Galileo Galilei–GG
Questions from the Science Assessment Review Panel (SARP) of ESA & Answers from the GG team

(Dated: October 31, 2017)

Premise: We appreciate all questions because they address key issues of the GG experiment in space and its relevance for fundamental physics.

I. QUESTION 1

Q1 The current levels of differential acceleration measurements performed on ground (cited in Table 2) require nearly 5 orders of magnitude extrapolation to the required flight GG resolution, with a novel flexure joint suspension that has not, to our knowledge, been used in small force measurements near the required level (10^{-15} N resolution for the 10 kg test masses). The reference to achieving statistical SNR = 2 in 3.5 hours (Sec. 2.72) is even more daunting, requiring differential acceleration noise near 5×10^{-15} m/s2/Hz$^{1/2}$, roughly the published LISA Pathfinder value. Ground testing, as we understand from the proposal, has fundamental limits due to the stiffer suspension required but also significant additional "technical 'noise that has not been resolved (factor 37)

What will be the strategy for verification of the required differential acceleration sensitivity for the GG apparatus, to limit uncertainties in the science return as early as possible in the project?

A1 Mechanical suspensions with low stiffness and high mechanical quality are used to suspend the mirrors of the laser interferometers in ground experiments to detect gravitational waves (GW), and to suspend the rotating torsion balance in experiments to test the weak equivalence principle (WEP). These experiments are ongoing since almost three decades and have both been highly successful.

The stiffness k can be reliably calculated. Instead, the mechanical quality factor Q is not predictable. However, it is well known that both k and Q can be measured with simple specific setups. As long as the natural frequency of oscillation, its amplitude, the clamping, the temperature and the pressure are the same as planned in the real experiment, the measured k and Q are totally reliable. We don't need the full experimental apparatus to measure k and Q. And for sure we don't need to fly them to find out! (CQG 16, 1463 (1999)).

Small force gravitational experiments are ultimately limited by thermal noise, and thermal noise from internal damping (ID) in the mechanical suspensions usually dominates; even in Microscope, which is based on electrostatic suspensions but uses a 7 μm gold wire for each test cylinder as an auxiliary 'dummy' spring.

Work carried out in relation to the suspensions of the mirrors in GW interferometers (PRD 42, 2437 (1990)) has shown that ID thermal noise force at temperature T depends on the frequency of interest. For a signal at frequency ω_{signal} the ID noise force (squared) is:

$$F^2_{th-id} = 4K_BT\frac{k}{\omega_{signal}Q(\omega_{signal})} \tag{1}$$

with K_B the Boltzmann constant. The quality factor $Q(\omega_{signal})$ is also frequency dependent. Nobody knows exactly how, except for the fact that it is higher at higher frequency. As a result, in a measurement limited by ID thermal noise the integration time to reach SNR=2 with a signal force $F(\omega_{signal})$ is:

$$\begin{aligned} t_{int-id} = SNR^2 \times \frac{F^2_{th-id}}{F(\omega_{signal})^2} = \\ = SNR^2 \cdot 4K_BT\cdot \\ \cdot\frac{1}{F(\omega_{signal})^2}\cdot\frac{k}{\omega_{signal}Q(\omega_{signal})} \quad . \end{aligned} \tag{2}$$

It is apparent that as long as the frequency of the signal is high, either by its nature (GW detectors), or by up-conversion (rotating WEP experiments), ID thermal noise is low and the integration time is short.

As far as the frequency of the signal is concerned, ground based GW detectors and rotating torsion balances are at opposite ends: the frequency of interest is between 40 and 2000 Hz in one case, end at 10^{-3} Hz (the rotation frequency of the balance) in the other.

By spinning at 1 Hz, the signal frequency of GG is 3 orders of magnitude above torsion balances (and above the Microscope signal by the same factor) and about 2 orders of magnitude below GW detectors. In orbit, with cylindrical symmetry 1 Hz rotation around the symmetry axis is realistic and feasible simply by angular momentum conservation after initial spin-up, with no additional motor/bearings noise.

At 100 Hz Virgo achieves its best sensitivity of about 6×10^{-20} m (40 kg mirrors suspended by 350 μm metallic wires); the sensitivity is even betters for the two LIGO (PRL, 119 114101 (2017), Fig. 2).

At 1 Hz, by detecting a displacement signal 7 orders of magnitude larger than that, of about 6×10^{-13} m, GG would reach its target of a WEP test to 10^{-17} (test cylinders of 10 kg, low stiffness U-flexures, $Q_{1\text{Hz}}$=20000) and improve over rotating torsion balance tests by 4 orders of magnitude.

At 10^{-3} Hz the rotating torsion balances have achieved the best WEP test so far, to 10^{-13} (70 grams balance, 20 μm suspension fiber, $Q_{1\text{mHz}} = 6000$). They have achieved ID thermal noise level and it has been found to be as theoretically predicted (CQG 29, 184004 (2012), Fig. 4).

A high precision test of WEP requires both rapid rotation (for up-conversion of the signal to high frequency) and low natural normal mode frequency ω_n (for high sensitivity). This means that, if possible, we should have $\omega_{spin} \gg \omega_n$.

It is well known from the theory of rotordynamics that a harmonic oscillator with $\omega_{spin} > \omega_n$ (known as super-critical rotor) must have two degrees of freedom, while it would be strongly unstable in 1D. Therefore, it was extremely important to demonstrate that Eq. (1) still holds in 2D, and −even more importantly− that in this case the strength of the up-converted signal is not reduced by the factor $(\omega_{spin}/\omega_n)^2$ as in 1D oscillators. Both issues have been demonstrated theoretically (PRL 107, 200801 (2011)); the non-attenuation of the signal has also been proven experimentally with the GGG demonstrator in the lab (CQG 29, 184004 (2012), Fig. 2).

We could therefore reliably calculate the integration time for GG. With ν_{signal}=1 Hz and $Q_{1\text{Hz}}$=20000 thermal noise from internal damping is only slightly larger larger than gas damping and Johnson thermal noise (both independent of frequency). By taking all of them into account we concluded that the integration time for GG to reach SNR=2 is of 2.4 to 3.5 hours, and therefore plan to have one full WEP test to 10^{-17} per day (PRD 84, 042005 (2014)). This ensures a very wide range of dynamical evolution, because the spin axis (hence the sensitive plane $\perp$ to it) is fixed in space while the normal to the sun-synchronous orbit changes by 1° per day. This allows the well known signature of a WEP violation signal to be separated with certainty from all systematic errors that have been identified over 40 years of investigation, by both US and European scientists since the first STEP proposal for a WEP experiment in space in the mid 1970s.

We note in passing that gas damping thermal noise could be reduced by allowing a 2 cm gap between the test cylinders. This is possible with the laser gauge readout because its sensitivity, unlikely that of a capacitive readout, does not require small gaps (600 μm in Microscope). A gap 33 times larger means patch effects 3 orders of magnitude weaker.

It is correct that for GG to reach SNR=2 in 3.5 h requires an acceleration noise of $4.5 \times 10^{-15}\ \frac{\text{ms}^{-2}}{\sqrt{\text{Hz}}}$. We are aware that this is about the level of acceleration noise measured by LISA pathfinder (LPF). However, there is a key difference, and that is frequency. GG must achieve this noise at 1 Hz, while LISA must ensure it at very low frequencies, also below 10^{-3} Hz (PRL 116, 231101 (2016), Fig, 1)).

The far reaching implications of this difference in frequency can be appreciated as follows. Imagine that GG were to target a violation signal to 10^{-17}, but at about 10^{-3} Hz (like in rotating torsion balances) rather than at 1 Hz. Let us be optimistic and assume that GG were able to realize U-flexures which, at 10^{-3} Hz, have $Q_{1\text{mHz}} \simeq 6000$ like the suspension fiber of the Eöt-Wash balances, though this is by no means an easy task. In such a case thermal noise from internal damping would dominate by far, since gas damping and Johnson noise, being frequency independent remain the same as calculated in PRD 84, 042005 (2014), Eq. (25). It is easy to conclude from (2) that for SNR=2 the integration time would be 1.3 years instead of 3.5 hours, ruling out the mission proposal altogether.

This shows how rapid rotation is extremely effective in drastically cutting the integration time; far more effectively than cryogenics. A cryogenic balance has been shown to be more sensitive than room temperature ones at zero spin. However, since it cannot rotate, it is in fact not competitive with a room temperature balance rotating at 10^{-3} Hz (E. G. Adelberger, personal communication (2017)).

In relation to a comparison with LPF it is worth considering the spectral density of the acceleration noise as measured in space by LPF and published in PRL 116, 231101 (2016), Fig. 1 (figure reported below). We notice that at 1 Hz the acceleration noise is higher than at 10^{-3} Hz by more than two orders of magnitude. Since in GG the signal is read at 1 Hz we might be lead to think that GG should face a similar large noise. However, if we use also the bottom plot in Fig. 1, which shows the extremely low relative displacement noise of LPF interferometer in space, we notice that, at 1 Hz the diplacement noise is about $\Delta r \simeq 30\cdot10^{-15}\ \text{m}/\sqrt{\text{Hz}}$, which translates into $\Delta a \simeq 30\cdot10^{-15}\cdot4\pi^2\cdot1^2 \simeq 10^{-12}\ \text{ms}^{-2}/\sqrt{\text{Hz}}$, in agreement with what is reported in the top plot. But this is not the transfer function in GG, where the signal (in the non spinning reference frame) is at the orbital frequency ($\nu_{orb} \simeq 1.7\cdot10^{-4}$ Hz), which is below the natural normal mode frequency in differential mode of the test cylinders oscillator ($\nu_{diff} \simeq 1.9\cdot10^{-3}$ Hz). Therefore, the oscillator responds at this frequency

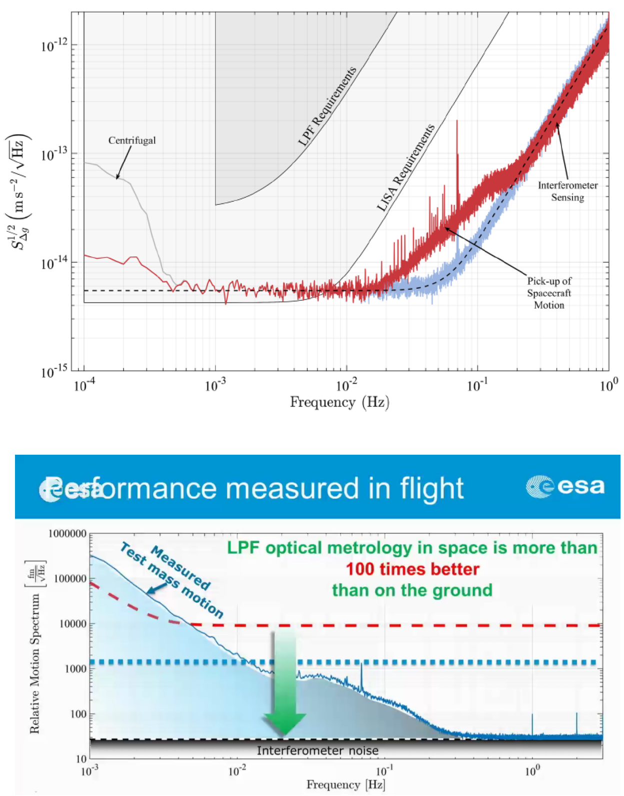

FIG. 1. *Top chart*: taken from Fig. 1 of PRL 116, 231101 (2016), Fig. 1. *Bottom chart*: relative displacement noise ($10^{-15}\,\mathrm{m}/\sqrt{\mathrm{Hz}}$) as measured in space by the LPF interferometer and reported by M. Hewitson at the LPF press conference in June 2016.

as: $\Delta a = 4\pi^2\nu_{diff}^2\Delta r$. Then, the key fact is that in a 2D oscillator this signal is not attenuated when up-converted to 1 Hz, as we have shown.

When aiming to test the WEP to 10^{-17} drag is 2.5 billion times larger than the target signal, and at the same frequency, and drag free control alone is not be viable. At 10^{-17} level no discussion on thermal noise would make sense unless the issue of drag effect has been taken care of. Since the target signal is differential while the effect of drag is common mode, GG has been designed as a balance with a high level of common mode rejection (by $\frac{1}{100000}$). This is feasible for two reasons. First, a balance in space is subject to a common mode acceleration many orders of magnitude smaller than on ground, therefore any unbalance yields a much smaller effect than on ground, which therefore can be easily compensated in GG (by changing the lengths of the coupling arms with inch-worm actuators). Second, with coupling arms of manageable dimensions, their lengths require adjustments by absolute amounts which are not challenging at all. We balance against the common mode effect of drag using its differential effect (which should ideally be zero) as the signal to be zeroed. We recall that balances on ground are balanced much better than we need for GG, to 5×10^{-10} (Metrologia 23, 87 (1986)). The capability for GG to reject common mode effects is crucial to reaching very high sensitivity to differential accelerations. It is by far better than in Microscope (whose concentric cylinders are not designed as a balance) and in LPF (whose test masses are 38 cm apart and one can only rely on the assumption that inside an accelerated frame test masses designed to be “the same” must feel “the same ” inertial acceleration).

Going back to thermal noise, a further striking confirmation of the dependence of ID noise on the frequency of the signal as given by (1) appears to come from Microscope, which is currently in orbit, successfully taking data to finally reach a WEP test to 10^{-15}.

Some preliminary results have been presented at the 656th Wilhelm und Else Heraeus Seminar in Bremen on ‘Fundamental Physics in Space’ (Microscope mission. A test of the Equivalence Principle in space, 23 October, 2017, by M. Rodrigues, P. Toubould and the Microscope team).

The frequency dependence of ID noise has also been demonstrated by Microscope in orbit. The team has reported that an increase of the spin frequency by about a factor of 4, from 7 mHz to 3 mHz (which means an increase of the frequency of the signal slightly less than that because $\nu_{signal} = \nu_{orb} + \nu_{spin}$), yields a reduction in the spectral density of the acceleration noise (in $\frac{\mathrm{ms}^{-2}}{\sqrt{\mathrm{Hz}}}$) by about 4 times, with a consequent reduction of the integration time by about 16 times. This occurs in the accelerometer with different composition test cylinders devoted to WEP testing, named SUEP, and it is very good news. However, it has been presented as partially unexplained because the $1/\omega_{signal}$ factor only in (2) and (1) would yield a factor of two lower spectral density of the acceleration noise, hence an integration time shorter by a factor of 4, not 16.

However, if random noise is dominated by ID thermal noise, the Q factor will also depend on the frequency; it is expected to he higher at higher frequencies and will therefore further reduce ID noise.

It is interesting to note that a quantitative measurement of Q for the same suspensions and the same apparatus at two different frequencies has been provided by the GGG demonstrator. Theory ensures that in super-critical rotors dissipation gives rise to whirl motion at the same frequency as the natural coupling frequency (except in cases of very high dissipation) and the time constant of whirl growth is proportional to losses at the spin frequency. Instead, when the system is not spinning oscillations at the natural frequency decrease in amplitude due to losses at this frequency (when the system rotates the suspensions are deformed at the spin frequency, while at zero spin only oscillations at the natural frequency produce losses). Since the spin frequency is higher than the natural one (by definition of a super-critical rotor) we should measure a higher Q from whirl growth during spin and a lower Q from oscillation amplitude decay when the system does not rotate. With the same system and the same experimental conditions (such as pressure and temperature) the measurement provides a reliable test of the theory of rotordynamics as well as a measurement of Q at different frequencies in a real complex experimental apparatus.

In GGG we have $\nu_n \simeq 0.074\,\mathrm{Hz}$ and $\nu_{spin} \simeq 0.16\,\mathrm{Hz}$. While in rotation, whirl motion at frequency $\nu_w \simeq \nu_n$ in the non rotating horizontal plane of the lab is found to grow with $|Q_{0.16\mathrm{Hz}}| = 2126$ (Q is negative). While not spinning, the amplitude of oscillations at frequency ν_n in the same plane is found to decay with $Q_{0.074\mathrm{Hz}} = 948$. Although this fact has been known for a long time, this is the first time that super-critical rotors are used in precision physics experiments where it can be demonstrated rigorously (PLA in press, online at http://dx.doi.org/10.1016/j.physleta.2017.09.027, see Fig. 8 and Fig. 9).

Thus, at a frequency higher by a factor of two, GGG measures a higher value of Q by a similar factor.

Should a similar improvement occur in the dissipation of the gold wires of the SUEP accelerometer, a factor of 4 higher spin rate would yield a higher Q by roughly the same factor, resulting in a 4 times smaller F_{th-id} and a 16 times shorter integration time.

The other fact reported by Microscope scientists is that the accelerometer with equal composition test cylinders, named SUREF, is less noisy than SUEP and does not improve when increasing the spin rate. This might be because in this case noise is dominated by gas damping and/or Johnson noise, which are both frequency independent and therefore would not be affected by the rotation rate. The very low $7\,\mu$m thickness of the gold wires involved (one for each test cylinder) appears to make it impossible to manufacture them from a single monolithic block. This results in the necessity to clamp 8 wire ends (2 for each wire) with glue, and such clamping, where losses are most likely to be higher, is not repeatable.

While a full understanding of this issue can come only from the Microscope team, it is apparent that up-conversion of the WEP putative violation signal to a frequency as high as possible is the key to a very high precision test of WEP in space.

It is worth noting that in addition to a spin frequency 333 times higher than the highest spin frequency of Microscope (and also less noisy because it does not need thrusters), GG can count on more massive test cylinders (hence lower thermal noise, as lower non gravitational accelerations in general). More importantly− a much higher $Q_{1\,\mathrm{Hz}} = 20000$ can be achieved at 1 Hz with monolithic suspensions made in CuBe. This value has already been measured by Virgo scientists for CuBe suspension wires at precisely 1 Hz oscillation frequency (PLA 255, 230 (1999), Fig. 2). We have measured a similar value for a monolithic helical spring made in CuBe (GG Phase A Study (1998-2000), Sec. 2.1.5) As recalled above, the GGG demonstrator, with two test cylinders coupled with CuBe monolithic cardanic joints has yielded $Q_{0.16\mathrm{Hz}} = 2126$ at a rotation frequency which is a factor 6.25 lower than GG. The GGG suspensions are necessarily much more complex in shape than the U-flexures planned for GG, because of local gravity, and the oscillation amplitudes are larger than they will be in space. In summary, the requirement of $Q_{1\,\mathrm{Hz}} = 20000$ at 1 Hz is realistic and can be measured in the for the actual U-flexures that will fly on GG.

As for the weakness of the suspensions (the weaker the better), we exploit another major advantage of space. In orbit the largest acceleration on GG is due to residual air drag and solar radiation pressure. The required nominal value is $2 \times 10^{-7}\,\mathrm{ms}^{-2}$ at the orbital frequency. (Note that this value is only about a factor 4.3 larger than the non gravitational acceleration of LPF which at the L1 point has no air drag but is anyway affected by solar radiation pressure).

In comparison with suspending bodies on ground against 1-g, a 100 kg mass inside GG requires the same stiffness needed on ground to suspend $\frac{2\cdot 10^{-7}}{9.8} \cdot 100\,\mathrm{kg} \simeq 2 \cdot 10^{-3}$ grams. As we can see, in orbit the limitation to the achievable weakness does not come from the mass to be suspended but rather from the necessity to fabricate monolithic suspensions in order to avoid large losses at clamping and also to ensure precision mounting, hence small off-centering errors. Torsion balances can manage $20\,\mu$m fibers but the $7\,\mu$m wires of Microscope appear to be too thin to be machined as a monolithic suspension. (Note that in GG the test cylinders and coupling arms will all be locked at launch; it has been verified that the weak U-flexures themselves can withstand launch accelerations.)

This leads us to the last part of question Q1, namely how we plan to verify that GG can reach the required sensitivity of $8 \times 10^{-17}\,\mathrm{ms}^2$ differential acceleration at 1 Hz.

As discussed above, mechanical suspensions are by no means an obsolete, old fashioned tool. To the contrary, they are at the heart of the most sensitive gravitational experiments ever. They are predictable (though not their losses) and measurable in the required experimental conditions. We have made sure that in GG U-flexures can be manufactured as monolithic ensembles (e.g. three U-flexures at 120° angular separation from one another, and not just as single monolithic U-flexure to be later assembled). They shall have enlarged ends where isolation can be provided (avoiding losses, because far from the bending portions) so that each flexure can serve also as an electric connection to the inch-worms placed on the coupling arms with no plastic or isolation material on the bending parts of the lamellae (which would degrade the Q). We recall that inch-worms are used to balance the two test cylinders so that they behave like a balance which can very effectively reject the remaining common mode acceleration of drag after drag-free control. The required precision is well within their current performance. U-flexure ensembles can be tested for measuring stiffness and losses. They can also be tested in combination with the test cylinders to assess the offset errors. With large test masses and no electrodes (whirl control electrodes are small and attached to the PGB shaft, not to the coupling arms of the test cylinders) offset errors can more easily meet the requirements.

The 'factor 37' recalled in Q1 comes from the fact that at 1-g the GGG suspensions provide a coupling stiffer than planned for GG in space, and this results in a sensitivity to differential accelerations a factor 1600 times worse (as the ratio of the respective natural frequencies squared). There is no way to bridge this gap with a full scale demonstrator in the presence of an acceleration 50 million times bigger than the largest one in orbit.

In its best long run, the sensitivity of GGG to differential accelerations at the frequency of $1.7 \cdot 10^{-4}$ Hz (the orbital frequency of GG, and the frequency of a violation signal from Earth before up-conversion), up-converted to the 0.16 Hz rotation rate of GGG, is $4.76 \times 10^{-12}\,\mathrm{ms}^{-2}$. Instead, GG must reach $8 \times 10^{-17}\,\mathrm{ms}^{-2}$ in order to meet its target. Thus, GGG is about a factor 59000 away from the target. Taking into account that a fraction 1600 of it is due to the stiffer suspensions on ground, it is concluded that GGG could be improved only for the remaining 37 factor.

After the implementation of a 2D weak joint for passive attenuation of local terrain tilts and seismic noise (CQG 29, 184004 (2012), Fig. 5), GGG appears to be limited by ball bearings noise. The possibility of implementing less noisy air-bearings has been studied. However, since neither terrain tilt/seismic noise nor bearings noise are present in orbit, there is really not much more to learn from GGG after such efforts in comparison with the great deal that we have already learned about a 2D harmonic oscillators made of weakly coupled concentric cylinders in super-critical rotation which was never realized before. Instead, we can improve and test the various components of the GG instrument at the level required by the space experiment.

One such component are the suspensions and their assembling with the test cylinders, for which we plan to follow the strategy outlined above. As reported in the proposal, we also have a well defined procedure for measuring the quadrupole mass moments of the test cylinders, for them to be adjusted until the requirements are met which make a well known systematic effect smaller than the signal.

The laser interferometry read-out for a 1 Hz signal is not a challenge and the noise level demonstrated at INRIM is already better than required. Adaptation to the specific system must be investigated but no major issues are foreseen. After the outstanding results of LPF laser interferometers are a proven technology of ESA.

After GOCE, LPF and Microscope drag-free control is also a well proven technology. For GG it must be adapted to 1 Hz, but the required level of drag compensation is less demanding than in Miccroscope, and much less than in LPF because GG is concerned with a signal which is differential in its nature and therefore relies on a very effective rejection of common mode effects (such as drag) by the test masses themselves. During Phase A-2 Study by TAS-I the spin rate of GG has been reduced form 2 to 1 Hz in order to be conservative on the performance of drag-free control.

Whirl control is conceptually similar to drag-free control: the effect to be controlled is at low frequency in the inertial frame while the sensors and actuators spin at 1 Hz. The required algorithms and a rotation sensor up to the task have already been developed at Phase A-2 Study level.

A crucial tool will be a new end-to-end simulator, along the lines of the one developed during Phase A-2 Study, based on the GOCE simulator, which already provides a correct representation of the system. The new simulator should incorporate all the major components according to the latest design and as tested and measured separately.

Based on the current status of the proposed GG mission, we conclude that −should GG be selected as a candidate mission for M5− by the end of the assessment study it will be possible to provide ESA with unquestionable evidence that GG will be able to reach the required differential acceleration sensitivity and meet its target within the timeframe of the M5 mission.

II. QUESTION 2

Q2. What is the current landscape for interpreting a possible null result in light of evolving theoretical work and the possibility of a null result at the 10^{-15} level by Microscope (with results likely in the very near future)? What are the leading theories and to what extent can they be confirmed or ruled out by a null result at the 10^{-17} level? The proposers are also asked to comment on the specific impact of their results on the dark matter question, alluded to in the proposal introduction.

A2. Leading theories in fundamental physics call into question the validity of the weak equivalence principle. These are: theories trying to go beyond General Relativity (GR) and the Standard Model of particle physics (SM), such as string theory; theories trying to explain Dark Energy (DE); theories trying to understand Dark Matter (DM).

In these vast areas of physics, and for all the theories involved, the level at which WEP has already been confirmed is a milestone which cannot be ignored. Microscope will soon make this milestone harder to face, forcing theories to come to terms with it.

This is because WEP experiments probe the possible existence of new physical interactions, potentially much weaker than gravity, which cannot be detected in direct experiments by other means.

String theory aims to unite relativistic quantum theory with GR. In string theory the dynamical nature of all coupling constants leads to WEP violation, but the level of violation would be in blatant contrast with current WEP tests. The majority among string theorists seek 'string vacua' which would stabilize all moduli fields at the minimum of some effective potential, thus preventing variation of the coupling constants and consequently forbidding any violation of the WEP (Damour, *Theoretical aspects of the equivalence principle*, CQG 29 (2012)). WEP experiments put to test this general assumption of string theory and could refute it altogether.

A different approach which some string theorists follow is to try to reconcile the existence of massless moduli fields with phenomenology (Damour, Piazza & Veneziano: *Runaway Dilaton and Equivalence Principle Violations* PRL 89 (2002) & *Violations of the equivalence principle in a dilaton-runaway scenario*, PRD 66 (2002); Damour & Donoghue, *Equivalence principle violations and couplings of a light dilaton* PRD 82 (2010)). This approach leads to an 'existence proof' of WEP violation below the currently tested level of about 10^{-13}, down to about 10^{-18}, spanning five orders of magnitude. Microscope at its target level would probe this conclusion over two orders of magnitude, from 10^{-13} to 10^{-15}. GG would probe it two orders of magnitude deeper, almost to the point of completely ruling it out.

Violation of the WEP is intimately related to a variation of the fundamental constants, in particular the fine structure constant α. It has been shown (Dvali & Zaldarriaga, *Changing α with Time: Implications for Fifth-Force-Type Experiments and Quintessence*, PRL 88, (2002)) that the time variation of α implies the existence of a very weakly coupled ultralight scalar field ϕ. Moreover, a dimensionless quantity λ_ϕ that quantifies this time variation, has been quantitatively related to the Eötvös parameter of WEP violation. It is concluded that realistic levels of λ_ϕ could be seen in WEP experiments at 10^{-16} level. While this level is still beyond reach for Microscope at 10^{-15} target sensitivity, its result will necessarily require a reassessment of the prediction made. With a WEP tests to 10^{-17}. With GG we will be able to infer a time variation of the fine structure constant and establish its order of magnitude.

In recent years dark energy has become a burning issue in theoretical physics, since its existence is no longer in doubt and its explanation in the form of a cosmological constant cannot be reconciled with the rest of physical theory. No wonder that ESA is building a space mission devoted to establishing the nature of DE. The number and scope of DE models and theories has been steadily increasing (see the Euclid theory paper: Amendola et al., *Cosmology and Fundamental Physics with the Euclid Satellite*, Living Rev. Relativ. (2013) 16: 6, and references therein).

Models of dynamical DE introduce new degrees of freedom, the simplest of which is a scalar field coupling to matter. In order to match the experimental data, the coupling is either assumed to decay from a large initial value and become as small as observed at the current epoch, or it is allowed to be large but a dynamical mechanism is postulated which acts on small scales to screen its effects and thus match the experiments (such as in the so-called chameleons).

The tightest constraints on such models are derived from solar system tests of gravity and, with much greater probing power, from tests of the WEP. One of the main objectives of Euclid is to discriminate between DE as the cosmological constant, and DE as a dynamical scalar field In dynamical DE models in which a scalar field couples to the standard model fields it would violate the WEP. In other models in which the scalar field is uncoupled, or only couples to DM, the WEP would not be violated. Thus, should Euclid observations favour the dynamical DE scenario, tests of WEP could discriminate between these two different classes of DE models.

With four orders of magnitude better precision than now, the role of GG would be of major importance.

In fact, a null result from Microscope at its design sensitivity (or even worse) may already have ruled out chameleon field theories based on screening mechanisms which would be effective on ground but not in space, where they allow a WEP violation even larger than the levels which have already been ruled out on ground (see Khoury and Weltman, *Chameleon Fields: Awaiting Surprises for Tests of Gravity in Space*, PRL 93 (2014) & *Chamaleon Cosmology* PRD 69 (2004); Mota & Shaw, *Strongly Coupled Chameleon Fields: New Horizons in Scalar Field Theory*, PRL 97 (2006); Joyce et al. *Beyond the Cosmological Standard Model*, Physics Reports 568 (2015)).

Tests of the weak equivalence principle are of fundamental importance also for dark matter, of which very little is known except that it accelerates ordinary baryonic matter and deflects photons. Our current understanding of the cosmos, of which non luminous DM is a significant fraction, larger than the luminous one, relies on the assumption that DM's only significant coupling to ordinary matter is via the gravitational interaction. This assumption should be tested by the most sensitive experiments (Stubbs, PRL 70, (1993); Smith et al., *Test of the Equivalence Principle for Ordinary Matter Falling toward Dark Matter*, PRL 70 (1993)).

Two test masses of different composition which are found to obey the WEP in the gravitational field of the Earth and the Sun may indeed fall differently toward the DM in our galaxy if DM couples to the ordinary matter of which they are made of via a new long range physical interaction other than gravity. This test has been carried out by the rotating torsion balances of the Eöt-Wash group (with the same sensitivity to differential accelerations as in the field of the Earth and the Sun). By using the best current models for the DM in our galaxy, they found a null result to a few 10^{-5} (Wagner et al., CQG 29, (2012)). This means that, should there be a long-range coupling of DM to ordinary matter other than gravity, it must be smaller than this.

Of course, any coupling of DM to DM only, can be investigated only at large astrophysical scales.

At the 10^{-17} target GG would improve the torsion balance ground tests in the field of the Earth by a factor of 100000 of which a factor 500 comes only by a stronger driving acceleration in orbit as compared to that on the suspended masses of the torsion balance; the remaining factor of 20 must be gained by making GG in space 20 times more sensitive to differential accelerations than the rotating torsion balance. We strongly argue that this is possible for GG because it has been designed to take full advantage of the space environment (absence of weight, no terrain tilt noise, rapid passive rotation of the whole lab/spacecrfat by conservation of the angular momentum..) while also ensuring a (peculiar) 'balance' design for the rejection of common mode effects as in ground balances. While the 500 factor comes for free, but only in the field of the Earth, the factor 20 holds in the field of source masses far away such as the Sun or the DM. GG will therefore improve the DM test of the torsion balance discussed above (as well as the WEP test in the field of the Sun) by 20 times.

With a 10^{-15} target in the field of the Earth, Microscope relies heavily on the factor 500 gain in orbit in order to improve over the torsion balance result at 10^{-13}, and can indeed achieve its target by being a factor of 5 less sensitive to differential accelerations than torsion balances. As a result, Microscope will not be able to improve over the torsion balance tests toward the Sun or the DM.

The possibility of a direct detection of dark matter by means of different-composition accelerometers has been suggested in recent years ((Carrol et al., *Implications of a scalar dark force for terrestrial experiments*, PRD 81 (2010); Graham et al., *Dark matter direct detection with accelerometers*, PRD 93, (2016)). Although this is a further interesting possibility, the wide variety of theoretical models and parameters involved suggests some caution.

As a general comment after this rapid excursion we would like to draw the attention of the panel on the fact that the number of papers being published which refer, in a way or another, to tests of the equivalence principle is extremely large. Moreover, discussing the various issue in more details is very difficult. Theories and theoretical models are complex subjects on their own; on the other hand, experiments which have attracted physicists since the time of Galileo are very subtle. As an example, in the work cited above by Carrol et al., PRD 81 (2010) it is assumed that if a space experiment improves the torsion balance ground test by 10^5 (it referred to STEP as proposed to NASA around that time), the torsion balance test toward the DM in our galaxy, would also improve by the same factor, which is obviously not the case as we have shown.

In our opinion, should GG be selected, a key task of the Assessment Study would be to bring together the best theorists and experimentalists in order to build up a common, deeper understanding of these difficult issues for a real progress in fundamental physics which −as we have argued in the proposal− can come only from space.

III. QUESTION 3

Q3. There is little discussion of the choice of materials to be used for the "violation" accelerometer. How does the Be / Ti baseline choice maximise the impact of GG as a WEP test?

A3. The question of which materials to use is not settled and the choice may be left for the assessment study phase should GG be selected. The issue has been given much attention by the Eöt-Wash group in their rotating torsion balance experiments and, in the context of space missions, in the STEP and Microscope projects.

The general idea is to span the largest possible volume in the space of the atomic properties which are the most likely sources of a WEP violation, such as the baryon number B, the lepton number L and the z component of isospin $I_z = N - Z$ (Fischbach & Talmadge, *The Search for Non-Newtonian Gravity*, Springer-Verlag, New York, 1998).

Eöt-Wash has studied Be−Ti and Be−Al test bodies. In STEP three materials, Pt/Ir alloy, Nb and Be, were planned, in a cyclic arrangement. Microscope uses Pt (10% Rh alloy) for the 'null' test body SUREF and Pt−Ti for the SUEP test body. In the GG proposal, the motivation for the initial choice Be−Ti was to build on the experience of the Eöt-Wash group and allow a ready comparison with their results.

On the other hand, the theoretical approach must be balanced against practical issues such as manufacturability, precise machining, and any other real-world issues that could hinder or mask the effect to be measured.

A unique characteristic of GG is that the concentric test cylinders spin around their own symmetry axis, hence any imperfections fixed with the cylinders (such as mass anomalies) produce DC disturbances in the rotating reference frame in which the signature of the violation signal is well known: 1 Hz frequency and Earth-pointing phase all the time.

As long as an anomaly is fixed on the cylinders, its effect could very well be much larger than the signal and yet don't affect its detection at all. Even an imperfection slowly changing with time, with a frequency component whose effect might turn out to be very close to 1 Hz in the readout, could still be separated from the violation signal because there is no reason why it should point to the center of mass of the Earth all the time (like the signal) since it has nothing to do with Earth.

This would be the case with patch effects, which after the GP-B mission of NASA have become a sort of nightmare for all high precision experiments in space. Although electric patch effects have been shown to be small in dedicated lab tests carried out with the GGG demonstrator, they will be further strongly reduced in GG by allowing a larger gap between the test cylinders; with a 2 cm gap instead of 600 μm as in Microscope, the effect is reduced by three orders of magnitude. Should an electric patch component at low frequency still be large enough to compete with the signal, it would be distinguished from it because of not being Earth-pointing.

The large mass of the GG test cylinders (10 kg) is another unique feature of GG, made possible by the considerations outlined above and ultimately by rotation around the symmetry axis. If imperfections are affordable, then why not using large masses? In WEP tests what matters are accelerations, not forces. But while gravitational and inertial forces do not depend on the mass (because of the equivalence principle!) all non-gravitational accelerations do, and are small for bigger masses. In particular, thermal noise and thermal noise due to internal damping is lower for more massive bodies.

Larger, more massive test cylinders also mean that it is easier to manufacture them with small errors. The total mass of the Eöt-Wash rotating balance is 70 g, of which 40 g for the test masses themselves, with four of them for each selected material, hence about 5 g for each mass. Meeting tolerances of 10^5 means the need to take care of $5 \cdot 10^{-2}$ milligrams. For the 10 kg test cylinders of GG the same tolerances require to control 0.5 grams, which is obviously much easier.

As a result, GG can test materials which are known to probe for composition dependence much more deeply than others, but so far could not be used because of manufacturing problems. For this reason during the GG Phase A-2 Study by the Italian Space Agency (ASI) carried out in 2008-2009, it has been considered the possibility for GG to use test cylinders made of Pb-C_2H_4 on the grounds that, with the same experimental sensitivity to differential accelerations it would be possible to test for WEP violation more deeply than by using 'standard' materials, possibly by one order of magnitude (Nobili et al., GG Phase A-2 Study, ASI (2009), Sec. 5.4). Indeed, the Eöt-Wash group is considering using Pb-C_2H_4 with their torsion balances (Wagner et al., *Torsion balance tests of the weak equivalence principle*, CQG 29, (2012)).

In 2013 a detailed analysis of the properties of different materials to be used for testing the WEP has been carried out (Hohensee et al., *Equivalence Principle and Bound Kinetic Energy*, PRL 111 (2013)) which allows us to quantify the expected advantages of various atoms/materials as far as testing the WEP is concerned.

The top chart in Fig. 2 has been published in their work. The other two plots have been kindly pro-

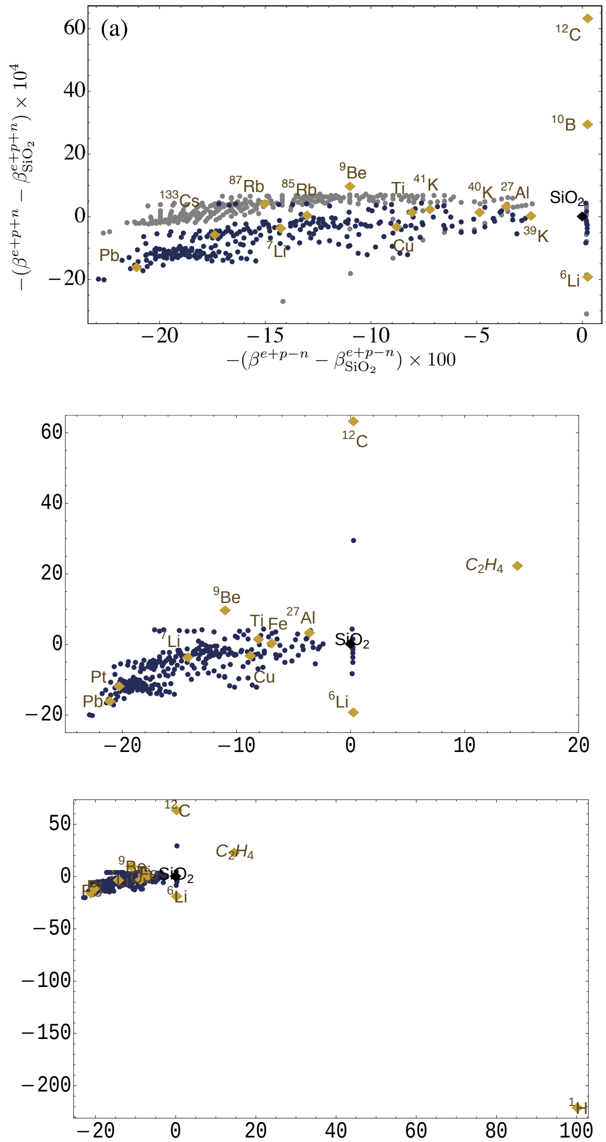

FIG. 2. *Top chart*: Assessment of the contribution of different atoms to a possible violation of the equivalence principle as reported in Hohensee et al., PRL 111 (2013). The spread along the vertical axis allows a direct quantitative comparison between different pairs of atoms; the larger the vertical separation, the larger the contribution to a possible violation. *Medium chart*: same plot as above (same units) with the addition by the authors of C_2H_4 for the benefit of the GG experiment (the zoom in is also helpful). It turns out that Pb and C_2H_4 would perform better than Be and Ti by a factor of 12. *Bottom chart*: the plot has been extended to the right to include the H atom, showing that H differs the most from all atoms and would therefore give the highest contribution to a violation of the equivalence principle.

duced by the authors for GG (Mike Hohensee is one of the supporting scientists of GG). The conclusion (from the middle chart) is that with test masses made of Pb-C_2H_4 instead of Be-Ti, assuming the same sensitivity to differential accelerations, the experiment would test the WEP a factor 12 more deeply. For GG, this would amount to reaching slightly better than 10^{-18}! The bottom chart is worth showing because it provides a beautiful plastic demonstration that the best atom to put to test for WEP violation, in combination with any other one, would indeed be the hydrogen atom. Of course this would be unpractical, but it has lead us (as well as the Eöt-Wash group, to seriously consider testing molecules such as C_2H_4 (polyethylene) because of its very high hydrogen content.

The choice for Be-Ti reported in the proposal has been motivated by the fact that it has best used in the best tests carried out so far with rotating torsion balances, and for this reason it is used in all theoretical work in which the authors try to compare their predictions (usually just estimates) with experimental results. By considering the same atoms being tested in GG with 10^4 times better precision we can readily establish the applicability of the same theories to GG.

This consideration aside, and also in view of the arguments given above, we consider the choice of material an open issue (and indeed a very important one) to be investigated during Assessment Study within ESA M5 should GG be given this opportunity. Should this be the case, it will almost certainly be possible to count on the expertise of the Eöt-Wash colleagues who are alraedy trying to realize test masses in C_2H_4 for their torsion balance tests of the WEP (Wagner et al., CQG 29, (2012)).

IV. QUESTION 4

Q4. What is the strategy behind the choice of the "control" accelerometer with two TM of the same composition (as opposed, for instance, to an accelerometer that inverts the geometric configuration of the two TM)? Will the control accelerometer have the same resolution?

A4. With two composition dipoles (each an accelerometer/balance constituted by two test cylinders of different composition) and four materials potentially at our disposal, the first question is whether having a 'control' accelerometer made of two test masses of the same composition is really beneficial with respect to adding another pair with different composition.

The Eöt-Wash group did not choose to test a null configuration, although they could have easily done so, by rearranging the four TMs normally configured as two composition dipoles into a quadrupole, hence averaging the effect to zero (Adelberger, personal communication (2017)). Rather, they inverted the composition dipoles on the pendulum to cancel systematic effects that followed the pendulum frame rather than the test bodies themselves. Having done that, systematic checks done at run time were deemed sufficient (Wagner et al., CQG 29, (2012)).

Flying only one composition dipole might limit the scientific return of the mission because a single composition dipole may not be particularly sensitive to the potential WEP violation due to an unfortunate 'mixing angle' (which depends on what form of the standard-model field the new scalar field couples to) (Wagner et al., CQG 29, (2012)). In addition to increasing the space of potential 'charges' probed by the experiment, two composition dipoles would be helpful in breaking this degeneracy.

In the configuration of GG (two nested coaxial pairs of TMs, each TM rotating about its own symmetry axis, all symmetry axes nearly coincident by construction and self-centering by supercritical rotation), inverting the position of the two TMs in a dipole is quite unreasonable (the denser and smaller test cylinder should naturally be the inner one of the pair).

In GG, we control the experiment by placing requirements on mass and shape of each TM, and by allowing for numerous systematic checks. As befits a balance, the TMs in each pair must have equal mass, although the tolerance need not be very stringent. Another requirement applies on knowledge of the TM mass moments (not very stringent, too, for the reasons mentioned under Q3 above), and a tested procedure is available to measure the moments of inertia of the test cylinders and their fractional differences, which are relevant in the error budget. Moreover GG can count on powerful checks. As presented in the proposal, the precession of the orbit normal around the spin axis allows checking for systematic effects that depend on the angle between the two (null checks). Above all, the integration time is short (see Q1 above) and allows plenty of time left for a thorough search for systematic errors based on rigorous celestial mechanics.

The added value of the null pair lies in establishing a posteriori the sensitivity of the GG ?balance?. This applies insofar as the null pair and the composition dipole pair can be made identical except for the composition; otherwise the sensitivity established for the null pair may not apply to the WEP pair. The nested (Russian doll) configuration of GG, despite its attractive features (unlike Microscope, all GG test cylinders are coaxial and have the same center of mass), leads to differences which cannot be avoided and will make the dipole pair and the null pair different by construction. All things considered, two dipole composition pairs may serve the objectives of the mission better. This discussion may be continued and concluded as part of the investigation planned for the Assessment Study, should GG be selected.