Psychology Science, Volume 49, 2007 (2), p. 116-131

Optimizing technical precision of measurement in computerized psychological assessment on Windows platforms

JOACHIM HÄUSLER[1], MARKUS SOMMER & STEFAN CHROUST

Abstract

Reaction times and response latencies are required to measure a variety of ability and personality traits. If reaction times are used to measure rather elementary cognitive tasks, the inter-individual variance in the measured reaction times are usually small in the sense that the central 50 percent of a norm population range within less than 100ms. Technical measurement errors therefore have the potential to seriously affect the validity of diagnostic judgments based on such measures. Thus the target of this paper is to investigate the magnitude of possible errors of measurement due to technical reasons and to suggest ways to prevent or at least consider those in the diagnostic process.

In Study I a highly precise 'artificial respondent' was applied to simulate reactions corresponding to a given percentile rank on 3 different tests (DG-Lokation CORPORAL, Alertness TAP-M, RT/S9 Vienna Test System) on 11 different computer systems. The result output of the tests was compared to the reaction times, actually provided by the artificial respondent. Results show, that there are detectable errors of measurement - depending on the hardware and software specifications of the computer system used. In the test DG-Lokation these bias caused an offset in the tests main variable of up to 20 percentile ranks.

In Study II a self-calibration unit which is part of the Vienna Test System (Version 6.40) was investigated, using the same experimental setup. After calibration, the bias detected can be reduced to the magnitude of about 1 percentile rank on all computer systems tested.

It thus can be concluded, that time critical computer based tests typically bear the risk of technical errors of measurement. Depending on how the test is programmed, the errors arising on some computer configurations can cause even severe changes in diagnostic judgment formation. In contrast, self-calibration proved to be an effective tool to permitting the user not only to control but also to ensure the precision of measurement, independent of the properties of the computer system he is administering his test on.

Key words: computerised testing, technical precision of measurement, self-calibration, reaction tests

[1] Mag. Joachim Häusler, Hyrtlstraße 45, 2340 Mödling, Austria; email: haeusler@schuhfried.at

Introduction

One of the major advantages of computerised psychological testing resides in the opportunities it provides for measuring ability and personality traits that cannot be assessed using traditional paper-and-pencil tests (cf. Klinck, 2006; Kubinger, 2006; Wagner-Menghin, 2003). For example, computerised measurements can be taken of the reaction times involved in various aspects of attention (eg. Perception & Attention Functions (WAF: Sturm. 2006)), executive functions or implicit self-concepts (using Implicit Association Tests IAT: Greenwald, McGhee & Schwarz, 1998).

As Table 1 makes clear, these measurement paradigms often provide very small inter-individual variance though. It is therefore crucial that time measurements are accurate at millisecond level.

In order to accomplish this goal computerised test systems have to enable an extremely precise measurement of response latencies. Even technical errors of measurement of the order of hundredths of a second have the potential to cause a significant shift in diagnostic judgments that are based entirely, or at least in parts, on the results of computerised tests. This raises the question of whether technical errors of measurement of relevant size can occur on modern computer systems. It is often argued that, within certain limits, increasingly powerful computer systems enable increasingly precise measurement of reaction times and response latencies (cf. MacInnes & Taylor, 2001; Forster & Forster, 2003). However, others have expressed concern that the more complex graphical user interface of Microsoft Win-

Table 1:
Width of the normal range (PR 25 - PR 75) for various reaction-time-based paradigms. The more complex the cognitive ability required by the test, the wider the inter-individual variance. Where very basal performance dimensions are concerned, measurement errors therefore have a particularly marked effect.

Dimension	Test	Width of the normal range	Source
Alertness	TAP-M	66 ms	Zimmermann & Fimm, 2005
Alertness (intrinsic visual)	WAFA	76 ms	Sturm, 2006
DG - Lokation*	CORPORAL	93 ms	---
Divided attention	TAP-M	226 ms	Zimmermann & Fimm, 2005
Divided attention (unimodal visual)	WAFG	200 ms	Sturm, 2006
Selective attention (unimodal visual)	WAFS	149 ms	Sturm, 2006
Stroop effect (reading interference)	STROOP	180 ms	Puhr & Wagner-Menghin, 1993
Reaction ability (visual stimuli)	RT	162 ms	Schuhfried & Prieler, 1994

*The norm tables of the CORPORAL test system are not given in the test manual; the details given are taken directly from the test's norm files.

dows operating systems makes direct control of stimulus display and reaction input more difficult. This presents a potential risk for the precision of measurement of reaction times and response latencies (cf. Myors, 1999; Plant, Hammond & Whitehouse, 2002; Plant, Hammond & Turner, 2004). Since there is a large and constantly varying range of computer systems on the market, and these computers are made up of the latest components from diverse suppliers, it is hardly likely that, on the hardware side, two different computer systems will yield identical measurement results. As an example, Plant, Hammond and Whitehouse (2003) showed that even changing the computer mouse applied, could have a major effect on time measurements.

The particular problem of technical errors of measurement at millisecond level is that the user is not in a position to witness if such errors exist. It is obvious that technical errors of measurement cannot be identified "with the naked eye". Moreover, methods of analysing the reliability of tests suffer from a similar problem. The most problematic aspect of technical errors of measurement is that they represent a (computer) system-dependent bias, which cannot be reduced by extending the length of a test. In a typical reliability study each participant works a certain number of items on one out of a number of computer systems. The internal consistency is calculated from the ratio of the variance within the subject's reactions to the variance between the different subjects (cf. Dawis, 2000). Since a respondent does not change computer system during the test session, the variance within his reactions is unaffected by any system-dependent bias; the variance between subjects, however, is increased by the system-dependent biases, since it is affected not only by the variation between the individuals tested but also by variation between the computer systems used. Rather than becoming apparent in reliability studies, technical errors of measurement even lead to overestimation of the internal consistency of the test investigated.

Typical reasons for technical errors of measurement and strategies for compensation

Technical errors of measurement can be considered as a sum of numerous small errors in different parts of the measurement process. They can in principle arise in connection with three different aspects of time-based tests: the presentation of the stimulus material, the internal timer and the detection of reactions. In modern computer systems the accuracy of the timer is usually given (MacInnes & Taylor, 2001). However, it is theoretically possible for interference from high-priority processes which are running in the background to lead to brief interruptions in the test program. To a large extent, this can be prevented by appropriate programming of the tests. It is nevertheless advisable to use the computers only for test administration and to ensure that during the assessment process no other programs are running in the background.

Technical errors in the detection of reactions are usually dependent on the input device used. It is perfectly possible for delays in the detection of the reaction occasioned by characteristics of the input device to be of the order of several milliseconds. The use of standardised input devices provided by the manufacturer of the test system is therefore recommended. While this does not guarantee that errors of measurement of this type will not occur, the user can at least expect that any of these errors will have been taken into account in the norms supplied with the test.

Aspects of the presentation of the stimulus material are also an important source of technical measurement errors. It is hardly surprising that a stimulus does not appear on the screen at exactly the moment at which the program gives the relevant order. The processing of the command and writing the data into the memory of the graphics card requires a certain amount of time, which depends on the performance of the computer and the graphics card. Moreover, programs intended to optimise graphics display can have a disruptive effect. An example is the Desktop Window Manager (DWM; cf. MSDN, 2007) in Microsoft Windows Vista. It must also be borne in mind that the screen is only updated at a particular refresh rate, and that this causes further delay. In case of synchronous screen display mode, test presentation programs can take refresh rates into account and can therefore precisely predict when graphical information will become visible on screen (cf. Xie, Yang, Yang & He, 2005); however, many commonly available video cards do not support the vSync signal. A test that makes use of synchronous display would be incompatible with these systems or would yield significant measurement errors.

For the majority of the error sources mentioned above the size of the measurement error can be estimated from the technical specification of the computer system used. Thus the delay in picture display has an expectancy value of $E = \frac{500}{f}$ with an equally distributed range of values between $[0; \frac{1000}{f}]$, where f is the refresh rate of the monitor. The factor of 1000 is used because of the different dimensioning of display delay (ms) and refresh rate ($Hz = \frac{1}{s}$). Provided that the test developer has in-depth knowledge of the construction of computerised tests, many of the relevant sources of technical errors of measurement can be similarly compensated for by taking their expectancy values into account. What remains are mostly uniformly or normally distributed random error components with an expectancy value of 0. Thus, given a sufficient amount of items the remaining error is only small.

Relevance of technical errors of measurement in applied psychological assessment

Technical errors of measurement can have far-reaching consequences for the validity of diagnostic decisions. Two areas of applied psychological assessment that are particularly affected by the technical accuracy of reaction time measurements are traffic psychological assessment and clinical neuropsychological assessment.

In several European countries individuals have to undergo a mandatory medical-psychological assessment in order to prolong or regain their driver's licence. The result of this assessment has extensive legal consequences for the respondents (cf. Bukasa, Chaloupka & Christ, 2001; Kroj, 1995). In accordance with legal regulations in Germany (FeV Annexe 5.2) and Austria (FSG-GV §18), psychologists have to administer several reaction-time-based tests to assess the client's reaction speed and attention (cf. BASt, 2000; Kroj, 1995; Schuhfried, 2005). In this regard one has to keep in mind that not only the tests but also the process of judgment formation is governed by legal regulations (cf. BASt, 2000; Schubert, Schneider, Eisenmenger & Stephan, 2005). According to these regulations, drivers with driver license class A or B have to reach a percentile rank of PR ≥ 16 in all ability tests ad-

ministered in order to be considered to be fit to drive. If the respondent's performance lies below this critical threshold in one or more of the tests administered the diagnostician needs to ensure that these deficiencies can be compensated for through the respondent's strengths in other driving-related ability or personality traits, or that they can be attributed to situational variables not related to driving. If neither of these possibilities provides an adequate explanation of the respondent's performance but data obtained in the anamnesis argue for the respondent's fitness to drive, the diagnostician can conduct a standardised driving test to enable the respondent to demonstrate his/her fitness to drive (BASt, 2000; Schubert et al., 2005). The guidelines for assessment thus propose a hierarchical structure of the diagnostic process. Technical errors of measurement in the initial stages may thus affect the fairness, reasonableness and even the validity of the entire process. Let us consider a case in which the respondent exhibits a performance in reaction-time-based measures that is above the critical threshold (PR $\geq$ 16). However, as a result of technical errors of measurement his or her performance is assessed as lying below this threshold. At best, the respondent simply has to undergo some further stages in the assessment process which would not have been necessary. This 'merely' leads to an increase in motivational and financial strain. However, in the worst case the technical error of measurement may lead to an incorrect diagnostic decision. Furthermore, one has to bear in mind that technical errors of measurement may exert differential effects depending on the specific configuration of the computer. At the very least this will affect the quality criterion of fairness (cf. Kubinger, 2006).

Another area of applied psychological assessment that is heavily affected by the precision of the technical measurement of reaction times and response latencies is clinical neuropsychological assessment. According to Strubreither and Mayl (2004) and Sturm (2000), two of its main tasks are to provide (1) detailed descriptions of cognitive and personality-related deficiencies following brain injuries and (2) necessary information for a theory-based intervention planning. As has been outlined by various researchers, attention disorders are among the most common symptoms in a variety of brain injuries (cf. Lezak, 1995; Stuss, Shallice, Alexander, & Picton, 1995; van Zomeren & Brouwer, 1994). Most commonly used measures of the various facets of attention rely heavily on reaction times as their main variables. However most of these measures feature a small inter-individual variance (cf. Table 1) making them vulnerable to technical errors of measurement. For instance, due to technical measurement errors one may overlook deficiencies in intensity components of attention (e.g. alertness, sustained attention) due to the smaller range of their inter-individual variation in the measured response latencies while deficiencies in selectivity aspects of attention (e.g. selective, focused and divided attention) may still be detectable. This does not only affect the validity of the descriptions of cognitive deficiencies of the client, but also affects the intervention plan. Sturm, Willmes, Orgass and Hartje (1997) demonstrated that specific trainings tailored toward the deficiencies of the individual clients are required to improve their impairments and that non-specific training programs might even lead to deterioration in performance.

The sections above highlight only two areas of applied psychological assessment that may be effected by technical errors of measurement of reaction times. Nevertheless, the discussion indicates that technical errors of measurement may have a serious impact on the quality of the entire diagnostic process. In the following section we will thus outline some means of handling technical errors of measurement that have been proposed in the literature.

Handling technical errors of measurement

There are many different suggestions on how technical errors of measurement should be dealt with. Schubert et al. (2005) proposed to circumvent the problem associated with technical errors of measurement by resorting to confidence intervals of the main variables and interpret them in the client's favour. This, however, is not helpful, since – as demonstrated above – the confidence interval contains no information about the technical error of measurement. Both the individual being tested and the decision-maker on whose behalf the testing is being carried out are entitled to expect the most precise measurement possible. Any adjustment of error ranges (whether they are established empirically or pragmatically) in favour of one of these two parties – and thus by definition at the expense of the other – is not a viable solution.

Kubinger (1984) proposed an approach in which the error risks are weighted according to the extent of the undesirable effects that could arise from the particular error; errors that have seriously damaging consequences are weighted more strongly than errors that have more minor consequences. This at least ensures that the social damage caused by technical errors of measurement is as far as possible minimised. An alternative solution is Mastery Testing (de Gruijter & Hambleton, 1984; van der Linden, 1990); here the undesirable effect of potential decisions errors is taken into account at test level. On the basis of a probabilistic test model and assumptions of the undesirable effect to be expected from the possible errors of measurement and the undesirable effect that arises from extending test length, a tailored test (Lord, 1968) is ended at the exact point at which the resulting undesirable effect is minimised – that is, when the shortest possible test length is combined with the lowest possible weighted resultant decision error risks.

The solution proposed by Schneider (2007) involves a hierarchical diagnostic process in which additional assessment methods are called on if there are concerns that a test result may be affected by technical errors of measurement; however, this method also fails to solve the problem. The main weakness of this approach is that the sole basis for deciding whether additional tests should be used is the test result itself, which may be affected by technical errors of measurement. The only way of resolving this dilemma would be to measure each required dimension by means of several independent tests based on different methods of measurement. However this process would be relatively uneconomical.

Another way of avoiding technical errors of measurement suggested in the literature is marketing the entire hardware as an integral component of the test system (e.g. ART2020, Bukasa, 1999). This would appear to guarantee that the technical error of measurement is identical to the one which occurred during the norm studies. The technical error of measurement is thus assumed to be compensated for in the standardisation process. However, problems arise in the practical implementation of this approach. Computer systems of a particular type do not usually remain on the market for more than a few months. Furthermore, there is no guarantee that two computer systems built by the same manufacturer to an identical specification actually consist of completely identical components. The "identity" of the complete systems cannot therefore be guaranteed. To expect two systems manufactured in different years to be identical is entirely unrealistic. This means that this approach leads to problems, once one needs to update the norms after a period of, for example, 5-8 years. In order to handle this problem within this approach, each batch would thus need to contain norms, which take the specific technical error of measurement of that batch into account.

Formulation of the problem

Based on the theoretical considerations and the practical relevance of the technical precision of measurement of response latencies outlined above, the following series of studies aims to investigate two main research questions: (1) whether system specific technical errors of measurement of reaction times occur on modern computers and (2) whether self-calibrating test systems would provide a sufficient means to fulfill even the highest demands on the technical measurement precision of reaction times.

Since computerised psychological assessment is primarily conducted on Microsoft Windows based platforms, the studies will be limited to this type of computer systems. Some of the technical problems, mentioned earlier, would apply to other operating systems as well. Research by MacInnes & Taylor (2001) states that technical precision of measurement might be even harder to achieve on Apple or Linux platforms than on Windows based computers.

STUDY I: Precision of measurement for uncalibrated computer systems

The initial study was designed to investigate whether relevant measurement errors due to technical reasons are to be expected at all. The measurement setup was tested on 11 computer systems that were representative of those in practical service at the time of the study. In selecting the computer systems care was taken to ensure a wide variation in particular hardware and software components. Three frequently used time-critical tests from different developers were installed on the computer systems and tested. These tests were the subtest 'DG-Lokation' from the CORPORAL test system (Berg & Schubert, 1999), the subtest 'Alertness' from the TAP-M test battery (Zimmermann & Fimm, 2005) and the 'Reaction Test' RT/S9 (Schuhfried & Prieler, 1994) from the Vienna Test System. All the computer systems applied meet the system requirements of the versions of the tests that were used.

Method

To test the precision of the reaction time measurement, an 'artificial respondent' is used. The appearance of the visual stimulus is detected by means of a very fast and highly sensitive photo diode. The signal is amplified and transformed into a digital signal using a threshold value detector. This triggers a highly precise delay circuit, which after a pre-set interval closes a contact, causing a button to be pressed on the input device supplied by the test manufacturer as part of the system. Using the test's norm table, the delay circuit is assigned a time interval that corresponds to the pre-defined percentile rank. Figure 1 (top) gives a schematic view of the setup of the artificial respondent.

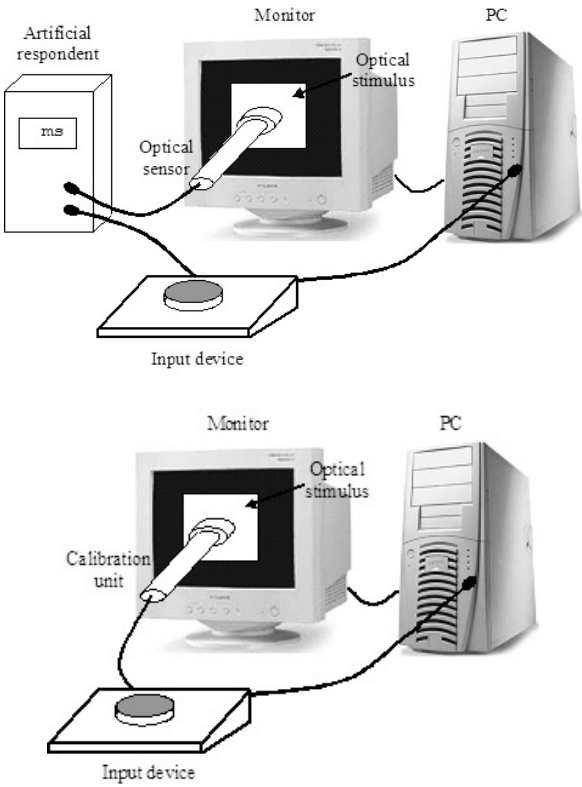

Figure 1:
Schematic setup of the 'artificial respondent' (top) and the calibration unit (bottom). The principal difference between the two systems is that the artificial respondent delays the transmission of the optical sensor's signal by an amount of time that can be set in milliseconds before activating the input device. It thus simulates the subject's reaction. The calibration unit, by contrast, transmits the signal without delay and thus provides a direct measurement of the time that elapses between intended and actual display of the stimulus.

The accuracy of measurement of the artificial respondent based on its technical specification is to be expected in the range of microseconds. For this study, target scores in the lower performance range were used, since erroneous measurements here can lead to judgments that have particularly severe consequences for the respondent. Since not all the tests used provide norms that are accurate to a precise percentile rank, reaction times in the region of a percentile rank of 21 or 33 were specified as target scores.

Due to the fact that computerised tests are expected to be fair on the level of single-case judgment, the measurement was aggregated over the length of one test run (CORPORAL: 128 items, TAP-M Alertness: 40 items, RT/S9: 28 items).

Results

Where reactions in the range of a percentile rank of 21 were specified, all tests yielded some measurement inaccuracies. As Table 2 shows, the measurement errors are substantially more significant for the CORPORAL test system than for the other two products.

Table 2:
Deviation of the normed test score reported by the test from the pre-set target value (at percentile rank 21) in percentile rank points for Corporal (DG Lokation), TAP-M (Alertness) and the Vienna Test System (Reaction Test).

	Measurement error in percentile rank points		
Hardware configuration	**Corporal**	**TAP-M**	**Vienna Test System**
Desktop PC 800MHz, 128MB RAM, WinXP Monitor Sony 446XS CRT 85Hz 1280x1024	-5	-3	+ 1
Desktop PC 1500MHz 512MB RAM Win2000 Monitor Nokia 920C CRT 85Hz 1024x768	-5	cannot be run	-1
Desktop PC 2400MHz 512MB RAM WinXP Monitor Samsung SyncMaster 192v 70Hz 1280x1024	-11	-3	+1
Desktop PC 800MHz 128MB RAM WinXP Monitor Samsung SyncMaster 192v 70Hz 1280x1024	-11	-5	0
Desktop PC 2800MHz 480MB RAM WinXP Monitor Samsung SyncMaster 193P 75Hz 1280x1024	-11	-5	0
IBM Thinkpad R51 Laptop 1500MHz 512MB RAM WinXP Monitor TFT 60Hz 1400x1500	-11	-3	-1
Acer Travelmate 517TE Laptop 366MHz 64MB RAM Win98 TFT 60Hz 1024x768	-11	cannot be run	-1
Acer Travelmate 722TX Laptop 500MHz 64MB RAM Win98 TFT 60Hz 1024x768	-11	cannot be run	0
Acer Travelmate 525TX Laptop 700MHz 128MB RAM WinME TFT 60Hz 1024x768	-11	cannot be run	-2
IBM Thinkpad 600E Laptop 500MHz 64MB RAM Win98 TFT 60Hz 1024x768	-11	cannot be run	-1
IBM Thinkpad 600 Laptop 500MHz 1600MB RAM Win98 TFT 60Hz 800x600	-16	cannot be run	-1

For the measurements taken with the CORPORAL test program the artificial respondent was set to a reaction time of exactly 707 ms. According to the norm score table, this corresponds to T=42 (PR=21). The reaction times recorded range between 712 and 776 ms, depending on the computer configuration used. This corresponds to percentile ranks between PR=16 and PR=5.

For the measurements taken with the TAP-M test program the artificial respondent was set to a reaction time of exactly 250 ms. According to the norm score table, this corresponds to T=42 (PR=21). The reaction times reported by the test program range between 259 and 262 ms, depending on the computer configuration used. This corresponds to percentile ranks of between PR=18 and PR=16.

For the measurements taken with the RT test program the artificial respondent was set to a reaction time of exactly 268 ms. According to the norm score table, this corresponds to a percentile rank of 20. The reaction times reported range between 266 and 278 ms, depending on the computer configuration used. This percentile ranks reported were between PR=21 and PR=18.

The results for reactions specified to be in the range of a percentile rank of 33 are very similar (see Table 3).

For the measurements taken with the CORPORAL test program the artificial respondent was set to a reaction time of exactly 684 ms. According to the norm score table, this corresponds to T=45 (PR=33). The reaction times reported by the test program range between 694 and 731 ms, depending on the computer configuration used. This corresponds to percentile ranks of between PR=10 und PR=21.

For the measurements taken with the TAP-M test program the artificial respondent was set to a reaction time of exactly 230 ms. According to the norm score table, this corresponds to T=46 (PR=34). The reaction times reported by the test program range between 239 and 243 ms, depending on the computer configuration used. The percentile rank reported was always PR=27.

For the measurements taken with the RT test program the artificial respondent was set to a reaction time of exactly 244 ms. According to the norm score table, this corresponds to a percentile rank of 33. The reaction times reported by the test program range between 243 and 254 ms, depending on the computer configuration used. The percentile ranks reported were between PR=29 and PR=34.

The results indicate that the technical accuracy of time measurement is influenced by a matter of milliseconds by the computer configuration applied. Particularly where tests have low inter-individual variance, technical errors of measurement can have a noticeable effect on the normed test scores. Nevertheless one can expect shifts of the order of a few percentile rank points in the normed test scores to affect the diagnostic judgment only in very rare cases.

However, the results reported here show that, depending on the computer configuration, very significant deviations are as well possible. They are probably caused, when a rather naive concept of programming is applied and typical sources of timing errors are not considered. Biases of that magnitude bear the risk to violate the criteria fairness and validity.

Table 3:
Deviation of the normed test score reported by the test from the pre-set target value (at percentile rank 33) in percentile rank points for Corporal (DG Lokation), TAP-M (Alertness) and the Vienna Test System (Reaction Test).

	Measurement error in percentile rank points		
Hardware configuration	Corporal	TAP-M	Vienna Test System
Desktop PC 800MHz, 128MB RAM, WinXP Monitor Sony 446XS CRT 85Hz 1280x1024	-12	-7	-1
Desktop PC 1500MHz 512MB RAM Win2000 Monitor Nokia 920C CRT 85Hz 1024x768	-12	cannot be run	-2
Desktop PC 2400MHz 512MB RAM WinXP Monitor Samsung SyncMaster 192v 70Hz 1280x1024	-17	-7	+1
Desktop PC 800MHz 128MB RAM WinXP Monitor Samsung SyncMaster 192v 70Hz 1280x1024	-17	-7	-4
Desktop PC 2800MHz 480MB RAM WinXP Monitor Samsung SyncMaster 193P 75Hz 1280x1024	-12	-7	-2
IBM Thinkpad R51 Laptop 1500MHz 512MB RAM WinXP Monitor TFT 60Hz 1400x1500	-17	-7	0
Acer Travelmate 517TE Laptop 366MHz 64MB RAM Win98 TFT 60Hz 1024x768	-17	cannot be run	-4
Acer Travelmate 722TX Laptop 500MHz 64MB RAM Win98 TFT 60Hz 1024x768	-17	cannot be run	-3
Acer Travelmate 525TX Laptop 700MHz 128MB RAM WinME TFT 60Hz 1024x768	-17	cannot be run	0
IBM Thinkpad 600E Laptop 500MHz 64MB RAM Win98 TFT 60Hz 1024x768	-17	cannot be run	-4
IBM Thinkpad 600 Laptop 500MHz 1600MB RAM Win98 TFT 60Hz 800x600	-23	cannot be run	-2

STUDY II: Precision of measurement for calibrated computer systems

Despite the deliberately wide range of the computer configurations used in Study I, there remains a risk that measurement errors could be even larger on other configurations, not represented in the study. It must also be recognised that these results are unlikely to be stable over time. The next generation of computer systems may use hardware or software that could

lead to significant measurement problems. Assessment of the measurement precision of different tests is therefore no more than a moderately effective quality assurance measure of the current state.

It would be more effective to incorporate the technology used in the artificial respondent directly into the user interface of the test system and to provide the user with a means of directly calibrating his computer system. Figure 1 (bottom) provides a schematic representation of such a calibration unit.

A similar approach was proposed by Plant, Hammond & Whitehouse (2002) in connection with the calibration of neuropsychological experiments. However, if computerised tests are to be calibrated, it is not sufficient to calculate the measurement error of the system configuration. The information obtained must be fed back into the test so that the measurements can be automatically corrected by the specific error. A calibration device that performs this function is available in the Vienna Test System 6.40 (2007). This study aimed to evaluate the effectiveness of this calibration function. As in Study I, an artificial respondent was used to calculate the measurement precision of different computer systems after calibration.

Results

Among the test systems used, only the Vienna Test System provides a self-calibration option. This study was therefore carried out only on the Reaction Test RT of the Vienna Test System. In principle it would be possible to calibrate the other two test systems using an external calibration system such as Black Box Toolkit (BBTK: Plant, Hammond & Turner, 2004) and to correct the measurements manually. It should, however, be borne in mind that there is little purpose in carrying out precise measurements at one workstation if the test has been normed at workstations that have not been calibrated and that are therefore affected by a bias of unknown size. Table 4 shows the measurement errors for the Reaction Test RT/S9 that remain after calibration.

It can be seen that after calibration only small errors of measurement remain. These are not caused by systematic inaccuracies of measurement, but by the random error of the individual reactions. Since these have an expectancy value of zero and a range corresponding to the reciprocal screen refresh rate, they are easy to estimate; as components of the test's internal consistency they are in any case taken into account in any study of reliability. Moreover, this residual error can by further reduced at will by increasing the number of items.

Discussion

The present article demonstrates that technical errors of measurement of reaction times do occur on modern computer systems. Some tests resort to reaction times or response latencies as main variables to enable the measurement of latent traits that cannot be assessed otherwise. However, some of these tests feature a very low inter-individual variance in reactions times. These tests are most vulnerable to the effects of technical errors of measurements which can give rise to significant judgments errors.

Table 4:
Deviation of the normed test score reported by the test from the pre-set target value for the Vienna Test System (Reaction Test) after calibration.

	Measurement error in percentile rank points	
Hardware configuration	Error Target value PR 20	Error Target value PR 33
Desktop PC 800MHz, 128MB RAM, WinXP Monitor Sony 446XS CRT 85Hz 1280x1024	-1	0
Desktop PC 1500MHz 512MB RAM Win2000 Monitor Nokia 920C CRT 85Hz 1024x768	0	0
Desktop PC 2400MHz 512MB RAM WinXP Monitor Samsung SyncMaster 192v 70Hz 1280x1024	0	0
Desktop PC 800MHz 128MB RAM WinXP Monitor Samsung SyncMaster 192v 70Hz 1280x1024	0	-1
Desktop PC 2800MHz 480MB RAM WinXP Monitor Samsung SyncMaster 193P 75Hz 1280x1024	0	+1
IBM Thinkpad R51 Laptop 1500MHz 512MB RAM WinXP Monitor TFT 60Hz 1400x1500	-1	+1
Acer Travelmate 517TE Laptop 366MHz 64MB RAM Win98 TFT 60Hz 1024x768	0	0
Acer Travelmate 722TX Laptop 500MHz 64MB RAM Win98 TFT 60Hz 1024x768	-2	-1
Acer Travelmate 525TX Laptop 700MHz 128MB RAM WinME TFT 60Hz 1024x768	-1	0
IBM Thinkpad 600E Laptop 500MHz 64MB RAM Win98 TFT 60Hz 1024x768	-1	+1
IBM Thinkpad 600 Laptop 500MHz 1600MB RAM Win98 TFT 60Hz 800x600	-1	+1

However, the results of the first study indicates that careful test development and detailed analysis of the potential sources of technical errors can compensate for a large proportion of these technical errors of measurement. The residual error range is rather small and relevant only where very high measurement precision is required. However, where test development has involved a more naïve approach to the problems of time measurement very large measurement errors are possible depending on the specifications of the computer systems applied.

The second study demonstrated that even the highest standards of measurement precision can be met comparatively simply by using self-calibrating test systems. The major advantage of this approach resides in the fact that it gives the diagnostician control over the computerised measuring equipment. Additionally, calibration systems provide the test manufacturer with technically high-quality means of evaluating their own tests. Error ranges and ways of compensating for them can thus be analyzed and clearly documented in the test manuals, further increasing the transparency of the test development process. Such systems also enable the user to check the measurement precision of the test and to keep the precision of measurement constant even in the face of major changes to the hardware used.

The Windows Vista operating system has further increased the technical demands made on highly precise time measurement. Tests that offered very precise measurement under Windows XP and earlier operating systems might run the risk of yielding significant technical errors of measurement under Windows Vista. An exploratory study has revealed a number of new sources of errors that could result in technical errors of measurement of up to about 100 ms. It is likely, therefore, that there will be an increasing need for sufficiently exact computerised tests measuring ability and personality traits, which feature a very small inter-individual variance in the reaction times or response latencies.

References

Berg, M.; & Schubert, W. (1999). Das thematische Testsystem "Corporal" zur Erfassung von Aufmerksamkeit [The thematic Test System "Corporal" for the assessment of attention]. *Zeitschrift für Verkehrssicherheit*, *45*, 74-81.

Bukasa, B. (1999). ART2020 - Das neue Multimedia Testgerät für die Fahreignungsbegutachtung [ART2020 - The new multi-media test device for traffic-psychological assessment]. In F. Meyer-Gramcko (ed.), *Verkehrspsychologie auf neuen Wegen. Herausforderungen von Strasse, Wasser, Luft und Schienen (I) [Traffic psychology on new tracks. Challenges of road, water, air and tracks (I)]* (pp. 381-401). Bonn: Deutscher Psychologen Verlag.

Bukasa, B., Chaloupka, Ch. & Christ, R. (2001). Die Besonderheit verkehrspsychologischer Tätigkeit [The features of the traffic-psychological occupation]. *Psychologie in Österreich, 21* (3), 116-121.

Bundesanstalt für Straßenwesen (2000). *Begutachtungsleitlinien zur Kraftfahrereignung [Guidelines for the assessment of respondents' fitness to drive]*. Bergisch Gladbach: Bericht der Bundesanstalt für Straßenwesen, Mensch und Sicherheit Vol. M 115.

Forster, K.I.; Forster, J.C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods*, *35*, 116-124.

Dawis, R.V. (2000). Scale Construction and Psychometric Considerations. In H.E.A. Tinsley & S.D. Brown (eds.), *Handbook of Applied Multivariate Statistics and Mathematical Modeling* (pp. 65-95). San Diego: Academic Press.

De Gruijter, D.N.M.; & Hambleton, R.K. (1984). On problems encountered using decision theory to set cutoff scores. *Applied Psychological Measurement*, *8*, 1-8.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. K. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*, 1464-1480.

Klinck, D. (2006). Computerbasierte Methoden [Computer-based methods]. In F. Petermann & M. Eid (eds.), *Handbuch der Psychologischen Diagnostik [Handbook of psychological assessment]* (pp. 226-241). Göttingen: Hogrefe.

Kroj, G. (1995). *Psychologisches Gutachten Kraftfahreignung [Diagnostic report on respondents' fitness to drive]*. Bonn: Deutscher Psychologen Verlag.

Kubinger, K.D. (2006). *Psychologische Diagnostik - Theorie und Praxis psychologischen Diagnostizierens [Psychological Assessment - Theory and Application]*. Göttingen: Hogrefe.

Kubinger, K.D. (1984). Nutzentheoretische Beurteilung differential-diagnostischer Entscheidungen [Profit-theoretical appraisal of differential-diagnostic decisions]. *Diagnostica*, *30*, 249-266.

Lezak, M. D. (1995). *Neuropsychological assessment* (3rd ed.). New York: Oxford University Press.

Lord, F.M. (1968). Some test theory for tailored testing. *Research Bulletin*, *69/4*. Princeton: Educational Testing Service.

MacInnes, W. J.; & Taylor, T.L. (2001). Millisecond timing on PCs and Macs. *Behavior Research Methods*, *33*, 174-178.

MSDN (2007). Desktop Window Manager. [Online] URL http://msdn2.microsoft.com/en-us/library/aa969540.aspx [15.05.2007].

Myors, B. (1999). Timing accuracy of PC programs running under DOS and Windows. *Behavior Research Methods*, *31*, 322-328.

Plant, R.R.; Hammond, N.; & Turner, G. (2004). Self-validating presentation and response-timing in cognitive paradigms: How and why? *Behavior Research Methods*, *36*, 291-303.

Plant, R.R.; Hammond, N.; & Whitehouse, T. (2002).Toward an Experimental Timing Standards Lab: Benchmarking precision in real world. *Behavior Research Methods*, *34*, 218-226.

Plant, R.R.; Hammond, N.; & Whitehouse, T. (2003). How choice of mouse may affect response timing in psychological studies. *Behavior Research Methods*, *35*, 276-284.

Puhr, U; & Wagner-Menghin, M.M. (1993). *Manual Interferenztest nach Stroop [Test manual: Interference test sensu Stroop]*. Mödling: Schuhfried.

Schneider, W. (2007). Zur Gültigkeit von Grenzwerten bei verhaltenswissenschaftlichen Testverfahren für die Frage nach der Eignung zum Führen von Kraftfahrzeugen [Validity of cut-off scores in behavioral tests for the fitness to drive]. W. Schubert & W. Schneider (eds.), *Stellungnahme der Deutschen Gesellschaft für Verkehrspsychologie zu dem Themenkomplex 'Objektivität, Validität und Fairness der im Rahmen der Fahreignungsdiagnostik eingesetzten psychologischen Testprogramme' [Statement of the German Society of Traffic psychology about the topics 'Objectivity, validity and Fairness in the context of tests in traffic psychological assessment]* (pp. 14-20). Berlin: DGVP.

Schubert, W., Schneider, W., Eisenmenger, W., & Stephan, E. (2005). *Begutachtungs-Leitlinien zur Kraftfahrereignung – Kommentare [Guidelines for the assessment of respondents' fitness to drive - Comments]*. Bonn: Kirschbaum Verlag.

Schuhfried, G. (2005). **Test** *Manual: Expert System Traffic*. Mödling: Schuhfried.

Schuhfried, G.; & Prieler, J. (1994). *Testmanual Reaktionstest [Test manual: Reaction test]*. Mödling: Schuhfried.

Strubreither, W., & Mayl, J. (2004). Neuropsychologie in Österreich: Entwicklung – derzeitige Situation – Ausblick [Neuropsychology in Austria: Development - State - Outlook]. In G. Mehta (ed.), *Die Praxis der Psychologie – Ein Karriereplaner [Application of psychology – A career-planner]* (pp. 187-206). Vienna: Springer.

Sturm, W. (2000). Aufgaben und Strategien neuropsychologischer Diagnostik [Duties and strategies in neuropsychological assessment]. In W. Sturm, M. Herrmann, & C. M. Wallesch (eds.), *Lehrbuch der Klinischen Neuropsychologie. Grundlagen – Methoden – Diagnostik – Therapie [Textbook of Clinical neuro-psychology. Fundamentals - Methods - Assessment - Therapy]* (pp. 265-276). Lisse: Swets.

Sturm, W., Willmes, K., Orgass, B., & Hartje, W. (1997). Do specific attention deficits need specific training? *Neuropsychological Rehabilitation*, *7*, 81-103.

Sturm, W. (2006). *Testmanual Wahrnehmungs- und Aufmerksamkeitsfunktionen [Test manual: Perception and attention functions]*. Mödling: Schuhfried.

Stuss, D. T., Shallice, T., Alexander, M. P., & Picton, T. W. (1995). A multidisciplinary approach to anterior attentional functions. *Annals of the New York Academy of Science*, *769*, 1991-211.

Van der Linden, W.J. (1990). Applications of decision theory to test-based decision making. In R.K. Hambleton & J.N. Zaal (eds.), *New developments in testing. Theory and applications* (pp. 129-155). Boston: Kluwer.

Van Zomeren, A. H., & Brouwer, W. H. (1994). *Clinical neuropsychology of attention*. New York: Oxford Univ. Press.

Vienna Test System 6.40 [Computer software] (2007). Mödling: Schuhfried. (http://www.schuhfried.at)

Wagner-Menghin M.M. (2003). Computerdiagnostik [Computerized psychological assessment]. In K.D. Kubinger & R.S. Jäger (eds.), *Schlüsselbegriffe der Psychologischen Diagnostik [Keywords in psychological assessment]* (pp. 68-82). Weinheim: PVU.

Xie, S.; Yang, Y.; Yang, Z.; & He, J. (2005). Millisecond-Accurate Synchronization of Visual Stimulus Displays for Cognitive Research. *Behavior Research Methods*, *37*, 373-386.

Zimmermann, P.; & Fimm, B. (2005). *Manual Testbatterie zur Aufmerksamkeitsprüfung - Mobilität [Test manual: Test battery for the assessment of attention - Mobility]*. Herzogenrath: Psytest.