DRIE ARTIKELEN OVER
DE VALIDITEIT VAN SPELSIMULATIES

Nijmegen, 2011

Samen spraak
spel simulaties

Informatie over Spelsimulaties

DRIE ARTIKELEN OVER
DE VALIDITEIT VAN SPELSIMULATIES

Inhoud

Vooraf 1

1 The validity of games 3

2 Validity of Games/Simulations: A Constructive View 13

3 The Validity of Laboratory Research in Social and Behavioral Science 23

4 Referenties 39

Vooraf

Validiteit is een belangrijk begrip in wetenschappelijk onderzoek, waarnaar veel onderzoek is gedaan en waarover veel is geschreven. Validiteit staat voor geldigheid en betreft de vraag of de uitspraken die je met je onderzoek doet geldig zijn, ofwel overeenkomen met de onderzochte werkelijkheid. Validiteit heeft dan betrekking op de meetinstrumenten die je gebruikt (c.f. inhoudsvaliditeit: meten we wel precies dat wat we beogen te meten), de manier waarop we de bevindingen interpreteren en tot conclusies komen (interne validiteit) en de mate waarin we de conclusies ook mogen toepassen op andere dan de onderzochte groepen of situaties (externe validiteit).

Wanneer het gaat om het uitvoeren van wetenschappelijk onderzoek, dan is er een hele bibliotheek aan literatuur beschikbaar rond het onderwerp validiteit. Allerlei vormen van validiteit worden uitvoerig beschreven en uitgewerkt. Een onderzoeker kan van die literatuur gebruik maken om de kwaliteit van zijn onderzoek te staven.

Anders is het wanneer we kijken naar spelsimulaties. Ook bij het toepassen van spelsimulaties speelt validiteit een belangrijke rol. We kunnen daarbij onderscheid maken tussen twee toepassingsgebieden van spelsimulatie:

- een spelsimulatie wordt toegepast als onderzoeksinstrument: personen ('respondenten') worden dan in een gesimuleerde omgeving geplaatst en aan het werk gezet, en de onderzoeker registreert dan hoe zij te werk gaan, op welke manier zij tot beslissingen komen, hoe de beslissingen er uitzien, et cetera. In dit geval moeten aan de spelsimulatie dezelfde methodologische eisen gesteld worden als aan elk ander onderzoeksinstrument, waaronder de validiteit. In feite is dan alles wat in de genoemde methodologische literatuur vermeld wordt onverminderd van toepassing op de spelsimulatie.
- een spelsimulatie kan ook worden toegepast als 'trainingsinstrument': de deelnemers moeten aan de hand van de ervaringen in de spelsimulatie komen tot meer inzicht in de werkelijke situatie of tot een goede keuze van de acties die men gaat uitvoeren. Ook in een dergelijke toepassing moet de spelsimulatie zodanig zijn, dat de conclusies die men op basis van de spelsimulatie verbindt ten aanzien van de werkelijkheid geldig zijn voor die werkelijkheid. Als deelnemers na afloop van een spelsimulatie allemaal zeggen: 'Dat was leuk en leerzaam, maar in de werkelijkheid gaat het er heel anders aan toe', dan is die spelsimulatie blijkbaar niet valide geweest voor die betreffende situatie.

Bij dit laatste soort toepassingen gaat geldigheid / validiteit van de spelsimulatie hand in hand met het realistische gehalte van de simulatie. Dat is iets wat de meeste bouwers en gebruikers van spelsimulaties wel beseffen (of ze zijn er door deelnemers nadrukkelijk op gewezen in de evaluatie). Maar wat het concept van validiteit in deze context precies inhoudt, daarover is minder bekend, en daarover is slechts mondjesmaat gepubliceerd.

In de afgelopen jaren hebben we een drietal (Engelstalige) artikelen gepubliceerd waarin het concept validiteit wordt uitgewerkt in de context van spelsimulaties. Wij zetten deze drie artikelen, als een soort drieluik, bij elkaar in dit rapport, in de hoop dat de gedachten over validiteit van spelsimulaties daardoor makkelijker toegankelijk zijn.

Het eerste artikel *'The validity of games'* gaat in vrij algemene termen in op wat het begrip validiteit betekent in de context van spelsimulaties en op welke manier de validiteit van een spelsimulatie bedreigd kan worden.

In het tweede artikel *'Validity of games/simulations: A Constructive View'* wordt het concept nader uitgewerkt naar een drietal aspecten van het ontwerpen en gebruiken van een spelsimulatie, namelijk het onderliggende concept, de spelsimulatie zelf, en de gedrag van de deelnemers in de spelsimulatie. Voor drie fasen –het ontwerp, de uitvoering en de debriefing- wordt een aantal vragen geformuleerd die behulpzaam kunnen zijn bij het bewaken van de validiteit op elk van de drie genoemde aspecten.

Het derde artikel *'The validity of laboratory research in social and behavioral sciencs'* gaat dieper in op het vraagstuk van de externe validiteit, namelijk het representeren van de werkelijkheid in een spelsimulatie. Om een beter zicht te krijgen op deze vorm van validiteit wordt discussie verbreed naar andere vormen van onderzoek, waarin de werkelijkheid vervangen wordt door een andere situatie.

Dr. Vincent Peters

Drs. Marleen van de Westelaken

1 The validity of games[1]

Vincent Peters
Geert Vissers
Gerton Heijne

Abstract

We are often confronted with a complex situation we have to learn about or we have to teach others about. One way to deal with complex situations is the simulation approach: build a simplified model of this reality, learn from or teach about this simplified model and, finally, translate the findings or knowledge back to the reality. Gaming is based upon this idea. If we want to make inferences about reality based upon experiences and knowledge acquired in a game, we have to be sure that the game model is a good, or in other words, a valid representation of the real situation. In this paper the concept of validity is explored in relation to games and simulations; four aspects of validity are distinguished that apply to simulations and games. These aspects are related to three applications of games. The paper finishes with factors that may threaten validity during the process of the game design; a few suggestions are made to avert these threats.

Keywords

Education; game design; gaming; research; simulation; threats for validity; validity

Introduction

In the field of research it often occurs that we have to answer research questions about situations that cannot be investigated directly. Research on future situations is a clear example. Vissers et al. (1995) mention other instances where the situation under study is inaccessible for researchers. Similar problems occur in teaching. In many situations it is impossible to teach or train students in the real situation, e.g. because the situation is too complex or because one is required to have certain knowledge or skills before one can be admitted to that situation (cf. the training of a pilot or surgeon). A third field is the development of new policies. If a policy maker

[1] Gepubliceerd als:
Peters, V., Vissres, G., Heyne, G (1988). The validity of games. In: *Simulation & Gaming. An International Journal of Theory, Practice, and Research*, March 1998, Volume 29, no. 1, p. 20-30.

wants to experiment with new policies to assess effectiveness and to explore possible negative side-effects, the real situation is not an appropriate place for these exercises.

In all these instances the researcher, the teacher or the policy maker can resort to gaming (or another form of simulation) to answer the research questions, transmit the desired knowledge, or get insight in the effect of policies.

1.1 The simulation approach

If we use simulations to learn from or teach about problems or situations, we first make a simplified model of the situation; next, we learn from or teach about this model; finally, we translate the findings or knowledge acquired in the model back to reality. The problem or situation that is the subject of our research, teaching or policy, is called the reference system. It is the point of departure for the simulation approach. In order to create a model, we describe the elements of the reference system and the relations between them in terms of another (known) system.

We can use various types of models. In case of a mathematical model, we use variables and mathematical functions to describe the elements and their relations; in a conceptual model we use concepts to indicate the elements and linking arrows to establish and describe the various relations; in a physical model we have physical objects and the spatial arrangements between them. In the case of gaming, elements of the reference system and their relations are represented by design elements like scenario, events, roles, rules, and accounting system.

The process of designing and applying a game can be represented as in Figure 1.

The arrow at the left indicates the process of the game design. The reference system has to be translated into a usable game. That is, we have to get a good understanding of the characteristics of the reference system and transform these characteristics into the elements that constitute a game. Next, the game is played by participants; this will result in new information and/or new knowledge and experiences. Depending on the kind of application and the objectives of the game, the output of playing the game can be of interest for the researcher or for participants themselves. For this, observations and experiences made in the simulation have to be translated back to the reference system. This is indicated by the arrow at the right. In the context of education and policy this is often referred to as debriefing (see e.g. Lederman, 1992). The arrow shows that the reference system can be considered as the target point for the gaming process. Since the reference system is also the starting point of the gaming process, we see that the simulation or gaming circle is closed.

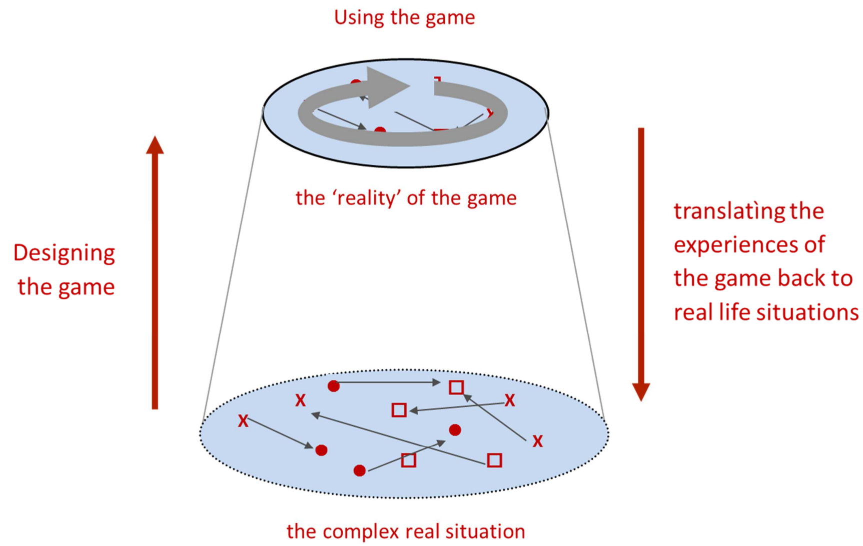

Figure 1 Designing and applying games

When games are applied in the described context, the basic assumption is that we are able to translate acquired knowledge and experiences from one system to another. The extent to what this translation will be successful depends, among other things, on the degree to which the game is a valid representation of the reference system. In other words, the strength of our conclusions about the reference system is determined by the validity of the game model.

1.2 The concept of validity

There is a vast amount of literature about the concept of validity, but this literature focuses mainly on the validity of experimental situations (cf. the concepts of internal and external validity) or on the validity of measurement instruments (cf. the concepts of content and construct validity) (Cook & Campbell, 1979; Cronbach & Meehl, 1955). These aspects of validity refer to the correspondence of a specific research method (e.g. the experiment) or the results of a research act (e.g. data gathering by means of a questionnaire) and the reference system, i.e. that part of reality the researcher wants to investigate. The concept of validity in relation to simulations and games as a simplified model of a complex reference system is hardly elaborated in the literature.

A very general definition of the concept of validity in relation to games is: the validity of a game is the degree of correspondence between the reference system and the simulated model thereof. This is not a very accurate definition since the concept of correspondence is not clarified. If correspondence means that each (relevant) element of the reference system has to be literally translated to the game model, we have a very narrow definition of validity; this definition implies that games that are based upon a metaphor cannot be valid. But if we relax the definition of the word correspondence, what does it mean then? What criteria do we have to apply to assess the correspondence? In addition, the question whether the correspondence is sufficient depends on the objectives of the game. For one purpose the game can seem to be a valid representation while it is not for another purpose.

With respect to the use of gaming in research, Raser (1969) has defined validity of models in the following way: 'a model can said to be valid to the extent that investigation of that model provides the same outcomes as would investigation in the reference system'. This definition does not stress the correspondence between the two models, but validity is based upon the results of the usage of the model. This utilitarian definition of validity can also be applied to other applications by replacing the word 'investigation' by terms like 'learning', 'taking decisions', and so on.

Raser (1969) has suggested four criteria for the validity of gaming as a research tool: psychological reality, structural validity, process validity and predictive validity.

The first criterion for validity, according to Raser, is *psychological reality*. A game is valid to the degree that it provides an environment that seems realistic to the players. If they fail to see the game as realistic, they possibly tend to show different behavior than they would do in real life situations, or they tend to take more risks. The result will be that behaviors in the game do not correspond to behaviors in the reference system.

Structural validity is the second criterion for validity distinguished by Raser. This criterion is formulated as follows: 'a game is valid to the degree that its structure (the theory and assumptions on which it is built) can be shown to be isomorphic to that of the reference system' (1969:144). Above we have pointed at the elements in the reference system and the relations between them. These elements (actors, information, data, laws, norms, etcetera) and the way they are connected should be reflected in the game model. The word isomorphic indicates that these elements and relations in both systems do not necessarily have to be similar, but there must be a congruency between them. Since modeling means that we try to build a simplified model of the reference system, it is not necessary that all elements and relations are represented in the game model. So, this aspect of validity implies that the most important features of the reference system should be included in the game model in a isomorphic way.

Process validity, the third criterion for validity, implies that 'a game is valid to the degree that the processes observed in the game are isomorphic to those observed in the reference system' (Raser, 1969:144). The previous criterion stated that there should be a congruency between the elements in the game system and the elements in the reference system. In a similar way, this third criterion states that there should be a congruency between the processes that take place in both systems. In this respect we can, for instance, think of flows of information or resources, interactions between actors, and negotiations.

The last criterion is *predictive validity*: 'a game is valid to the degree that it can reproduce historical, outcomes or predict the future'. This criterion refers to the accuracy of the outcomes of the game: are we able to make a good estimate or prediction of what happens in the reference system? We can assess the validity of a game by trying to reconstruct known situations. The results of the game can then be compared with the result in reality. If this so called 'postdiction' proves to be sufficient we feel more confident about the game and its predictions about future situations.

These four criteria for validity help us to get a better understanding of the general concept of validity. Raser has described them in relation to simulations and games for research purposes. We will look at three different applications of games to see whether and to what extent these four criteria are applicable.

Table 1, derived from Geurts & van Wierst (1991), gives a short characterization of three applications of games. We will explain each application shortly and, next, see how the criteria for validity apply to the specific situation.

	Function of the game	**Dominant communication**	**Desired outcome**
Research	offer a stimulus	model → researcher	data to answer research questions
Teaching	conveying medium	game → players	cognition and skills
Policy	create conditions	players → players	policy options and solutions

Table 1 The characterization of three types of application of games

1.3 Gaming as a research tool

If a game is applied as a research tool, a researcher has one or several research questions about the reference system, though without being able to collect necessary information in the reference system itself (e.g. because it is inaccessible or the questions concern a future situation). A game model is constructed to collect the desired information and the game is played. After gathering the information before, during and after the game, and after analyzing it, the researcher has to translate the findings to the reference system, i.e. the researcher has to draw conclusions about the original problem. The position of the game in the research process can be compared with other situations where the researcher can collect the desired data: the real life situation, a test situation or an experimental situation. The need to draw valid conclusions about the reference system on the basis of information gathered in a game, makes high demands upon the game. The game should be constructed in such a way that it is plausible that participants behave in more or less the same way as they would have done in reality. Therefore, the game should appear as realistic to participants. There must be a rather strong resemblance between the game model and the reference system (structural and process validity), and the researcher must also have indications that the outcomes of the game (i.e. the data gathered) are of a high quality, i.e. have a high predictive validity towards 'reality'.

1.4 Gaming as a teaching tool

The second application of the gaming approach is the situation in which we want to teach people about the reference system or how to act in a new situation. They have to acquire insight in the reference system and/or they have to learn skills. If this reference system is too complex we can use a game to provide students with new knowledge or to offer them the possibility of training new skills. After the learning or the practice students will have to apply their new knowledge or skills in the real situation.

A special feature of the application of gaming as a teaching tool is that the teacher and the game designer know beforehand what has to be learned by the participants; in other words the desired output (knowledge and skills) of the game is known and so are the standards that must be met by the participants. The learning elements needed to acquire the desired knowledge and skills must be included in the learning environment (i.e. the game) and they must be conveyed to the participants. If the knowledge and skills have to be applied directly in 'reality', the game environment should have a strong resemblance with that reality (cf. the training of a pilot in a flight simulator). If, on the other hand, the game pertains to more general knowledge and skills (cf. a game about negotiation skills), there is more latitude for game design. The fourth criterion, predictive validity, seems less important in this application. Since the desired output is known in this kind of applications, the earlier mentioned utilitarian definition of validity seems to apply very well: the game is valid to the degree that the learning objectives are achieved by the participants.

1.5 Gaming as a policy tool

In the third application, gaming as a policy tool, the game is designed as an environment in which participants can explore possible policy options to solve a problem or to improve a situation. They are placed in a situation where they can invent options, experiment with them, consider the results of these options, and compare options in terms of effectiveness, efficiency, and so on. Playing the game shows these results to the players within a very short period of time, in contrast to reality where it may take months or even years before the impact of a policy becomes perceivable. The game environment should at any rate be open, meaning that it should not guide the participants to only one solution. Rather, the environment should challenge participants to explore several solutions. It is obvious that against this background the concept of validity is valued different from the two former applications. The reference system should be represented in the game model, but not in a very restrictive way: participants should be able to explore new strategies and behaviors. Since the results of the game give information about the reference system, the outcomes of the game must have some predictive power towards the reference system.

This exploration of the concept of validity in relation to different uses of simulations and games has shown that there are several aspects one has to bear in mind when talking about the correspondence between game model and reference system. These criteria seem to apply to all situations where gaming is used. But the exploration has also has shown that it will not be possible to give general guidelines about how to apply each of these criteria. The value of each criterion, and the balance between them, is highly dependent on the specific objectives of the project of which the game is a part.

We started this section with the observation, that there is hardly any literature on the validity of games. In fact, the book of Raser, cited several times in this paper, is the only book found by us in which this subject is treated systematically. Some authors have used the ideas of Raser, but there have been no genuine renewals of the thoughts about the validity of games. Raser's book was published in 1969, five years before one of the bibles on gaming was published: Duke’s 'Gaming: The future's language' (1974). Since then thinking about gaming, the game design process and the applications of games have gone through an enormous evolution. Thinking about the validity of games seems to be still in the same phase as in 1969. We think it is time that Raser's important work on the validity of gaming should be updated in light of the new developments in gaming.

In the rest of this paper we will identify some factors that might threaten the validity of games, and we will describe some guidelines that might be useful in avoiding these threats. In this paper we will limit ourselves to factors that are related to the process of the game design, i.e. the arrow at the left in Figure 1.

1.6 Threats of validity

The design process of a game is based upon three principles, namely reduction, abstraction and symbolization. In the process of translating the reference system into a simplified game model we apply these three principles. Reduction means that we make a selection of elements from the reference system that have to be included in the game model: we include the elements that seem relevant to us and we leave out the elements that are less important. The second principle, abstraction, implies that the elements included in the game model are not necessarily as detailed as they are in reality: we deliberately simplify them in order to make our model less complex. The last principle, symbolization, deals with the fact that the elements and relations of the reference system are molded into a new symbolic structure, namely into scenario, roles, rules, symbols, which are the most important basic elements of a game. Some game elements may quite resemble their counterparts in reality, but other elements may undergo a metamorphosis and have a complete different appearance in the game model.

During the design process we can make errors concerning each of the three principles just mentioned. We can wrongly decide to leave out some very essential elements or relations from the simulated model, or we can include elements of minor importance in it; both errors result in the wrong aspects being emphasized in the game model. Or we can introduce in the game model too vague or too detailed elements, which may have the same result. In addition, we can transform elements into such a symbolic structure that the participants fail to see the link with the reference system.

Making these kind of errors can be imputed to several causes. One possible cause is that a designer fails to take full account of the objectives of the game. Another cause might be lack of thorough knowledge of the reference system on the part of the game designer; in that case, the designer will not be capable to estimate the relative importance of the elements of the reference system correctly, and thus, runs the risk of making the wrong decisions concerning the inclusion or exclusion of elements. Furthermore, making wrong selections may be caused by a designer being too strongly focused on the game model and the eventual game; as a consequence, a designer may be guided by the opportunities and/or the restrictions of the game, instead of the features of the reference system and the objectives of the game.

All errors mentioned jeopardize the extent to which a game model corresponds to the reference system, in other words these errors are a threat to the validity of the game. During the design process we can take some measures that help us to prevent us from making such errors. We will mention a few possible precautions and checks here.

The first guideline concerns the design process. In fact it is very simple: work systematically. This advice might seem like hammering on an open door, but it is a first requirement for a good result. We can be a bit more precise in this. Working systematically means a thorough analysis of the reference system; this analysis should focus on the structure of the reference system as well as on the processes in that reference system. A second precautionary measure is to make clear deductions and small steps during the design process. I.e. one should not translate the reference system at once into a game model, because in such a way one cannot sufficiently be aware of how elements of the reference system are expressed in the game model. It is advised that one should discuss the decisions one has made with other persons, especially with the client for whom the game is designed. A participative way of working, in which the client is highly involved in the design process will give the opportunity of a constant discussion of the steps and the decisions. The methodology, described in Greenblat & Duke (1975) and later adapted and elaborated by many other game designers, offers the necessary support for working systematically.

Another way to improve the validity of a game is to check the validity explicitly, that is to present the concept of the game to other persons and ask them for their opinion about the correspondence between the game model and the reference system. There are two possibilities for doing so. One can discuss the validity of the game with other game builders and ask them to judge the game from their expert view. Alternatively one can discuss the game concept with experts on the subject of the game. This discussion should encompass all four aspects of validity distinguished by Raser, that have been described in a previous section. This way of discussing the game and its validity with experts (either on the game building process and on the content) is referred to as 'peer debriefing' in traditional methodological literature (Guba & Lincoln, 1982). The other option is to present the game concept to future game players (since the game is not ready yet, we have to use future players) and ask them for their opinion about the validity of the game; this procedure is referred to as 'member check' (LeCompte & Goetz, 1982). Since one can expect that the experts and the future players will focus on different aspects of the game and its validity, these two approaches can be considered as complementary.

The third way to check the validity explicitly is to test the game extensively. Most games are tested before they are released and applied officially. However, these tests tend to focus mainly on the logistics of the game: are all descriptions clear to the participants, do we have sufficient forms, can the players accomplish their tasks within the available time, and so on. But we should also use these test runs to confront, if possible, the games explicitly with the reference system. It can be very useful to have the test runs attended by observers concentrating exclusively on questions concerning aspects of validity.

We have pointed out a few measures a game designer can take to improve the validity of a game. There are other measures, concerning the phases of using the game and the debriefing, that have not been discussed in this paper.

We have explored the concept of validity in relation to games and simulations. This exploration has shed some light on the phenomenon, but it has also made clear that there is need to further clarify the concept of validity.

2 Validity of Games/Simulations: A Constructive View[2]

Geert Vissers
Vincent Peters
Gerton Heyne
Jacques Geurts

Abstract

Justification is an integral aspect of any process of knowledge acquisition. The assumptions made, the methods applied, the reasoning used, must be sound. In the domain of social research, justification issues are captured by the concept of validity. This concept, though, largely derives from a tradition in which great value was attached to standardized, generally applicable instruments, psychological tests in particular. Accordingly, validity was primarily seen as an attribute of instruments. In current research practice, this close relationship between validity and instrument seems to survive, despite later developments in thinking about validity. In addition, validity is widely understood in terms of 'threats'. However, research findings are not valid in the absence of error, but they are valid in the presence of tenable assumptions, proper methods, cogent reasoning. Therefore, research may benefit from prospective guidelines rather than from criteria to be used in retrospect. This paper argues that instrument-oriented and reactive conceptions of validity, likely to be disadvantageous in general, are certainly so in the field of simulation gaming. A pro-active approach is advocated instead. A model is presented that comprises an activities-dimension (design, application, debriefing) and a dimension that involves different 'products' of using a simulation game. These products may serve as a starting point for justification. The model provides researchers and practitioners in the field of simulation gaming with a heuristic scheme for thinking about validity in a way that suits the (ongoing) research project at hand, concerning not only the design of a simulation game but also its application and results, and the added value of simulation gaming in a particular research project.

[2] Gepubliceerd als:
Vissers, G., Peters, V., Heyne, G., Geurts, J. (1998). Validity of games/simulation: a constructive view.
In: Geurts, J., Joldersma, C., Roelofs, E. *Gaming/simulation for policy development and organizational change*, ISAGA '97, Tilburg: Tilburg University Press, 353-359. Proceedings of the annual conference of the International Simulation and Gaming Association (ISAGA) of 1997 in Tilburg.

2.1 Introduction

Assessing the validity of a simulation game is a demanding, yet hardly satisfactory task, if validity is taken the conventional way. One point is that conventional approaches to validity have developed outside the field of simulation and gaming, and must be adapted to meet the specific demands of this field. More important, however, is that such an attempt to adapt is likely to be disappointing, because of inconvenient properties of the concept of validity as it is currently used in the wider field of social, psychological and educational research.

Validation usually means that some more or less standardized, widely accepted validity types are used to assess The Validity of, usually, instruments for data collection. This practice (and the types involved) seems a direct inheritance of psychological test theory, the domain of research in which the concept of validity was first developed. However understandable from a historical viewpoint, this practice and the associated types have clear disadvantages:

There is a collection of multifarious validity types to draw upon: Types vary in degree of generalization; some are complementary; some are incompatible; some are defined in a variety of ways; some are obsolete. This collection may give rise to confusion and selection problems. Researchers who resort to textbook advice will probably be left with the types endorsed by the American Psychological Association in 1954 already (APA, 1954), viz. content, predictive, concurrent, and construct validity, perhaps accompanied by internal and external validity, notions coined by Campbell only a few years later (Hammersley, 1991), and sometimes Brunswik's (1955) ecological validity. Already the APA-endorsed types reflect different research traditions (Ebel, 1961), not to mention Campbell and Brunswik. Thus, a type-approach to validity may remain confusing, even if the small list presented by many a textbook is relied upon. Further problems are that subjects not captured by one of the types may readily be overlooked, and that more recent contributions to validity - like the tendency to conceive of validity as a quality of propositions rather than of instruments (Hammersley, 1991; Joint Committee on Standards for Educational Evaluation, 1994) tend to be ignored.

Yet, many available validity types do have one thing in common: They reflect a clear preoccupation with research instruments, particularly instruments for data collection and the way these are used. Much has been written about the validity of tests, questionnaires, and research settings (including experimental situations), mostly in an attempt to specify the instrument's domain of generalizability. In other words, prevailing validity types tend to emphasize the characteristics shared by different people, populations, situations, behaviors, or whatever the research unit, instead of the distinctive characteristics of the subject under study in a particular case.

In short, validity continues to be associated with separate research instruments; other aspects of research nor the integration of instruments and other research aspects is given much attention. Furthermore, the type-approach to validity hardly allows an answer to the question whether (or to what extent) a generally applicable instrument is adequate for a specific research project. Often the question to be answered is not: 'Is this instrument valid', but rather: 'Is this instrument sufficiently valid for present purposes'.

The comments made refer to common validation practice, which is often straightforward, more than to the argumentation underlying the various types, which is often very careful. Still, even these argumentations show a certain preoccupation with instrument, and a tendency towards generalization. These remarks apply also to Raser's (1969) discussion of validity, that is probably the most systematic attempt to adapt conventional notions of validity to the field of simulation gaming. Concentrating on simulation as an instrument for scientific research, Raser suggests four validity criteria - viz. psychological reality, structural validity, process validity and predictive validity. These are broad concepts, defined by Raser in a way that does not meet many a researcher's first impression. For example, the sentence 'Perhaps all that is required is that the structure seems realistic to the players, that it conforms to their ideas as to what constitutes reality, not that it accurately reflect what is actually "out there"' (Raser, 1969: 151) refers to process validity, not to psychological reality. This latter aspect concerns two other questions: (1) Is a simulation game so involving that subjects forget they are conducting activities that have no future repercussions? (2) Does a particular game provide the possibilities for involvement and action needed to stimulate a wide range of responses? Consider the way many researchers conceive of 'psychological reality', as an illustration of the classical problem that concepts become detached from the argumentation that brought them about.

2.2 Arguments in favour of a constructive approach

'Reactive' seems a proper characterization of current validation practices. The phrase 'threats to validity' is recurrently used, suggesting that 'validity' is kind of a natural state of affairs in research, but also that this state is at risk since there are many errors waiting to be made. Afterwards, it must be checked that these errors are successfully avoided. Thus, considerations concerning validity contribute marginally to the research process, only indicating that some decisions should not be made.

A more constructive approach to validity is possible. If validation relates to questions concerning the suitability of a research instrument or a set of instruments, the choice (or creation) of a proper research setting, the quality of reasoning, the tenability of conclusions, then validity can be described as an aspiration that guides the steps in a process of knowledge acquisition. The word 'aspiration' brings, in Homans' words, men back in. Rather than an intrinsic quality of instrument or reasoning, validity comes to be seen as an attribute of the connection between the steps in a research process and the researcher's objectives and wider goals. These goals are indeterminate: Explanations ask for further explanations, theories can be extended, new perspectives can be adopted, new facts may present themselves. The point here is not that validity is necessarily temporary (though it is), but that it is supposed to support and strengthen ongoing research, rather than serving as a closure mechanism. This applies to the empirical cycle at large, as well as to its constituting cycles called 'research projects'.

The next section presents a model for constructive validation in the field of simulation gaming. The model is constructive in the sense that it may support the ongoing research process; it does not confine to afterwards judgment. The field of simulation gaming seems very suitable for the task of developing and investigating a new approach to validity. As compared with other domains of research, this field is less burdened with a validity tradition. That is an omission to be filled, but it seems also an advantage: In the absence (more or less) of such a tradition, usual impediments to change such as vested ideas and doctrines of method may scarcely present themselves. Another reason, equally important, is that a constructive approach to validity seems of great value in the field of simulation gaming, more perhaps than in other domains of research. This is partly because of the intricate position of simulation participants - who are both (passive) subjects of observation and (active) agents in the production of the simulation's reality - and partly because the events and processes in a simulation game are often hard to foresee. If somewhere, early justification and a procedure for interim adjustment (as an alternative for improvisation) are needed in the domain of simulation gaming.

One more reason needs to be given in favor of constructive validation in the field of simulation gaming. If conventional, instrument-oriented notions about validity are applied in the field of simulation gaming, whether or not in adapted form, the conclusion becomes inevitable that isomorphic simulation models (that copy a certain reference situation, albeit in a simplified way) are superior to metaphoric models (that deliberately introduce rather unfamiliar rules and designations). There are situations in which metaphoric models are to be preferred, however, for instance when a simulation aims at exploring how actors will behave in the absence of (experienced) constraints of some particular reference situation.

2.3 A model for reasoning

We have discussed several drawbacks of validity types, e.g. that these types may direct attention away from important aspects in a particular research process, that they may elicit reactive behaviors, and that they may be used without reference to the arguments they derive from. We will not offer a new set of types or criteria instead. Such a set would run the same risks as previous sets have, and it might even (who knows) become part of the existing collection of validity types, which presumably would not lessen confusion. We rather present a model that we hope will provide researchers and practitioners with elements for developing their own line of reasoning with respect to quality of knowledge, as acquired in the course of a particular research project. We start presenting a table that combines a list of successive activities in a simulation gaming project with a list of 'products' that can be taken as sources for evaluation or justification.

		Activities in the process of knowledge acquisition, specified for simulation gaming		
		design	application	debriefing
Products of research, as sources of evaluation criteria	concept	××	×	×
	instrument	×	××	×
	behavioral outcomes	×	×	××

Including a list of activities serves more than a single purpose. It exhibits the constituting stages of the wider process of knowledge acquisition through simulation gaming, each of which can be judged on its own merits. This judgment may pertain not only to separate means and tools applied in a particular stage but also, equally important, how these fit together. A sequence of stages/activities thus draws attention to the fact that stages are or should be linked. Recognition of these links allows 'backward mapping' (Elmore, 1985), which means asking whether a certain stage is a likely result of actual processes in the preceding stage.

For instance, assumed that a simulation game has to serve specific learning objectives, a sequence of 'backward questions' might start with debriefing: What learning processes are to take place in (or as a result from) debriefing? For these processes to take place, what requirements have to be met in the application stage? What kinds of design characteristics are helpful or even necessary for these requirements to be met? Such a sequence of questions may assist the very designing process, but it may also contribute to justification.

The example, though illustrating the prospects of 'backward mapping', also implies that a list of activities is somewhat arbitrary, in the sense that the last stage is not the end of the process (nor is the first stage its beginning). This means - whether forward or back mapping is used - that at least some criteria for justification are external to the model, and can neither be defined nor controlled by the researcher. It follows that it is beyond a researcher's competence to make statements on The Validity of some simulation game or of the conclusions drawn from it - to which it must be added that a researcher is obliged to make all statements on validity that can be made from a researcher's vantage point. In the next section, we will discuss the role of some other actors. Finally, perhaps needless to say, the stages mentioned in the table may be replaced with other stages, if necessary.

The discussion thus far refers mainly to 'outcomes' as a source of evaluation criteria. Other sources criteria must be considered as well. Let's return to the discussion of 'backwards questions' and ask - for instance - how to choose if two different designs are expected to produce comparable (and desired) processes in the application stage. We must be able to compare the qualities of both designs on the basis of other criteria than 'outcomes', for that criterion fails to distinguish. Theory is a likely candidate. A given design will represent theoretical concepts, assumed cause-and-effect relationships, a province of content, and in these respects it can be compared with another design. In Elmore's vocabulary, this may be called 'forward mapping': the logic to get from a starting point to a result. Theory is indeed such a starting-point; it is already present before any particular designing attempt is made.

Theory may turn out to be inadequate in the course of a research project. That is not Elmore's first problem, though his approach does not prevent its being considered. Instead of focusing on theory alone (as common in epistemology), however, Elmore stresses 'reversible logic', "so commonplace that we often don't recognize it, much less exploit it. The logic is essentially this: To get from a starting point to a result, we don't just set an objective and go there. We begin at either end and reason both ways, back and forth, until we discover a satisfactory connection" (Elmore, 1985: 35). In the course of this iterative process, theoretical notions may come to be adjusted.

Elmore discusses 'reversible logic' in an attempt to show that even if you know where you are and where you want to go, the process of getting there is often more complex than it seems. We may agree on that, only adding that often we do not know exactly where we are, or where we want to go. Put in other words, theoretical notions may not be explicit and articulate enough to offer criteria to evaluate design (if we concentrate on that stage). If such is the case, we may fall back on methodological criteria, e.g. simplicity, robustness, transparency of procedure, replicability, absence of external disturbances, and many of the subjects mentioned by Cook and Campbell under the heading of internal validity (Cook & Campbell, 1979: 50-59).

Thus we arrive at the three broad sets of evaluation criteria included in the table: concept, instrument (or methodological requirements), and result. A final remark on these three sets concerns the possible suggestion of an order of appearance. Adopting the idea of reversible logic means accepting both forward (theory-derived) and backward (outcome-derived) mapping. Neither of these is subordinate, at least not in general. And in the course of reversible logic, methodological requirements must be met to make sure that reasoning itself is plausible. In short, we propose to view the three sets of criteria as equally important: each must be considered. In the table we have used asterisks to indicate that a given set of criteria may apply to a given stage in particular, but also that no set of criteria is irrelevant to some of the stages.

2.4 Concrete questions

It is conceivable that the argumentation given thus far does justify the table in terms of concept (offering a theoretically cogent picture of processes in a research project), but not in terms of instrument. How may justification proceed?

The gist of our argument is that ready-made prescriptions for validation do not obtain. We will nevertheless present a number of concrete justification issues, if only to demonstrate that the active approach to validity we advocate does offer support when validation attempts are actually made. In doing so, however, we have to acknowledge that validation issues often do not fit into a single cell in the table, but rather concern the relationship between adjoining cells. We will therefore not force justification issues in separate cells, but instead use the table as an instrument to elicit questions that are relevant in the context of justification (and in the context of discovery, for that matter).

A first series of questions relates mainly to the early stages of applying simulation gaming (problem formulation, creating or choosing a game design, assessment of likely processes):

1. Does the 'schematic' or conceptual model that will be used as a starting point for game design adequately represent existing knowledge and theoretical notions, and does it capture real-life circumstances?

2. Does the game design either include the subject (problem) to be dealt with in course of playing the simulation game, or does it allow for this subject to develop in the course of playing the game?
3. Are the number of variables (reduction), the level of abstraction, and the choice of symbols in the game design adequate, both in view of the conceptual model that is to be represented and in view of the subsequent stages of playing and the drawing of conclusions? (Peters et al., 1995)
4. Is there reason to expect inadvertent steering by the game design, e.g. through the account system used, the amount of time available in relation to the number of tasks, the nature of tasks and roles, characteristics of the initial situation, or boundary rules?
5. Is it likely that the game design allows the processes to be studied to develop in the application stage? Is it unlikely that less complex, less time-consuming, or less laborious (e.g. already existing instead of tailor-made) games would allow such processes to develop? Is it necessary that the game design copies some reference system in considerable detail?

A second series of questions relates to intermediate stages:

6. Is the conceptual model, and the derived game design, likely to produce events and processes in the application stage that can be productively referred to in debriefing? In particular, does the design allow for an acceptable degree of dynamics?
7. Does the simulation design fit the population of participants (and is such a fit necessary): is the design offered understandable, is it fair in the eyes of participants, is it engaging, is inadvertent steering of processes a likely result of combining the design with a particular population of participants?
8. Are participants properly selected and prepared, are positions and tasks assigned to them in a way that meets the game's objectives?
9. Does deliberate or inadvertent steering of processes occur in the application stage, by participants (experimental demand) or by the experimenter/facilitator (experimenter bias)? Have preventive measures been taken?
10. Are 'environmental' circumstances likely to affect the course of processes in the application stage in an undesirable way: e.g. learning effects, maturation, 'instrumentation' (Cook & Campbell, 1979: 52), sample mortality, demoralization, real-life connections between participants, 'obtrusive' observation, physical location?

A third series of questions relates to the final stages of using a simulation game:

11. Are, in retrospect, the events and processes that occurred in the simulation relevant with regard to the initial problem formulation? Have events, processes, or outcomes been observed or experienced that seem incompatible with the simulation game's objectives? How are these to be handled?
12. Is it possible to relate these events and processes back to some recognizable factors, by outside observers, and by participants themselves?
13. Do these factors provide clues for action?
14. Does either the nature of insights to be gained by participants in (or as a result from) the debriefing stage, or the way to help them arrive at such insights, make demands on the debriefer's own perception of (a) what went on in the simulation game, (b) how the simulation game relates to real-life processes, (c) how people may be enabled to transfer experiences from simulation game to real life?
15. If collective learning is aimed at, in addition to individual learning: have specific measures been taken to align the debriefing process to this objective? Have measures been taken to keep all participants aboard?

Our list - not pretended to be even roughly complete - may give an indication of the kind of questions a researcher has to consider in the course of a justification attempt. Many of the included questions refer to preceding as well as subsequent points of interest (that is, they have both a backward and a forward tendency). Or course, we formulated these question, but we did not in a deliberate attempt to make them unsuitable for classification in one of the cells. It is hardly possible, if possible at all, to make a statement on the quality of, say, some simulation design, without taking into account what it is supposed to reflect, and what it may give rise to.

2.5 Conclusion

We have defended the claim that a constructive approach to the question of validity is necessary and possible. When offering a line of argument to substantiate this claim, we emphasized simulation gaming as a research instrument, according to the idea that research can be considered an exemplary form of knowledge acquisition. That does not necessarily restrict the domain of application of constructive validation.

Any process of knowledge acquisition demands justification. That justification is not always the systematical attempt it is (or ought to be) in research does not disprove our contention. It may be that different forms of knowledge acquisition require different criteria for justification. It may be that different 'parties' involved in a research process will resort to different criteria. It may be that regarding justification a distinction must be made between the contexts of 'applied' and 'fundamental' research', that is, between a regulatory or policy cycle and an empirical cycle (Vissers et al., 1995). We suggest that such questions will be given due attention, since easy answers are suspect.

We offered a model (a table and its explication) that involved stages/activities and 'products' in a simulation gaming project, accepting beforehand that both stages/activities and 'products' may have to be refined, changed, replaced in order to meet the circumstances of a concrete project, be it research or another process of knowledge acquisition. We only pointed out that justification often cannot confine to separate cells. If anything, we hope that the argument presented will cause the awareness to sink in that a simulation game as it is played ('game-in-use') provides a far better basis for judgment than paper design ('game-in-the-box').

3 The Validity of Laboratory Research in Social and Behavioral Science[3]

Geert Vissers
Gerton Heyne
Vincent Peters
Jac Geurts

Abstract

The validity of artificial situations is often questioned, and particularly so the possibility of transfer of findings to the real world. Such questions, or doubts, may stem from a rigid distinction between real and artificial situations or from too strict a notion of representation. This article will argue that 'the real world' does not provide unambiguous criteria for representation and that, moreover, many experiments and simulation games do not have to represent 'the real world' in any direct way. Both issues are usually treated under the heading of external validity, which means compliance to conventions that dominated thinking about validity over decades. These conventions need to be reconsidered. Quality standards for research must not be rigid, nor should be applied in a way that ignores the characteristics of a particular research project. Fixed notions about validity may prevent a researcher from adapting validation procedures to the circumstances at hand. The article takes issue with a conception of external validity as surface resemblance between artificial and real situations, advocates an active, non-routine approach to validity questions, and encourages individual researchers to develop a line of reasoning on these questions instead of adhering to standards that may not suit their particular research.

Key words

Validity, external validity, simulation, gaming, laboratory experiment, experimental research

[3] Gepubliceerd als:
Vissers, G., Heyne, G., Peters, V., Geurts, J. (2001). In: *Quality & Quantity*, 35: 129–145.

3.1 Introduction

Of all aspects of the validity of artificial situations such as psychological experiments and simulation games, transfer to the real world is probably the one brought up most frequently. The question is whether, or to what extent, observations and experiences attained in artificial situations do apply to the real world. One who is engaged in designing and applying simulation games is likely to be asked frequently in what way and to what extent the behaviors and processes in a game (and the inferences made on that basis) can be applied to real-life situations. The literature on validity, despite its methodological depth, does not always contribute to adequate answering of this question, and worse than that, it may even suggest answers in an unsatisfactory direction.

Section 2 argues that present validity concepts and procedures need to be reconsidered. Three reasons are given, the first of which is that thinking about validity originates from psychology in the 1950s, broadly speaking, and that many current validity types can be traced back to that period, still reflecting conventions and objectives prevailing in psychological research at the time. Secondly, the concept of validity traditionally denotes the requirements to be fulfilled before the outcomes of scientific research are trustworthy. Because any type of research involves specific demands, numerous validity concepts and procedures have been proposed over the years. And even more concepts and procedures have appeared as a result of changing conceptions of 'good research'. The result is a plethora of validity types, whereas a line of reasoning that might guide decisions in designing and conducting research, or that might help justifying such decisions, is virtually absent. Thirdly, this accumulation of validity types is inconsistent and out of balance. Some aspects of research are the subject of elaborate validation procedures while other aspects are fully ignored. It is argued that researchers, rather than complying to the prevailing 'validity system', may treat validity as a context-dependent quality, that is, as a quality of the fit between the way instruments are used and conclusions are drawn in a specific research setting.

As discussed in section 3, this approach is particularly needed in relation to 'the' external validity of artificial settings. With respect to laboratory experiments and simulations games, the phrase 'lack of external validity' often seems little more than a way to express skepticism about the possibility to base conclusions that make sense in the real world on processes and behaviors observed (or experienced) in a situation that 'is not real'. Such skepticism rests on the supposition that a clear distinction can be made between real and artificial, it ignores the fact that created situations may differ in degree of representation, and it seems to assume that any experiment aims at direct transfer of findings to the real world.

In the final section the argument is broadened. Validation, rather than the application of an approved set of well-defined procedures, can be conceived of as a justification attempt stemming from the aspiration to demonstrate the soundness of assertions and claims. Whether the attempt is successful is to be judged by others, which means that validity is recognized again as part of a broad empirical cycle – and validation as a researcher's contribution to this. A notable implication is that validation and validity are not necessarily confined to the realm of scientific research.

3.2 Validity: Different Traditions, One Standard

3.2.1 Two traditions

In the 1950s, discussion about validity was mainly confined to psychological and educational testing (first tradition). Initially, these domains of research were absorbed by the prospect of tests that would, in due course, allow prediction of behaviors and achievements, in school, at work, or in the military. Accuracy of prediction was the first concern: "The behavior to be predicted was called criterion, and validation was all about the correlation between test and criterion: the degree of predictive validity. The contents of the test did not need to relate to the criterion" (Swanborn, 1981: 220). Soon this preoccupation with prediction weakened, however, due to stagnation in improving the quality of prediction. Attention shifted to theoretical and conceptual issues, which had a clear effect on the way validity was understood: "Now the question presents itself whether a measurement instrument does adequately reflect some general attribute from the renowned theories in a given field of research. The concepts 'content validity' and 'construct validity' make their entry" (Swanborn, 1981: 221).

Thus, validity was first conceptualized in direct support of operationally aimed test research, which was followed by validity concepts reflecting a more theoretical orientation. Various validity types emanated from each perspective. The resulting large number of types was soon viewed as a source of confusion, the more so because of a perceived lack of compatibility between validity types stemming from the two perspectives. The American Psychological Association (APA) intervened: "Prior to 1954 (...) the concept of validity was in considerable disarray. There were almost as many definitions and varieties of validity as there were people interested in psychometric theory. Something had to be done. Naturally a committee was formed (. . .)to look into the matter, and in 1954 the APA's Technical Recommendations for Psychological Tests and Diagnostic Techniques were published" (Ghiselli et al., 1981: 267).

The Technical Recommendations did not favor either perspective. Rather than choosing between quality of prediction and theoretical accuracy, the Committee endorsed four validity types: content, predictive, concurrent, and construct validity (APA Committee on Test Standards, 1954). Two of these (predictive, concurrent) derived from the operationalist perspective while the other two reflected a theoretical orientation. Although the concept of construct validity was not beyond dispute (Bechtoldt, 1959; Campbell, 1960), this has turned out to be the validity type to receive much attention after the Technical Recommendations were published (Cronbach and Meehl, 1955; Cook and Campbell, 1979; Embretson, 1983; Cronbach, 1989; Messick, 1995).

In due course, then, the concept of validity was adopted in other disciplines. Originating from the field of testing it became incorporated in research that utilized methods like questionnaire, observation, and experiment. Soon already, Brunswik coined the notion of 'ecological validity', arguing that experimental designs should

involve a representative sample of situations and of subjects (Brunswik, 1955). The issue of generalization was also addressed by Campbell's concept of 'external validity'. Together with its counterpart 'internal validity' this concept marks a second grand tradition in thinking about validity. In this second tradition, highlighting experimental and quasi-experimental research, the focus of attention changed in much the same way as in the earlier, test-oriented tradition: from instrument to concept. In a review of Campbell's work, Hammersley observes that validity used to pertain to instruments but has come to refer to assertions: "The first question to be asked is: what is it whose internal and external validity is to be assessed? In Campbell's 1957 article the answer is experimental designs. Much the same seems to be true in Campbell and Stanley. However, in Cook and Campbell 'validity' and 'invalidity' are introduced as terms referring to propositions, and thus presumably it is the conclusions drawn from experiments rather than the experimental designs or the experiments themselves whose internal and external validity are under scrutiny" (Hammersley, 1991).

Thus with respect to validity at least two large traditions can be distinguished, each encompassing operational or instrumental as well as more theoretical approaches. Over the years many validity types and concept have been added to the basic validity concepts developed in either tradition, often in an attempt to deal with specific validity questions. As it happened, all available types and concepts gradually became treated as if they were part of a single large collection of validity aspects, supposed to range simply from highly instrumental to highly theoretical.

This collection – which we refer to as 'the standard classification of validity' – has important drawbacks, the most important of which may be its lack of balance. Detailed validation procedures are available for some parts of the research process

– most notably for instrument design and data collection – but not for all. Striking omissions are the absence of guidelines for assessing the choice of instrument and setting and for the selection of data collection and data analysis methods. In the specific case of experimental research, further omissions relate to the selection and instruction of participants (Friedman, 1967) and evaluation or 'debriefing' (Peters et al., 1998). To be sure, these subjects are discussed in the methodological literature. The point is that extensive treatment of methodological considerations is too often followed by a concise list of questions to be answered for the purpose of establishing ‘the validity’. Note that this is a characteristic of the way ‘validity’ is handled, not of the basic idea of validity.

Textbook treatment of validity, in addition, tends not to refer to initial reasons for introducing a particular validity type, thus suggesting that different theoretical backgrounds can be ignored and that face value interpretation will do. Consider the example of face validity, a concept that returns in many recent texts on methodology, none of which does touch on Mosier’s (1947) distinction between validity by assumption, the appearance of validity, and validity by definition (Ebel, 1961; Nevo, 1985). In the absence of such specifications, face validity may seem little more than an euphonious word to disguise that private judgment is taken as a source of evidence.

Finally, contradiction, or at least lack of consistency, can be mentioned. Contradictions may stem from the fact that validity types derive from various research perspectives which, as indicated, reflect different theoretical positions. But contradictions can be found within a given perspective as well, as illustrated by the assumed inverse relationship between internal and external validity: “Generally, the higher the internal validity, the lower the external validity” (Swanborn, 1993: 214; see also Cook and Campbell, 1979: 89–90). Later in this paper it will be argued that there is reason not to speak of external validity in such a generalizing way. At present we only observe that, according to their originators, internal and external validity are concepts referring to requirements that cannot be fully reconciled.

In short, taken as a system, ‘the standard classification of validity’ is flawed. This observation may focus attention to the fact that the classification is indeed used as a standard. A standard may provide a basis for thought and a common language, which are useful qualities, but it may also serve as a substitute for thought. The abundance of validity types, often poorly defined, lack of methodical balance, and the presence of somewhat incompatible criteria make validation a demanding task. Confusion is the likely result of lack of clarity about separate validity types and about their interrelations. Researchers trying to evade such confusion by following textbook conventions are likely to detract from the merits of their own work, forcing it into the procrustean bed of generic validity requirements while at the same time having to accept that attention is directed away from important aspects of their particular research.

The solution advocated here is to consider validity as a relational attribute, that denotes the suitability of research instruments – their nature and the way they are used, which also includes the mix of instruments employed – in relation to the research questions to be answered. Researchers who engage in a validation attempt may first of all pay due attention the situational aspects of their subject of inquiry, rather than relying on validity types designed for context free measurement. Then, the old question ‘is this instrument valid’ is replaced by the question: ‘Is this instrument sufficiently valid to answer the research questions at hand’.

Such a change of perspective may not only apply to ‘the validity of instruments’, but also to ‘the validity of assertions’. According to the Joint Committee on Standards for Education Evaluation, validity is about “the soundness or trustworthiness of the inferences that are made from the results of the information gathering process” (Joint Committee, 1994: 145), a viewpoint that echoes Cook and Campbell’s emphasis on the validity of assertions (Hammersley, 1991). After listing several instruments and procedures for collecting information, the Committee goes on to declare that “Validation is the process of compiling evidence that supports the interpretations and uses of the data and information collected by using one or more of these instruments and procedures”. Thus, the Joint Committee adopts the broadened definition of validity that was issued in 1974 by APA’s National

Council on Standards for Educational and Psychological Tests (“Validity refers to the appropriateness of inferences from test scores or other forms of assessment” quoted by Ghiselli et al., 1981), in 1971 by Cronbach (“What needs to be valid is the meaning of interpretation of the score, as well as any implications for action that this meaning entails”, quoted by Messick, 1995), in 1965 by Crow and Noel (“Validity be measured by asking: how useful to the purpose for which it is to be gathered is the information produced by this method, as compared to some alternative method? ”, quoted by Raiser, 1969), and in 1949 already by Edgerton (“By ‘validity’ we refer to the extent to which the measuring device is useful for a given purpose”, quoted by Ebel, 1961).

Building on the view that validity is best considered a context-dependent quality that relates to instrument, interpretation, or assertion, the next section will concentrate on external validity and transfer. More specifically, it will examine the possibility of making inferences about the real world on the basis of processes that take place in an artificial setting.

3.3 On ‘the’ external validity of artificial situations

Generalization of experimental results was already a serious issue when Campbell and Stanley (1963) introduced the term external validity, but seems to have gained importance since, even to the extent that, presently, negative connotations like ‘not real’, ‘not valid’, or even ‘forged’ must be ignored in a decision to engage in experimentation. This section, in an attempt to redress some of the balance, will defend that simple claims on either the external validity or invalidity of artificial settings like psychological experiments and simulation games are suspect. Direct transfer of findings from experiment to the real world is not always aimed at, and when such transfer is aimed at, the line between artificial and real situation may not be clear, because of various degrees of 'artificiality' and because real-life situations are all but unequivocal.

3.3.1 The need for direct transfer

It is an often overlooked fact that direct transfer of findings to the real world is not always aimed at. Mook (1983) argues that artificial settings may be used for exploring or testing theoretical notions rather than for making direct predictions about real-life behaviors. He lists four alternatives to what he calls the 'analogue' model of research: "First, we may be asking whether something can happen, rather than whether is typically does happen. Second, our prediction may be in the other direction; it may specify something that ought to happen in the lab, and so we go to the lab to see whether it does. Third, we may demonstrate the power of a phenomenon by showing that it happens even under unnatural conditions that ought to preclude it. Finally, we may use the lab to produce conditions that have no counterpart in real life at all, so that the concept of 'generalization to the real world' has no meaning".

Even if generalization of results is aimed for it may not be necessary, Mook adds, that the laboratory setting resembles a real-life setting as much as possible. Referring to Milgram's experiments on obedience to authority he argues that "there are cases in which the generalization from research setting to real-life settings is made all the stronger by the lack of resemblance between the two".

Mook's argument that resemblance to real life is not always necessary can be extended. Berkowitz and Donnerstein (1982) point out that such resemblance, if already deemed necessary, is not a subject that allows for clear distinctions and general evaluation criteria: "The meaning the subjects assign to the situation they are in and the behavior they are carrying out plays a greater part in determining the generalizability of an experiment's outcome than does the sample's demographic representativeness or the setting's surface realism".

Thus, correspondence between artificial setting and some specific real-life reference situation is not necessarily a validity criterion, and when it is, the kind of correspondence required cannot be assumed to be self-evident.

3.3.2 Degrees of artificiality and the equivocality of real-life situations

If generalization from the lab to the real world is aimed at, the question of degree of correspondence between experiment or simulation game and real-life reference system will have to be answered. The often repeated criticism is that so many real-life elements are not and cannot be represented 'in the laboratory'; that findings derived from it cannot but give a distorted picture of 'real life' situations, behaviors, and processes. This criticism rests on two assumptions: (1) that unequivocal descriptions can be given of ‘reality’, and (2) that a clear boundary can be drawn between real life and the laboratory. Both assumptions are disputable.

Aronson et al. (1985: 443–444) take issue with the idea of a clear boundary between ‘the field’ and ‘the lab’. They give an example that starts from a simple event: “Suppose that you are a young man walking along a street in New York and that a rather attractive young woman carrying an armload of books and papers approaches you. Just as you and she come to within ten steps of each other, she stumbles slightly, dropping her books and scattering her papers. Unknown to you, a social psychologist, sitting in a car parked at the curb, is observing whether or not you stop to help the woman retrieve her books and papers, how long you stay at the task, and so on, as a function of the physical attractiveness of the woman”.

Did the psychologist prepare the situation, asking an attractive young female student to stumble and drop books and papers in front of young men in a New York street? Probably so, but not necessarily. Suppose that you are a psychologist studying the impact of physical attractiveness on human responses. You sit in your car parked at the curb, studying the responses of those passing by a disabled

looking beggar (who is one of your students). All of a sudden you see a rather attractive young woman carrying an armload of books ....”

Does the artificial nature of a situation fully depend on the intentions of the observer? And if so, on which observer’s intentions? It makes a difference, of course, if an observer made preparations, e.g., if a psychology teacher asks students to participate in an experiment. But what if you are a psychologist, observing responses to beggars, and suddenly you see a young attractive woman stumbling, and so forth, without knowing that she is a student of one of your colleagues?

Aronson and his colleagues do not discuss the range of possibilities between fully unplanned and prepared stumbling. instead, they offer a number of increasingly ‘controlled’ situations:

- Suppose that you are walking along a street on the campus of Columbia University in New York and that a rather attractive young woman carrying an armload of books and papers, is walking to you. Just as you and she ....
- Suppose that you are a student at Columbia University and you have volunteered to participate in an experiment (...). You enter the psychology building at the appointed time and as you are walking down the corridor, an attractive young woman is walking toward you carrying an armload ....
- Suppose, after signing up for an experiment in psychology, you arrive at the psychology department and are told to wait in a room for the experimenter until another student

> who will also be a subject in the experiment arrives. A few minutes later an attractive young woman enters; she is carrying an armload ....

This sequence of situations (that is not even complete) clearly demonstrates that ‘degrees of preparedness’ are possible. It is difficult to make a sharp distinction between ‘the field’ and ‘the lab’. That is also the conclusion to be drawn from examining the practice to treat field experiment and quasi-experiment as distinct research methods. According to Cook and Campbell (1979) ‘passive observational approaches’ (referred to as a class of ‘non-experimental’ methods) are “methods (that) try to infer causal processes based on observations of concomitances and sequences as they occur in natural settings, without the advantage of deliberate manipulation and controls to rule out extraneous causal influences”. Implied in this definition is that a passive observational approach turns into a form of experiment as soon as the researcher makes a deliberate attempt to influence the course of events in a natural setting. Also implied in the definition, and even more important, is the assumption that a researcher may choose not to manipulate and control – in other words: not to influence the course of events in a research setting.

‘Investigator effects’ have long been recognized in behavioral and social science, at least since Mayo and his colleagues conducted the studies known as the Hawthorne experiments (Phillips, 1976). Started as a study on the impact of physical environment on team output (Schulz and Schulz, 1994), unanticipated results produced by a less than perfect experimental design enabled the researchers to find that a situation is changed by an observer’s presence. This finding has in spired organization theorists and psychologists to explore the role of social and psychological factors in the workplace, and later the role of social interaction, expectations, and people’s needs in a variety of social settings. Phillips (1976) links the Hawthorne studies to Rosenthal (1966), who showed that experimenters may systematically influence their ‘subjects’ through personality characteristics and techniques. Such systematical influence need not be deliberate. Rosenthal and Jacobson (1968) emphasized that simple perceptual clues (pupils’ test results) offered to classroom teachers sufficed to produce significant gains in an intelligence test after 8 months, whereas such gains were not deliberately pursued (Weick, 1995).

Thus taken, the Hawthorne effect’ is not a methodological flaw or bias (as too often asserted) but an aspect of interpersonal relationships, which is a category that includes most social and behavioral research, field research and experimental research alike. Research is not flawed by the Hawthorne effect itself – or by related phenomena like ‘experimental demand’ (Orne, 1962; Orne and Evans, 1965) or ‘evaluation apprehension’ (Rosenberg, 1969) – but by the failure to take such phenomena into account when findings are reported. For present purposes, the important point is that it will be difficult for a researcher to make simple and unambiguous statements about the presence of absence of manipulation and control in a research setting. This means that the distinction between experimental and non-experimental methods, in Cook and Campbell's terms, is vague and uncertain, and possibly unstable over time.

We may conclude that words like 'experiment' and 'natural setting' reflect convention rather than intrinsic qualities of a research setting. Because of the strong connotations of words like 'real', 'natural', and 'artificial', however, a dichotomy phrased in these terms may well convey the message that such intrinsic qualities are present.

Let us now turn attention to the alleged unequivocally of real-life situations. The above argument does not eliminate the possibility that the distinction between real and artificial is only seemingly gradual because of the variety of experimental settings, ranging from Sidowski's 'minimal social situation' (Weick, 1979) or Heap's 'minimum-structure simulation game' (Heap, 1971) to very detailed representations of real-life reference situations. In other words: There is only one reality, one could argue, but there are many forms of representation. While the second part of the argument is easy to accept, the first must be rejected. Studies in the sociology of science and technology show that scientific knowledge does not simply reflect 'nature' or 'reality'. In a formulation by Barnes, "theories are imposed upon reality rather than deriving from it" (Mulkay, 1980: 69). This applies not only to theory: "Even the level of 'fact' – of experiment and observation – is social, and different groups of scientists in different circumstances have been shown to have produced radically different 'facts' " (MacKenzie and Wajcman, 1985: 8). Another example is offered by Dolin and Susskind (1992: 37) who, when preparing the National Energy Policy Simulation, observed that between conflicting groups "there was

sharp disagreement over the price that a barrel of oil would have to reach before the domestic economy would begin to feel the effect".

Such disagreement cannot be waved aside as an incident. It is consistent with the observation that perception is theory-laden (Hanson, 1972). In empirical research, many factors have been shown to produce more or less systematic differences in perception, e.g., professional group (Strauss et al., 1963), organizational position (Zajonc and Wolfe, 1966; Gregory, 1983), social background (Lupton and Cunnison, 1964), or type of position in public administration (Lipsky, 1980). These broad factors cannot fully explain perceptual differences. For that, at least 'factors' like culture (Swidler, 1986; Young, 1989) and past developments would have to be added.

For the present argument, perceptual differences do not have to be explained. It suffices to observe that more or less systematical differences do exist in perceived 'reality'. The implication is that answers to the question of similarity (or isomorphism) between artificial setting and real life will differ between people. Answers to this question will depend on respondents' perceptions of the real-life reference situation (even to the extent that people disagree about what real-life situation is represented in the laboratory). Answers will also depend on respondents' perceptions of what happened in the laboratory, the recognition of which has brought Berkowitz and Donnerstein (1982: 249) to the assertion that "the meaning the subjects assign to the situation they are in and the behavior they are carrying out plays a greater part in determining the generalizability of an experiment's outcome than does the sample's demographic representativeness or the setting's surface realism".

Thus, when generalization is aimed at, the question is whose perceptions are chosen to define the real-life situation that will serve as the criterion for 'goodness of fit' of the artificial research setting, but also whose perceptions are chosen to define the nature of events in the experiment.

3.4 Towards Non-Routine Validation

Above it was argued that prevailing validation conventions do not meet the requirements of many a concrete research project and that, in the particular case of research using artificial settings, the prospects of generalization are not simply proportionate to degree of direct representation. Whether external validity requires direct representation (of elements in the reference system) depends on the objectives of the research being conducted. It is a researcher's responsibility to justify the extent to which findings from the lab apply to the real world. (Of course, it is a joint responsibility of the members of the larger scientific community to scrutinize the arguments given.)

Making the case for researchers developing a line of reasoning in defense of the research that they are conducting, we will make two suggestions that may support this task. The first, addressing (deliberately) artificial situations, elaborates the idea that theory rather than direct representation is a proper basis for generalization.

The argument may have wider (more general) significance, however, since this problem of generalization also presents itself in field research, e.g., when a particular 'unit' is studied for the purpose of making more general claims. Although artificial situations are the present paper's main focus, a second suggestion applies to any type of research. Conclusions must rest on solid data, which requires sound data collection procedures. Andsoback. The obvious implication is that validation includes a strong element of backward reasoning.

3.4.1 Theories as vehicles for generalization

As mentioned, it is often supposed that generalization means direct representation. For research using an artificial setting this (overt or tacit) supposition implies that a given real-life reference system should be reflected as much as possible, that is, in all cases that generalization is aimed at. Therefore, in such cases 'isomorphic' models can be preferred over 'metaphoric' models. Although a high degree of representativeness can be useful, there is reason to temper the quest for representative models. Bass and Firestone (1980) argue that generalizability cannot be equated with representativeness, and representativeness does not only, or even primarily, concern physical and demographical characteristics: "Simply knowing that certain field studies (or laboratory studies) have been conducted primarily or exclusively on certain types of subjects implies nothing whatever about the possible generalizability of research findings from these studies, unless one has some theoretical rationale or empirical basis for expecting similar or different relationships as a function of setting or subject population". Without such a theoretical or empirical basis it is impossible to decide "what are and what are not relevant variables that must be accounted for in extrapolating results". This position is in agreement with Schlenker and Bonoma's (1978: 27) claim that "Theories (...) are the vehicles which allow generalization to the real world. No one experiment and no series of data can be generalized directly to anything. Only theoretical propositions can allow generalization".

Regarding the subject of representation, Bass and Firestone (1980) observe a 'physical' bias. They mention three basic categories that pertain to generalizability of research results: setting, person, and response. Setting aspects include "not only objective stimulus components such as job descriptions but also subjective task definitions". Person aspects involve demographic factors "but also ability and personality variables of individuals as well as the mix of these variables as they may obtain in functional work groups". Within the category of response aspects a distinction is proposed between cognition-judgment and action-performance. The distinction made does clearly show that (theoretically, practically) relevant phenomena are readily neglected if generalization is automatically associated with demographic attributes or with the more 'tangible' elements of a research setting.

The recognition that 'setting' is not the only, and not always the most critical part in transferring experimental results to the real world has been given a

surprisingly lucid interpretation in the field of simulation and gaming. Having noticed that the term 'simulation game' is often used to refer merely to the set of instruction books and other materials necessary for play, Geurts and Van Wierst (1991) propose a definition that stresses playing itself ('game-in-use') instead of setting elements ('game-in-the-box'). Obviously, such a change of definition has implications for validation. Even when the 'objective stimulus components' of an artificial setting seem to reflect elements from a real-life reference system, it is the way these components operate when used by actors (both in the artificial and in the real-life setting) rather than 'paper resemblance' that needs to be emphasized in a process of (external) validation.

3.4.2 Backward reasoning

When it is possible to make a distinction between 'game-in-the-box’ and ‘game-in-use’, the question presents itself how these two relate. How strongly does ‘game-in-the-box’ influence game-in-use’? Is this influence itself dependent of circumstances? Which other factors codetermine? These questions focus attention to the connection between successive stages in a research process, a connection that is important since the processes and outcomes of ‘game in use’ (to refer to the above example) have to meet research objectives, and must be judged in that light. Therefore, the constituting stages in a research project should not only be judged on their own merits; also important is that the fit between stages is examined. Projected outcomes can be used as a starting-point for validation through backward reasoning.

This remark applies to research in a strict sense but also to other activities using an artificial setting. For example, if a simulation game has to serve specific learning objectives (Peters et al., 1998), a sequence of ‘backward questions’ might start with debriefing objectives: What learning processes should be the result of debriefing? For these processes to happen, how does the discussion in the final debriefing session proceed best? For a given type of debriefing discussion, what precautions have to be taken (duration, group size, debating methods, type of discussion leadership, and so forth)? What requirements must be met in the preceding stage? Which design characteristics are helpful or even necessary for these requirements to be met? Such a sequence of backwards oriented questions may offer a framework for justification.

Backward reasoning is a useful heuristic for structuring the task of making validity claims and selecting materials for that (Vissers et al., 1998). If we concentrate on research, such backward reasoning will have to start from research objectives, which presupposes that objectives are specific enough to allow the formulation of criteria that can be used for observing the fit between objectives and research outcomes. To make objectives specific often means: to make them explicit. Research starts from expectations, though sometimes largely implicit ones. It is always possible to make implicit expectations explicit, to refine them if they are too

general, and to rephrase them in a way that testable suppositions are obtained. It is not possible to offer a general procedure that could help researchers to arrive at test-able suppositions. One reason is that the variety of research objectives (and related methods) is large, a second that much depends on the kind of theory that guides a particular research project. Research objectives vary for a number of reasons, as becomes particularly clear when ‘objectives’ is rephrased as ‘aspired products’. Then, the question is: What is the ‘product’ this particular project is supposed to deliver? Here the word ‘product’ does not refer primarily to the medium used for information transfer (book, paper, presentation) but to the type of information to be transferred.

In this respect, a pertinent question relates to the kind of audience being addressed. Here, a relevant distinction is between peer scientists and other addressees (within the latter category further distinctions can be made, e.g., between ‘experienced consumers of research’ and others). Further important questions are: Is the product meant to be directly applicable (e.g., advice, systems design, blueprint, research instrument, training schedule), or should it contribute to the audience’s knowledge base (analysis, assessment, forecast), and is the product aiming at developing courses of action within a prevailing paradigm (‘improvement’) or at exploring the prospects of a new paradigm (‘change’).

Such questions, though still quite general, may help to specify criteria for evaluation of the product delivered, which includes evaluation of the research steps needed to arrive at this product. To this conclusion, we add two further remarks. One is that we have used words like ‘testable suppositions’ and ‘criteria for evaluation’, thus avoiding to speak of ‘hypotheses’. Apart from the fact that the word ‘hypotheses’ has strong scientific connotations (which might contribute to overlooking ‘other audiences’), it would be misleading to suggest that operational objectives are only attainable in the case of well-developed theory. Objectives may range from a first attempt to specify criteria that research outcomes have to meet to very elaborate criteria that include exact measures for ‘goodness of fit’.

A second remark is that research objectives may not be stable. In research, as in any course of purposeful action, initial intentions and expectations tend to change during the process. New insights emerge, often as a result of unanticipated responses by those being investigated. Such a change of objectives is legitimate, provided that the initial objectives and the reasons for change are described in detail, so that other researchers are able to review the full line of argument.

3.5 Conclusion

Over the years the basic idea of validity has been almost snowed under with definitions, types, standards, and routines. The present article aimed at revitalizing the basic idea. Validation, then, returns to be a challenging aspect of research, or knowledge acquisition in general, instead of being an obligatory task that remains to be done after the interesting part of the work is finished. We have made the case

for relating validity to assertions instead of understanding it as merely a property of a research instrument. Instead of asking ‘is this instrument valid?’ the question ‘Is this instrument sufficiently valid for this specific purpose?’ seems far more relevant.

Attention was focused on validity in relation to artificial situations like experiments and simulation games. Addressing this relation leads almost inevitably to the question of external validity.

We concurred with Douglas Mook that not any experiment or any simulation game needs to have external validity. Transfer from experimental situation to real life is not always an objective, that is, when transfer is taken to mean that findings can be directly applied to a real-world reference situation. But when direct transfer of findings is aimed at, one cannot ignore the blurred distinction between artificial and real. 'The real world' cannot be used as a simple criterion for isomorphism, because no clear boundary can be drawn between 'real' and 'experimental', or between 'real' and 'artificial', and because, as a rule, perceptions and judgments of 'the real world' will not be shared.

We argued that attempts to assess the validity of artificial situations tend to pay undue attention to superficial setting and person characteristics and, at least with regard to simulation games, too little attention to 'game-in-use'. It is important, both in the context of design and in the context of justification, to acknowledge the difference between 'paper' model and operating model: Evoking a particular response is not a quality of design, but of the interplay of design and experimental subjects or players.

In short, we questioned various rules of thought that seem to dominate current validation practices. These rules, we contend, are precluding rather than fostering validation.

Whether undertaken during the research process or afterwards, validation is not the activity of determining one or more aspects of validity. Rather, it is all that results from the aspiration to design and justify the steps in a (research) project. The word 'aspiration' emphasizes that only provisional claims can be made with respect to validity, be it of a psychological experiment, a simulation game, some other method, or of assertions based on any such method. Validity is context-bound and its realization is uncertain. One can never be sure that objectives are fully realized, requirements fully met, and all this in the most efficient way.

This means that validity, or validation attempts, may not gain much by adding further procedures to 'the standard classification', or by refining existing ones. In our view, research is (and researchers are) more likely to benefit from conceiving validation as a line of reasoning that supports the design, or selection, and the use of research instruments.

We have made an effort to specify some elements of such a line of reasoning. With regard to external validity of artificial settings we stressed the importance of a 'theoretical detour', and we argued that 'operational' rather than 'paper' models should be taken as a starting point for comparing 'artificial' and 'real-life' setting.

To these arguments we add here that there is no reason to adopt a defensive stance towards the critique that 'artificial situations' lack external validity. If, in a particular case, the external validity of an experiment or simulation game is questioned, we have a subject of serious discussion, not a critique that includes the privilege of reversing the onus of proof. Finally we made the case for backward reasoning. Any research project has objectives (whether or not fully elaborated and explicated) that can be used to reason backward, asking whether a stage in the process is properly prepared by the preceding stages. Again this is not a question of 'process logic' that can be answered without reckoning with the audience being addressed or the kind of 'product' to be delivered.

However, 'backward reasoning' is only one part of the story. Previously we argued that in a research project it is always possible to identify objectives by making implicit expectations explicit. But we warned that such objectives are not fixed, since researchers may learn from the research they are conducting, which may cause initial expectations (and derived objectives) to change. Research, then, comes to be seen as an iterative rather than a goal-directed process: Expectations guide objectives, objectives guide research decisions, research decisions guide what you observe and learn, and this, in turn, guides expectations. The consequence is that research cannot be fully planned in advance; many decisions will have to be taken in the course of a research process. Elmore (1985: 35), considering the case of policy analysis, makes a similar observation. Forward and backward reasoning are mutually influencing, he argues, which results in 'reversible logic': "The logic is essentially this: To get from a starting point to a result, we don't just set an objective and go there. We begin at either end and reason both ways, back and forth, until we discover a satisfactory connection". Now, researchers may benefit from realizing that back and forth reasoning is an inevitable and therefore acceptable part of any process of knowledge acquisition, not a sign of poor problem definition that has to be suppressed in a final report. This realization, if it keeps researchers from complying with the rule that 'validated instruments' be used, will enhance the quality of research. It will make room for considering the prospects of various methods and instruments, which also means that due attention is paid to the fit between successive stages in a research project. All this may restore the sense of reasoned curiosity that is beneficial for good research, and it may encourage an open discussion about assumptions and methods, a discussion that will sometimes include the suitability of artificial settings.

4 Referenties

- Albinski, M. (1978). *Onderzoek en Aktie*. Assen, Van Gorcum.
- APA Committee on Test Standards (1954). Technical recommendations. Psychological Bulletin 51: 200–238.
- Aronson, F., Brewer, M., Carlsmith, J.M. (1985). Experimentation in social psychology. In: Lindzey, G., Aronson, E. (eds), *Handbook of Social Psychology, I*. New York: Random House, pp. 441–486.
- Bass, A.R., Firestone, I.J. (1980). Implications of representativeness for generalizability of field and laboratory research findings. *American Psychologist* 35: 463–464.
- Bechtoldt, H.P. (1959). Construct validity: A critique. *American Psychologist* 14: 619–629.
- Berkowitz, L., Donnerstein, E. (1982). External validity is more than skin deep: Some answers to criticisms of laboratory experiments. *American Psychologist* 37: 245–257.
- Bougon, M.G. (1992). Congregate cognitive maps: A unified dynamic theory of organization and strategy. *Journal of Management Studies* 29: 369–389.
- Brunswik, E. (1955). Representative Design and Probabilistic Theory in a Functional Psychology. *Psychological Review,* 62, 193-217.
- Campbell, D.T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin* 54: 297–312.
- Campbell, D.T. (1960). Recommendations for APA tests standards regarding construct, trait, or discriminant validity. *American Psychologist* 15: 536–553.
- Campbell, D.T., Stanley, J.C. (1963). Experimental and quasi-experimental designs for research and teaching. In: Gage, N.L. (ed.), *Handbook of Research on Teaching*. Chicago: Rand McNally, pp. 171–246.
- Cook, Th., Campbell, D.T. (1979). Quasi experimentation. Design analysis issues for field settings. Chicago: Rand McNally.
- Cronbach, L.J. (1989). Construct validation after thirty years. In: Linn, R.L. (ed.), *Intelligence: Measurement, Theory, and Public Policy.* Chicago: University of Illinois Press, pp. 147–471.
- Cronbach, L.J,. Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin* 52: 281–302.

- Dolin, E.J., Susskind, L.E. (1992). A role for simulations in public policy disputes: The case of national energy policy. *Simulation & Gaming* 23: 20–44.
- Duke, R. (1974). *Gaming: the future's language.* New York: Sage Publications.
- Ebel, R.L. (1961). Must All Tests Be Valid? *American Psychologist*, 16, 640 647.
- Elmore, R.F. (1985). Forward and Backward Mapping: Reversible Logic in the Analysis of Public Policy. In K. Hanf, K., Toonen, Th. (eds.) *Policy Implementation in Federal and Unitary Systems.* Dordrecht, Martinus Nijhoff.
- Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin* 93: 179–197.
- Friedman, N. (1967). *The Social Nature of Psychological Research*. New York: Basic Books.
- Geurts, J., Wierst, P. van (1991). Spelsimulatie: oefenen met complexiteit. [Gaming: exercising complexity]. In: *Spelsimulatie in managementopleidingen* [Gaming in management training]. Deventer: Kluwer, 1-16.
- Ghiselli, E.E., Campbell, J.P., Zedeck, S. (1981). *Measurement Theory for the Behavioral Sciences.* San Francisco: W. H. Freeman.
- Greenblat, C., Duke R. (1975). *Principles and practices of gaming/simulation.* Beverly Hills: Sage Publications.
- Gregory, K.L. (1983). Native-view paradigms: multiple cultures and culture conflicts in organizations. *Administrative Science Quarterly* 28: 359–376.
- Guba, E.G., Lincoln Y.S. (1982). Epistemological and methodological bases of naturalistic inquiry. *Educational Communication and Technology Journal*, 30, 233-252.
- Hammersley, M. (1991). A Note on Campbell's Distinction between Internal and External Validity. *Quality and Quantity*, 25, 381-87.
- Hanson, R.N. (1972). *Patterns of Discovery.* Cambridge: Cambridge University Press.
- Heap, J.L. (1971). The student as resource, uses of the minimum-structure simulation game in teaching. *Simulation & Games* 2: 473–487
- Henshel, R.L. (1980). The purposes of laboratory experimentation and the virtues of deliberate artificiality. *Journal of experimental Social Psychology* 16: 466–478.
- Joint Committee on Standards for Educational Evaluation (1994). The Program Evaluation Standards: How to Assess Evaluations of Educational Programs. Thousand Oaks, Sage.
- LeCompte, M.D., Goetz J.P. (1982). Problems of reliability and validity in ethnographic research. Review of educational research, 52, 31-60.

- Lederman, L. (1992). Debriefing: toward a systematic assessment of theory and practice. *Simulation & Gaming: An international Journal of Theory, Practice, and Research*, 23 (2), 145-160.
- Lipsky, M. (1980). Street-Level Bureaucracy: Dilemmas of the Individual in Public Services. New York: Russell Sage Foundation.
- Lupton, T., Cunnison, S. (1964). Workshop behavior. In: Gluckman, M. (ed.), *Closed Systems and Open Minds: The Limits of Naivety in Social Anthropology*. Chicago: Aldine, pp. 103–128.
- MacKenzie, D., Wajcman, J. (1985). Introductory essay. In: MacKenzie, D., Wajcman, J. (eds), *The Social Shaping of Technology*. Milton Keynes, UK: Open University Press.
- Messick, S. (1995). Validity of psychological assessment. Validation of inferences from persons' responses and performances as scientific inquiry into score mean. *American Psychologist* 50: 741–749.
- Mook, D.G. (1983). In defense of external invalidity. *American Psychologist* 38: 379–387.
- Mulkay, M. (1980). Sociology of science in the West. *Current Sociology* 28, 3, Trend Report: The sociology of science in East and West: 1–184.
- Nevo, B. (1985). Face validity revisited. *Journal of Educational Measurement* 22: 287–293.
- Orne, M. (1962). On the social psychology of the psychological experiment. *American Psychologist* 17, 776–783.
- Orne, M.T., Evans, F.J. (1965). Social control in the psychological experiment: Antisocial behavior and hypnosis. *Journal of Personality and Social Psychology* 1: 189–200.
- Peters, V., Vissers, G., Heyne, G. (1998). The validity of games. *Simulation & Gaming* 29: 20–30.
- Peters, V., Vissers, G., Meer, F.B. van der (1998). Debriefing depends on purpose. In: Geurts, J., Joldersma, C., Roelofs, E. (eds), *Gaming/Simulation for Policy Development and Organizational Change*. Tilburg: Tilburg University Press, pp. 399–404.
- Phillips, B.S. (1976). *Social Research, Strategy and Tactics*, 3rd edn. New York: MacMillan.
- Raser, J.R. (1969). Simulation and Society: An Exploration of Scientific Gaming. Boston, Allyn and Bacon.
- Rosenberg, M.J. (1969). The conditions and consequences of evaluation apprehension. In: R. Rosenthal, R., Rosnow, R.L. (eds), *Artifact in Behavioral Research.* New York: Academic Press, pp. 279–349.

- Rosenthal, R., Jacobson, L. (1968). Pygmalion in the Classroom, Teacher Expectation and Pupils' Intellectual Development. New York: Holt, Rinehart & Winston.
- Rosenthal, R. (1966). Experimenter Effects in Behavioral Research. New York: Appleton.
- Schlenker, B.R., Bonoma, T.V. (1978). 'Fun and games': The validity of games for the study of conflict. *Journal of Conflict Resolution* 22: 7–38.
- Schulz, D.P., Schulz, S.E. (1994). *Psychology and Work Today.* Upper Saddle River, NJ: Prentice-Hall.
- Strauss, A., Bucher, R., Ehrlich, D., Sabshin, M., Schatzmann, L. (1963). The hospital and its negotiated order. In: Friedson, E. (ed.), *The Hospital in Modern Society.* New York: Macmillan, pp. 147–169
- Swanborn, P.G. (1981). Methoden van sociaalwetenschappelijk onderzoek. Meppel: Boom.
- Swanborn, P.G. (1993). External validity abandoned? *Quality & Quantity* 27: 211–215.
- Swidler, A. (1986). Culture in action: Symbols and strategies. *American Sociological Review* 51: 273–286.
- Vissers, G. (1994). *The Production of Strategy.* Delft: Eburon.
- Vissers, G., Heijne, G., Peters V. (1995). Spelsimulatie en bestuurskundig onderzoek. [Gaming and research on public administration]. *Bestuurskunde*, 4 (4), 178-187.
- Vissers, G., Peters, V., Heyne, G., Geurts, J. (1998). Validity of simulation games: A constructive view. In: Geurts, J., Joldersma, C., Roelofs, E., (eds), *Gaming/Simulation for Policy Development and Organizational Change.* Tilburg: Tilburg University Press, pp. 353–359.
- Weick, K.E. (1979). *The Social Psychology of Organizing,* 2nd edn. New York: Random House.
- Weick, K.E. (1995). *Sensemaking in Organizations.* Thousand Oaks, CA: Sage.
- Young, E. (1989). On the naming of the rose: Interests and multiple meanings as elements of organizational culture. *Organization Studies* 10: 187–206.
- Zajonc, R.B., Wolfe, D.M. (1966). Cognitive consequences of a person's position in a formal organization. *Human Relations* 19: 139–150.

Colofon

Drie artikelen over
De validiteit van spelsimulaties

Samenspraak Advies Nijmegen

Postbus 31006
6503 CA Nijmegen

telefoon 024 3555662
email info@samenspraakadvies.nl
website www.samenspraakadvies.nl