when we try to understand the meaning of other people's behavior and of their own. (Kelley 1973) Ironically, his idea has not yet caught on in educational and psychological research.

Brunswik's idea thrilled me. It let me design my tool for measuring moral competence, the *Moral Competence Test*, as a multivariate experiment or an *Experimental Questionnaire*. (Lind 1978; 1982) Such multivariate experiments are long in use and have proven to be a powerful instrument for testing hypotheses about causal processes. Researchers have hardly used them yet for making traits, like moral competence, visible.

4.2 The *Moral Competence Test* (MCT)

Except for its experimental design, the MCT is simple and short. It fits on two pages. The MCT presents two dilemma stories to the participants. While the protagonists are hypothetical, the issues of the stories are real. One is about a doctor who is asked by his lethally ill patient to help her die, and one about two workers who break into their manager's office to secure evidence for their illegal eavesdropping on the workers. The participants are asked to mark their opinion on the protagonists' decision in each story on a 6-point scale from right (+3) to wrong (-3). By marking their opinion on the protagonist's decision, they publicly take sides on an issue. This sets the stage for the actual experiment.

After each story the participants are confronted with six arguments for and six against the protagonist's decision—and, therewith, also for and against their opinion on this decision. The participants are asked whether they accept or reject these arguments. They must mark their answers on a scale from -4 (completely reject) to +4 (completely agree).

Each argument voices a certain moral orientation, namely one of the six Types of moral orientation, which Kohlberg used to define his six Stages of moral development. Note, however, that the arguments do not represent different *Stages* but only different *Types* of moral reasoning. Since we use two dilemma stories, two groups of arguments, six arguments in each group, there are altogether 24 arguments which the participants is to judge.

We took the *Doctor* story from Kohlberg's *Moral Judgment Interview* (MJI). Therefore, we could take many arguments from the sample interview samples, which the authors of the MCI scoring manual have provided. (Colby et al. 1987) The arguments for the *Workers* story had to be created new. Contrary to our expectations, we it was more difficult to formulate arguments of a low type than at a higher level. For example, it took great effort to formulate serious Type-One and Type-Two arguments for and against mercy killing.

Experimental Questionnaire

Here come some technical explanations that you can skip if you like. Readers familiar with experimental psychology will have discovered that the arguments of the MCT are, as Brunswik proposed, a behavioral experiment with a 2 x 2 x 6 factorial multivariate design. The three factors are (I) two different decision contexts (workers, doctor), (II) two opinions on each decision (pro and contra), and (III) the six Types of moral orientation which are voiced in each argument respectively. The so-called "dependent variable" is defined as the participant's rating of the arguments on the scale from -4 (totally reject) to +4 (totally accept).

The crux of the experimental design of the MCT is that the three factors are "orthogonal," that is, they are contained in *each* argument, rather than being operationalized each in a separate test. This is necessary because they represent inseparable *aspects* of behavior, not separable *components*. In other words, each aspect is present in each argument and is systematically varied and combined. This "orthogonal" design of the MCT makes it possible to "read" the participants' minds from the pattern of the individual responses as shown in the above graphs. If participants' accept all supporting arguments, and reject all opposing arguments regardless of the arguments' moral quality, we say that their pattern exhibits a lack of moral competence. Their pattern of ratings shows that they do not use their own moral principles to judge these arguments and do not examine, and possibly revise, their opinion if their opinion disagrees with their own moral principles. Rather they utilize arguments only to cement their stance on a certain issue. Conversely, if participants rate arguments in regard to their moral quality, we say that their pattern exhibits high moral competence. They seem to be able to apply their own

moral principles to the evaluation of arguments regardless their opinion-agreement. This does not mean that they have no opinion on the issues that are discussed. However, usually there opinion is less extreme and more debatable.

Application of the Moral Competence Test

The MCT requires only basic reading skills. We can use it with people over the age of about ten years. If the test-takers do not understand some words, the supervisors may help as long as they do not influence the participants. Participants usually like to fill it out. They say that the test is very interesting. They rarely skip questions. It seems that this happens mostly by accident. In the electronic version, we remind the participants to check their answers for completeness before they sent them off.

The MCT can be completed as a group test (e.g. in class at school), as a single test filled-out by individuals, sent by postal mail, or filled out online. To make sure that the participants fill it out carefully, some provisions should be made. Participants are more motivated to fill out the test carefully if they feel that they are participating in an important study, for example when they are asked to help to evaluate the effectiveness of a moral education program. Researchers should always report the return rate in their publications because it can have an impact on the average C-score. A high dropout rate correlates with low moral competence. (Krebs & Rosenwald 1977) So if many participants dropout, the C-score of the study group will increase. Filling in the standard version with two dilemmas lasts between eight and fifteen minutes depending on the reading speed of the participants. No time limits should be set for the MCT. When a participant uses too much time for deliberating on the best answer, the test-administrator can gently encourage a faster completion.

> Users of the MCT frequently ask me whether one can shorten the MCT by dropping one of the dilemma-stories, or substitute the stories by other stories, which have more "face-validity" for a particular research question or for a particular population. My answer has been always no. If we change the standard MCT, we cannot compare the findings anymore. It is a common mistake to believe that the measurement of a psychological concept should or can be adapted to a particular research question or population. Imagine if we would change our yardstick to align it with our objects. No comparisons could be made, and no scientific knowledge gained. This does not preclude critical reflections on the suitability of the standard MCT for one's research intentions and one's research subjects, or the need for new dilemma stories. I have been involved in some attempts to construct a new story. They all failed. We could not construct a new test, which was nearly as valid as the one we have. Only one attempt has succeeded so far, namely the attempt by Patricia Bataglia. (Bataglia et al. 2007; Bataglia 2009) This new test is best used as an addition in research studies, not as a substitute for the standard MCT. It contains the Judge Steinberg story, which teachers often use in KMDD sessions. Hence, it should not be used for measuring the efficacy of a KMDD session when this story is also being discussed in the session.

Calculation of the moral competence score (C-score)

The C-score is typically calculated for the *standard* MCT, which contains the two dilemma stories described above. Most researchers use only this C-score. Thus, we can use only it for comparing findings from different studies. Researchers who use other scores should call them differently in order to avoid confusion.

We can calculate the C-score also for each dilemma story separately, for example for studying the phenomenon of *moral segmentation*. Moral segmentation means that a person shows a much lower moral competence in one dilemma story than in the other one. Conventionally we speak of moral segmentation when the difference between the two C-score is bigger than 8 points. Moral segmentation is usually observed in the Doctor story, which deals with mercy killing. This seems to be connected to the religious background of the participants like Catholicism and Islam, which forbid mercy killing. It is noteworthy that religion does not only influence the participants' stance on mercy killing but also their readiness (or ability?) to think about it. (Lind 20000d; Bataglia & Schillinger 2013; Hegazi & Wilson 2013) Segmentation has also been observed in military contexts when military dilemma stories were used. (Wakenhut 1982; Senger 2010) Note: We should not compare the C-score for each dilemma story directly with the overall C-score for the MCT. It also ranges from 0 to 100, but for mathematical reasons it is always larger than the standard C-score.

Interpretation of the C-score

For research purposes the moral competence which we can see in the graphics need to be translated into a number. This number cannot summarize all information contained in the graphical display of the judgment pattern of an individual. However, it should contain the essential information, namely the degree of moral competence. We have designed the C-score to do this. In short, it quantifies the proportion of the variation of the individual responses, which can be accounted for by an individual's ability to judge the arguments in regard to their moral quality. A score of "0" means no proportion at all, "100" that hundred percent of an individuals response variance is "determined" by his or her moral competence. About how this number is calculated (by a multivariate analysis of variance) informs the MCT web site.

All studies using the MCT should report the average C-scores of the participants, not only broad categories or "transformed" data. Such practice makes studies less comparable, if not incomparable. I suggest that studies report the numbers with one digit behind the decimal point. Usually we do not need more digits. If we use the C-score in graphs, we should let the axis, on which it is shown, start from zero and end at forty. Average C-scores are rarely higher. If so, the axis must be longer of course. However, the axis should not be shorter if the average C-scores are lower. This would not be wrong but it tends to create a wrong impression in the reader. Especially when the differences found in various groups of the study are very small, shortening the axis would make the difference look much bigger than they actually are. When I find such reports, I always redraw the graphs in order to get a truer picture of the results.

Most researchers are only interested in the differences they find between the groups they have studied but rarely in the absolute level of moral competence, which their data shows. Yet also, the absolute level is interesting. It is interesting for the reader who wants to compare it with other studies on this topic. In addition, it could be interesting for the researcher as a reference for interpreting his or her findings. For example, if the overall level of moral competence is very low or the variation of scores in the study sample is very small, no big differences can show. The interpretation of a C-score as "low" or "high" should also take into account where the gross of C-scores is located on the absolute scale. A C-score of 15 may look high if most research participants have much lower C-scores. However, on a scale from 0 to 100 it is a rather low score!

On the basis of experimental studies about the importance of moral competence for various forms of moral or social behavior (obeying the law, keeping promises, helping people in

distress, rejecting violence as a means of political action etc.) it seems that a minimum level of moral competence is needed to live a moral life without being controlled by an authority see Chapter 5). In the next chapter, I will discuss the importance of moral competence for individuals' socio-moral behavior and for living together in a free society. These studies suggest that we should set this minimum at a C-score of 20. Above this competence level, our behavior seems to need less external control. Therefore, I pledge that we should invest into the promotion of moral competence of everyone so that everyone can develop beyond his or her moral competence beyond this level. I believe that this is possible and that we can afford this.

The interpretation of the size of the C-score and the sizes of differences is difficult and beyond the scope of this book. I will discuss how much effect an educational intervention should have in part 2. Beyond this, we are not able to set differentiated standards for interpreting the C-score. Some considerations are in place. This interpretation must consider many factors. Beyond a certain absolute level an increase of the C-score may not automatically mean an increase of honesty, law-abiding, or helping behavior. When people could develop a very high moral competence and can weigh in many moral principles and many facts when they make a decision, they may sometimes feel forced to lie, to transgress a law or to deny help. If we want that people develop the ability to make their own judgments, that is judgments based on internal, moral principles we should welcome this.

Usually situational factors should not be able to push the C-score upward since the C-score cannot be faked upward. (Lind 2002) Yet these factors may lower the actual C-score that a person could get. Certain forms of "priming" can lower the C-score. If the researcher lets the participants fill out an achievement test before the Moral Competence Test, they might think that they are expected to give "correct" answers to the test. Maybe priming accounts for the extremely low moral competence scores of fact of gifted students in a study in which an IQ test and the MCT were administered. (Hettinger 2009) The mechanisms of this priming are not fully understood yet but I assume that achievement-primed participants feel that they should take sides on the protagonists' decision in the MCT, and thus judge the arguments more in regard to their opinion-agreement. Time pressure can also lower the C-score, probably because the participants have not enough time to think. Even if the researcher does not prescribe a certain time limit, other situational factors may put the participant under time pressure, like the need to catch the bus at the end of a class. It seems that the application of the MCT by paid interviewers at the doorsteps also lowers the overall C-score because they might pushes the interviewee to hurry. Interpreting difference of C-scores needs also some consideration. If we can be sure that situational factors influence all participants' score in the same way, we can neglect them. They might react to such factors to different degrees. Nevertheless, mostly such comparisons seem to be legitimate and give valid results. Yet if the situational factors vary much between participants we must be careful not too overestimate differences. If participants do not feel well or are tired, or for many other reasons they might get a lower C-score than they would if the test-taking situation was normal.

For this and other reasons, we should not interpret individuals' C-score. We did not make the MCT for this and we do not know how situational factors influence individual test scores. Aside from this technical concern, we must not judge or select people based on their C-score for moral and practical reasons. Moral reason: Judging people because of their moral competence invites prejudices and distracts from the need to provide them with better opportunities for moral competence development. Practical reason: If we use the MCT for judging and selecting people, they will try to cheat the test as an act of self-defense. Through this, the MCT would become unusable within a short time, like achievement tests, which need

to be revised with a few years interval. Our policy of confining the use of the MCT to research and efficacy studies has contributed to its longevity. There is hardly any other test, which has been in use nearly unchanged for more than forty years.

Classifying the MCT

We have based the MCT on a new theory of measurement. It is an *Experiment Questionnaire*. (Lind 1982) This newness invites misrepresentations. It is not a "recognition test" although the participants have to "recognize" the questions and arguments which are to be rated. The MCT measures participants' moral competence, of which they are mostly not aware. The C-score does not stand for consistency score but for competence score. The term 'consistency' is meaningful only if we would define it in reference to a specific criterion. There is no "consistency per se" (there are many kinds of consistency). Therefore, we could say that the C-score reflects the consistency of responses *concerning the moral quality of the arguments*. However, it does not reflect consistency *concerning their opinion-agreement*. The C-score is a pure measure of moral competence. It does not reflect moral attitudes, values, orientations or preferences. However, the MCT allows also to measure moral orientations independently from the C-score.

Six scales of moral orientations

Besides the C-score, the MCT provides scores for each of the six Types of moral orientations built into the test. We can measure each Type by averaging a participant's ratings of the four arguments that represent it. (Remember that there are two arguments for each of the two dilemma stories.) Averaging means that we add the ratings together and then divide it by four. The scales for each Type of moral orientation range from -4 to +4 like the response scales on which they are based. However, note that in some publications, researchers do not divide the sums of the scales by four, so that the measured values range from -16 to +16. We usually represent the scores for the six Types of moral orientation as profile (for illustration see graph in Section 3.2, above).

I do not recommend older ways of indexing the affective aspect of moral behavior anymore. For a while we reported a P-score (preference score) for the MCT like the one which is used for the Defining Issues Test. (Rest 1979) It was used, as far I know, only in two studies. (Heidbrink 2010; Biggs & Colesante 2015) Yet we phased it out because the P-score lets us only distinguish between the preference for a "postconventional" Type of moral orientations and a non-conventional Type. The six Types of moral orientation scales give a more differentiated picture. Moreover, they do not confound affect and cognitive aspects.

The validity of the Moral Competence Test

The MJC is both theoretically and empirically valid. Actually the MCT is one of the few psychological tests which can be tested on their theoretically and empirically validity. We can do such tests only (a) if we clearly define the object of measurement in psychological terms and (b) if we base it on an elaborate psychological theory. Only in that case we can derive validity criteria for a test or measurement.

Although all researchers in the field of psychological and educational measurement regard the validity of a test as most important, they do not report the validity or report substitutes, which

indicate the reliability but not the validity of a test. They do not define their measurement objects psychologically but only statistically. For tests that are based only on statistical models, we can merely check their measurement error or "reliability" (retest correlation, internal consistency and standard error) but not their validity. However, precision of measurement is not the same as validity. A measurement can be very precise but not measure what it should measure. Unless we base a psychological test on a substantial psychological theory, we cannot know whether it is valid. I do not discuss here any of the dozens of types of validity which are used in research literature. Ironically, these types of validity have a validity problem themselves. All are based only on statistical models, instead of on substantial theories that have been well researched. Their multitude shows that they are not based on cumulative research but on the motto, "anything goes."

As I have shown in the previous chapters of this book, in contrast to most other tests, we have based the Moral Competence Test on a clear definition of its object and on an elaborate and well-founded psychological theory. Therefore, we can examine its theoretical and empirical validity.

Theoretical validity

The MCT is theoretically valid by virtue of design and construction. Its items (arguments) have been written to conform as closely as possible to Kohlberg's well developed six Types of moral orientation. Several distinguished experts of Kohlberg's theory reviewed them. Based on their reviews we revised the arguments. Of course, we did not submit the MCT to statistical item selection like conventional tests. If we would take out items from the MCI, which do not fit a certain statistical model, we may increase some statistical coefficients, but at the same time undermine its theoretical and empirical validity.

When we designed the MCT, we followed closely the theory on which we based it. As we have seen, it contains a moral task, which is rooted in experimental psychology and moral philosophy. The test-design of the MCT follows closely Brunswik's proposal to use a multivariate design. Its moral task is based on Habermas' imperative of communicative ethics (Solve conflicts through discourse instead of through the use of violence!), on Keasey's findings regarding role of opinion-agreement, and last but not least on Kohlberg's structural concept of moral competence. Consequently, Kohlberg acknowledged the MCT's capacity to "assign a pure structure score for an individual […] I believe," he wrote, "this to be a highly promising approach." (Kohlberg 2010, p. xvi, originally published in 1985) Interestingly, he taught the MCT in his seminars as I could witness during a class to which he invited me.

Empirical validity

The MCT is also *empirically valid*. The empirical validity of the MCT can be checked by five non-trivial hypotheses and because the MCT is in line with all although these hypotheses and its predictions are highly improbable. These hypotheses do not concern external criteria like data from another test. If valid tests were available, we would not have to construct a new test. These hypotheses have rather been derived from the theory of moral competence and well confirmed by research.

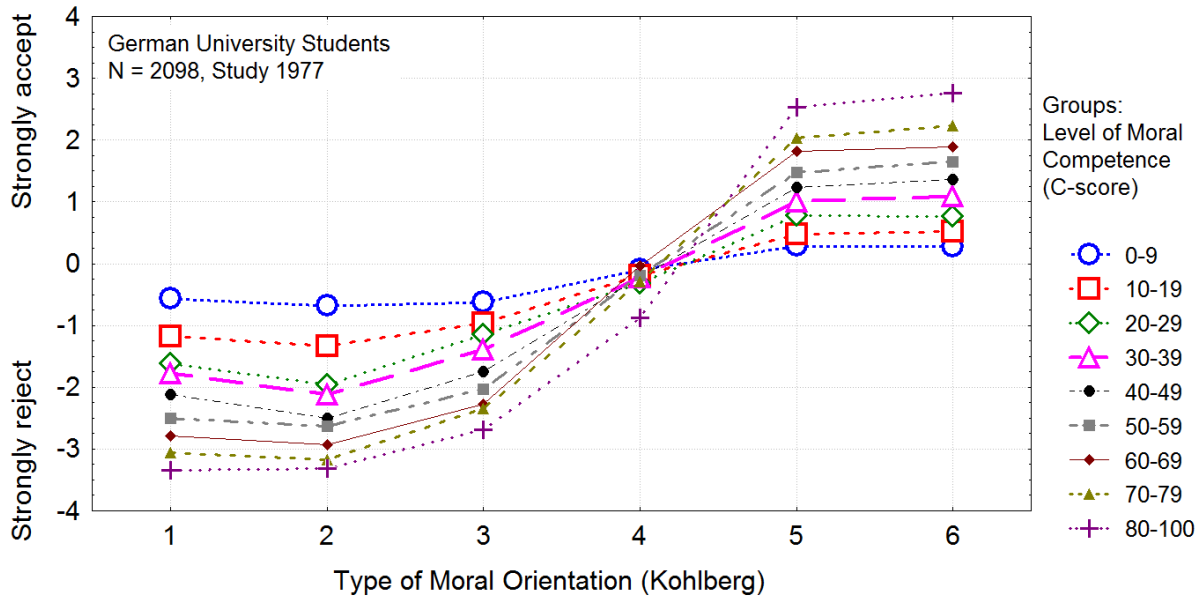
> *Nonfakeability*: Participants should not be able to simulate their competence test scores upwards as is possible in moral preference tests like the DIT. (Emler et al. 1983) In fact, we submitted the participants to the same experiment, as Emler did. In our experiment,

they showed no ability to fake their moral competence score upward. (Lind 2002; Wasel 1994; see also Chapter 5)

Preference hierarchy: Participants should prefer "higher" types of moral orientation to lower types for discussing moral issues. (Kohlberg 1958; 1984; Rest 1969) All studies with the MCT confirm this hypothesis, yet with some slight deviations (see below).

Simplex-structure of Type inter-correlation: Participants' preferences for "neighboring" types of moral orientation (e.g. type Four and Five) should correlate more highly with one another than with more "distant" Types of moral orientation (e.g., type Four and Two). (Kohlberg 1958, pp. 82-84) MCT data confirm this hypothesis even more clearly than the data from Kohlberg's studies with the Clinical Interview method. (Lind 1985)

Affective-cognitive parallelism: This hypothesis predicts that participants prefer higher types of moral orientation the more clearly, and reject lower types of moral reasoning the more clearly, the higher their moral competence is. (Piaget 1976) In fact, so far all studies have confirmed this prediction. (See the following Graph)



Graph: Empirical confirmation of affective-cognitive parallelism: the higher the moral competence of the participants the more they accept high-type moral orientations and reject low-type orientations when discussing a decision in a moral dilemma-story. (Lind 2008b)

Ruling out non-moral hints: When writing the arguments we tried not to give the respondents a non-moral hint for their ratings. We put the arguments in a random order, not in the order of their Types. We tried to write all arguments of about the same length (higher Type arguments often require longer statements). Recently Biggs and Colesante (2015, p. 511) have shown that all arguments of the MCT also have similar grammatical complexity. This adds to the findings that the ratings of the arguments are independent of linguistic cues. In contrast, the scores of the Moral Judgment Interview shows a substantial correlation (r = 0.30 to 0.40) with verbal fluency and complexity. (Kohlberg et al. 1984, p 329)

In order to appreciate these findings regarding the validity of the MCT better, one needs to do a little algebra. In educational and psychological research predictions are often formulated so that their confirmation "on the statistical significance level of $p = 0.05$" is almost guaranteed. For example, the "prediction" that two groups differ in regard to their moral competence is confirmed if the researcher collects enough data to reach this significance level. This is

because by enlarging the sample we can make the precision of a test so high that even tiny differences become statistically significant. They are usually true like the prediction that the weather next day will change or stay. In other words, such "predictions" do not advance our knowledge.

In contrast, the predictions derived from moral psychology are very risky. For example, the prediction that a participant will prefer different Types of moral orientation in a specific order is anything but trivial. The chance of being confirmed by chance is very low. We can calculate this chance in the same way we have calculated above the probability of Kohlberg's hypothesis of developmental invariance. There are 7! (Factorial of seven) = 1 x 2 x 3 x 4 x 5 x 6 = 720 different orders or permutations of Types. Hence the probability of one specific ordering (1, 2, 3, 4, 5, 6) is $p = 1/720$ or 0.0014. However, the theory does not only predict the preference order of one person but of all persons in a study. If the study comprises ten participants, the probability of confirming the hypothesis of a particular order becomes very, very small, namely $(0.0014)^{10}$ (read 0.0014 to the power of ten). Not all participants' responses confirm this prediction. The average preference for the Types of moral orientations deviates from the predicted order. Type-Five moral orientations are preferred slightly more than Type-Six ones, and Type-One more than Type-Two. A closer analysis reveals that the preference order fits better for the Doctor story than for the Workers story. These deviations do not call the validity of the MCT in question but Kohlberg's normative theory that the adequacy of moral orientations is the same for all types of dilemmas. Type-Six moral orientations might be adequate for dilemma in which ultimate moral principles are at stake, like the principle to preserve life. In the Workers story, people seem to think that this dilemma rather invokes Type 5 moral principle like the principle to preserve mutual trust in a relationship.

In sum, the Moral Competence Test has been submitted to rigorous tests of its theoretical and empirical validity. It meets all criteria even though these are more rigorous than the criteria usually used in educational and psychological measurement. (Lind 1978; 1985; 2002; 2008b)

"Reliability"

As I have mentioned, we did not base the MCT on conventional test theories, therefore we did not seek to optimize it for "reliability" in the sense which Classical Test Theory, and its successors use this concept. These theories falsely regard structural properties of test data as properties of the test instead of the participants who take the test. If we had submitted the MCT to "item analysis" and "item selection," as Classical Test Theory recommends, we would have certainly compromised its validity. Interestingly, in spite of the absence of any item selection the MCT shows a very high "retest reliability" of 0.90. (Lerkiatbundit et al. 2006, p. 101) Therefore, theory-base test construction may beat statistics-based test construction on its own turf. Nevertheless, this high correlation is not an attribute of the test but that the moral competence seems to be very stable.

One cannot prove the stability of a test through correlation but only by inspecting the test itself. The MCT has been stable since over forty years. Beside some minor editorial corrections, the MCT did not need to be changed. Thus, we can compare MCT data from recent studies directly with data from studies done in the late 1970ties, without the need to employ questionable statistical assumptions regarding their comparability.

4.3 Why ordinary tests fail to measure moral competence

"For a large part of the early 20th century it was generally held that moral phenomena cannot be studied scientifically," wrote Kurtines and Gewirtz in their handbook article on moral development. (Kurtines & Gewirtz 1995a, p. 3) Indeed, there are many approaches to the measurement of morality or of some of its aspects. Most are far off the target. Some are at least proxies and therefore allowed some valuable insights into the nature, relevance and teachability of morality, though they had also severe limitations. Some lack any definition of the "moral phenomena" which they studied so we cannot say whether they were valid. Some studies look at people's behavior, but only from the outside. They are measuring norm conformity, but not moral competence.

Behaviorist studies of moral character

I have already mentioned the early approach to assessing moral character by Hugh Hartshorne and Mark May (1928). They set up a whole series of experiments in which they tested their subjects' character. They gave them hard academic tests and observed them through one-way mirrors: Would they cheat or would they stay honest when an authority did not monitor them? Because of their counts of cheating behavior, they assigned an honesty scores to each participant. This approach is straightforward and seems to make sense. The less people cheat the stronger seems to be their moral character, and the more they cheat the lower seems to be their character developed. However, how can we know that such a relationship exists? The authors have been aware of this difficulty: "It is very difficult to build up a trustworthy empirical criterion for character tests." (Hartshorne & May 1928, p. 137) "At this point," they write some pages later, "we stumble against the problem of what we are really trying to measure after all." (p. 141) What is their solution? "If [the test] can be shown to be reliable, then it is ipso facto a valid measure of the particular behavior in question in the particular types of situation embodied in the test. […] Its validity is the square root of its reliability." (p. 142 and footnote) Although this solution is very common in educational and psychological measurement, it clearly misses the point. Reliability or precision of measurement does not tell us at all whether we really measure what we intend to measure. A thermometer may be very reliable. Nevertheless, it gives invalid data if we use it for measuring the length of a table.

To be fair, the authors eventually recognized the weakness of their measurement method. At the end of their book, they admit that it was a mistake to measure character by focusing only on an isolated act and to define this act externally cheating without considering the participants' intentions. (Hartshorne & May 1928, p. 377) Loevinger seconds: "Helping each other on examinations, which adults call 'cheating', does not seem to them to be a serious offense, as Piaget and others have shown. For them it is not evil per se but only because adults say so. Hartshorne and May, in doing one of the earliest studies of moral development, did not take seriously the child's own moral sensibilities and concentrated on offenses like cheating. To this day, many experimental studies of morality concern themselves with offenses trivial and inconsequential with no moral content whatsoever, such as choosing a smaller and immediate rather than a larger and delayed reward or looking at a toy the experimenter told them not to look at." (Loevinger 1957, p. 58) Thus, Hartshorne and May did not measure an internal trait (honesty) but external norm-conformity. Nevertheless, their measure showed some interesting correlations, which we cannot discuss here in detail. For example, religiously educated adolescents showed more cheating than adolescents who attended "progressive" schools. It seems that students who have learned to control themselves cheat less when they are not controlled than students who did not have such an opportunity.

Their failure to confirm the existence of moral character has led many psychological and educational researchers to eliminate human traits from their research agenda altogether. This shows when we browse through the subject index of many contemporary textbooks psychology. One of the most prominent scholars who deny the existence and relevance of moral traits are Philipp Zimbardo and Stanley Milgram. They believe that humans commit atrocities not because they lack morality but because they are forced to do so by their environment. "Moral factors can be shunted aside with relative ease by a calculated restructuring of the informational and social field." (Milgram 1974, pp. 6-7) But when Lawrence Kohlberg (1984) repeated Mailgram's experiments, he found that only participants with low moral competence obeyed the command to apply electroshocks to other people. In other words, traits count. I was told by an ear-witness that Milgram did not accept Kohlberg's finding because it was, as he admonished, based on clinical interviews.[8]

Clinical interviews

As an alternative to statistics-based tests, Jean Piaget and Lawrence Kohlberg have developed the so-called clinical interview method to study the structure of moral cognitions. However, both scholars were quite aware of the ambiguities and limitations of their method. On the one side, Piaget wanted to understand people's moral behavior. However, on the other side it "is the moral judgment that we propose to investigate, not moral behavior or sentiment." (Piaget 1965, p. 7) "A great danger, especially in matters of morality," Piaget admitted, "is that of making the child say whatever one wants him to say." (Piaget 1965, p. 8) "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 [...]. We do not claim to have solved the problem completely. Only direct observation can settle it." (p. 115)

Kohlberg also had his doubts about assessing moral behavior through interviews: "From the point of view of structural theory, a subject need not be self-consciously aware of his own stage structure." Nevertheless, he insisted, "Structural consistency […] should be found only by a psychologist abstracting structural features from spontaneous responses." (Kohlberg 1979; p. xiv) Thus, he was caught in a methodological dilemma. On the one hand, he criticized "behaviorist conceptions of moral conduct," because these "typically define conduct as moral if it conforms to a socially or culturally accepted norm." (Kohlberg 1984, p. 392) On the other hand, he envisioned a behavioral method of assessing moral competence or structure: "[Structure] is a construct rather than an inference, and is warranted only on the grounds of 'intelligible' ordering of the manifest items. One might say that the hypothetical structure is the principle of organization of the responses." (p. 408) Therefore Kohlberg asserts, "The test constructor must postulate structure from the start, as opposed to inductively finding structure in content after the test is made. [...] If a test is to yield stage structure, a concept of that structure must be built into the initial act of observation, test construction, and scoring." (Kohlberg 1984, pp. 401-402)

These quotes show that Piaget and Kohlberg themselves were not fully convinced by subjective measurement method like the clinical interview. Apparently, they thought that their method was the lesser evil.

Kohlberg fought a brave fight against statistics-based measurement of moral competence. "Dust bowl empiricism," he called the DIT. However, eventually he gave in to the ever-growing pressure from mainstream psychological research. (Kurtines & Greif 1974; Nico-

[8] Prof. Uwe Gielen, personal communication.

layev & Phillips 1979; Gielen & Lei 1991; Emler 1996) He agreed to revise his clinical interview method according to that test theory. (Colby et al. 1987) Apparently, he did not fully realize how incompatible Classical Test Theory (CTT) and his cognitive-structural concept of moral competence were: while Kohlberg originally defined moral competence internally ("based on internal principles"), he now sounds like an advocate of external scoring: "I include in my approach a normative component. [...] This normative concern has led me to rely upon philosophic as well as psychological theory in defining what I study. 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, p. 400) While he championed long for a structural approach to the measurement of moral competence, he and his team accepted the assumption that each isolated response of the interviewees reflects his or her moral competence: "My colleagues and I [...] have required each item in the manual to clearly reflect the structure of the stage to which it is keyed." (p. 403) He did no longer believe that a single item could not reflect structure. "Structure is an arrangement and organization of interrelated elements in a material object or system, or the object or system so organized." (Wikipedia, Sept. 2019) Originally he insisted that a measurement instrument must be designed for "testing theoretical propositions derived from the cognitive-developmental theory." (Kohlberg 1984, p. 195) Later he accepted statistical indicators of measurement error as a sign of test validity. "Test reliability and test construct validity is one and the same thing." (p. 424)[9] Yet statistics-based test theory is at odds with cognitive-structural theory of moral competence.

Statistics-based test theory

Most, if not all, current tests are based on statistical assumptions about human nature instead of on psychological theories. The mostly chosen statistical theory as a model for test design and test analysis is so-called Classical Test Theory (CTT). Item Response Theory (IRT) replaces this theory more and more. This theory extends CTT in important ways, but shares its basic assumptions. Therefore, I will discuss here only the former.

Classical Test Theory prescribes how we should design a test of human traits, how we should score it, how we should select the test-items and evaluated a test's quality. This theory is usually is expressed in this simple mathematical formula: X = Y + e. "X" denotes the participants' response to a test question or a test task, "Y" denotes the human trait which is believed to determine this response (e.g., moral competence) and "e" denotes some random processes intervening between the trait and the response. Thus, the term "e" accounts for error of measurement: "These 'random errors' are the only errors that will be explicitly considered in test theory. […] By random error we mean errors that average to zero over a large number of cases." (Gulliksen 1950, p 6) In other words, this theory rests on the idea of a one-to-one relationship between human traits on the one hand and test behavior on the other. The error term "e" is ideally zero or close to zero, if we repeat the measurement often enough, that is, if the test contains many questions or tasks ("cases") which measure the same trait.

It does not need great expertise to see that this model of the trait-behavior relationship is wrong. Hardly any human act is so clearly determined by only a single trait (motive, intention, ability etc.) that it can be represented by this model. "The problem of making inferences about a single trait from a set of responses all of which are multiply determined is a sub-

[9] My critical evaluation of Lawrence Kohlberg's research does not diminish my respect for his great achievements in moral psychology and education. Only late I realized how big the pressure might have been on him. When I told him at a conference in Starnberg, Germany, in 1978 how sad I was about his bowing down to his critiques, he said: "Let's have a beer together." We continued to talk about this at several occasions.

stantial one." (Loevinger 1957, p. 647) Most of people's acts are determined by more than one trait and the same act can be determined by a different trait in different people. This does not mean that traits do not exist. Rather it means that the statistical theory of a simple response-trait relationship is wrong. As Campbell concluded from his review of experiments, "in no case should a single overt behavior be regarded as the criterion of a disposition." (Campbell 1963 p. 162)

From a psychological point of view, the second term in their formula is also wrong. Advocates of Classical Test Theory argue that the lack of correspondence between isolated acts and hypothesized traits is due to purely random processes ("e") which intervene between a trait and the participants' responses to a test item. This "measurement error" or "unreliability," they say, can be easily reduced or even eliminated by repeating the measurement, that is, by including in a test several items, which measure the same trait. However, their assumption of random measurement fails the reality check. Purely random processes (like the throw of a dice, or the repeated measurement of a physical object like a rod) produce data, which are shaped like a bell-curve. However, test scores hardly ever produce such a bell-curve. (Micceri 1989; Walberg et al. 1984; Wuttke 2007)

Even if there was random error, the suggested cure is not feasible in psychological research. In contrast to physical measurement, we cannot repeat the measurement of a trait repeated in the strict sense of this word. If the researcher would ask the same question twice or would present the same task a second time, the participants would strike. They would think that it was a typo or that the researcher would question their trustworthiness. If their answers change, it would be because their intentions or motivation has changed but not because they made a random error when they answered the question. Therefore, researchers do not use identical questions or tasks for assessing the reliability of their test but use what they believe are *similar* items. However, if the items change, we should not be surprised that the participants' answers change, too, not because of random error but because the items changed. Repeating the measurement many times in order to increase the reliability and precision of the test, comprise its validity in two ways. First, with each repetition with similar items, the similarity of the items decreases. There is only a limited way to ask, "How do you feel?" Second, because of and during the repetitions the measured trait may change: For example, participants can become tired, become bored, or change their attitude toward taking the test. In the face of such insights, one should wonder why researchers still used Classical Test Theory and its successors for psychological and educational measurement.

To conclude, tests based on statistical theories are not suited to measure moral competence. Probably they are not suited to measure any human trait. They focus on isolated responses to test-questions, thus ignoring the relational properties between an individual's responses. Therefore, Mischel and Shoda have proposed a "new conception of personality in which such patterns of variability are seen not as mere 'error' but also as reflecting essential expressions of the same underlying stable personality system." (Mischel & Shoda 1995, p. 246) This is what we have successfully done by making the Moral Competence Test forty years ago.

Self-report questionnaires

Some researchers have suggested measuring morality and character through self-description questionnaires (Wang et al. 2015). For example, for assessing the virtue of 'honesty' (*trustworthiness*), they provide statements like "One can rely on me to tell the truth." In response, the participants are able to select on a 5-point scale, "That's not at all like me" to "That's me exactly." While it seems to some degree to be possible that people can estimate their own

moral competence it is very questionable whether they respond objectively to such questions. Wang and colleagues found no confirmation for their hypothesis that their activities with the Scouts improved their character. On the contrary, they found a decrease in test scores with age for non-Scouts. This strongly suggests that their instrument does not measure morality, but rather participants' adjustment to the (alleged) expectations of researchers; this adaptation usually decreases with age, only not in Scouts, who may be more obedient to authority. Here we may recall what Jane Loevinger wrote. "Self-reports concerning personality traits are subject to such massive systematic distortions as to make the virtually worthless as direct measurements of personality traits." (Loevinger 1957, p. 646)